Spring 2018

The effects of speaking rate manipulations on the perception of voicing contrasts

Sarah M. Howell
James Madison University

The Effects of Speaking Rate Manipulations on the Perception of Voicing Contrasts

An Honors College Project Presented to

the Faculty of the Undergraduate

College of Behavioral and Health Sciences

James Madison University

by Sarah M. Howell

Accepted by the faculty of the Psychology Department, James Madison University, in partial fulfillment of the requirements for the Honors College.

FACULTY COMMITTEE:

Project Advisor: Michael D. Hall, Ph.D.
Professor, Psychology

Reader: Tracy E. Zinn, Ph.D.
Professor, Psychology

Reader: Jeffrey T. Andre, Ph.D.
Professor, Psychology

HONORS COLLEGE APPROVAL:

Bradley R. Newcomer, Ph.D.,
Dean, Honors College

PUBLIC PRESENTATION

This work was accepted for presentation, in part or in full, at the JMU Department of Psychology Student Symposium on 4/23/18.

Table of Contents

Acknowledgments

First and foremost, I would like to thank Dr. Michael Hall for pouring so much time and energy into this project in order to help me see it through to completion. This project has been challenging and extremely rewarding, and I could not have accomplished this much within the past two semesters without Dr. Hall's tireless work and insightful advising. I would also like to thank the students in his lab for helping run participants and for their support throughout this process.

Furthermore, I would like to thank those on my committee who served as draft readers. Dr. Tracy Zinn and Dr. Jeffrey Andre provided valuable feedback and asked insightful questions that contributed to the finalized product of this thesis project.

I further thank the Psychology department as well as the Honors Program at James Madison University for providing me with the support and resources necessary for this project. I am especially thankful to have had access to the department's auditory perception lab, which offered me the materials needed to construct stimuli and collect data.

Lastly, I want to thank my professors, mentors, friends, and family for supporting me throughout this past year. I truly could not have accomplished all of this if it weren't for their unwavering support.

Abstract

Phoneme-level research involving speaking rate has typically relied on a single method of synthetically manipulating rate of speech by compressing the vowel portion of a syllable. This does not mimic what occurs during natural speech production, and therefore could be influencing the perception of voicing contrasts. An experiment was conducted to address this problem by constructing a continuum of voice onset times for the velar place of articulation and then subsequently altering the rate of speech using three methods of manipulation: compressing the vowel, compressing the consonant and vowel proportionate to what occurs naturally, and compressing the total duration of the syllable. Each method of rate manipulation was evaluated at three speeds. The original continuum served as a control condition reflecting normal speed. Medium and fast versions of the continua were also presented, where utterances following the aspiration noise that specified voice onset time were .75 and .50 times the length of the original production at normal speed, respectively. Participants rated each stimulus on a scale of one to six (*1* = most /ga/-like - *6* =most /ka/-like). As expected, categorical functions were obtained across continua. There was an observed tendency for stimuli with short VOTs at faster speeds under consonant-vowel and total compression conditions to be rated/categorized as more /ka/-like relative to stimuli at the normal speaking rate. This pattern was not apparent for the vowel manipulation condition, which suggests that vowel compression is an appropriate method to manipulate the speaking rate of voicing contrasts. Total compression did produce consistent responses across speeds around the phonemic boundary and voiceless region of the continua, indicating that total compression could be appropriate for manipulating voiceless consonants with longer voice onset times. Taken collectively, the data seem to show that compressing the consonant

portion of a syllable, even to a small degree, can limit the perceptual information that is necessary for categorization.

Keywords: speech, rate, voice onset time, voicing, compression, categorical perception

Introduction

The effectiveness of computer-generated speech systems (e.g., text-to-speech systems) relies upon basic research in speech and language perception to identify critical parameters for manipulation and control. For example, basic research can identify average vowel and consonant durations in continuous speech at varying speeds (e.g., Kuwaraba, 1996). These kinds of observations could then be used as a model for creating synthetic vowels and consonants.

Real-world applications of computer-generated speech require an understanding of the perceptual consequences of contextual variables, such as speaking rate, in order to make synthetic speech as intelligible as possible. Basic research could improve artificial speech production for digital forms of text, such as eBooks, by modifying how the words are presented at faster rates so that it sounds as natural as possible. Visually impaired listeners rely upon these systems, and often choose to drastically speed up the rate of speech in order to retrieve information more quickly (Borodin, Bigham, Dausch, & Ramakrishnan, 2010). Developers of text-to-speech systems must ensure that their algorithm uses the most effective method of rate manipulation in order to maximize the utility of their product.

The effect that rate of speech has on phonemic perception has been investigated over several decades (e.g., Volaitis & Miller, 1992; Eimas & Miller, 1982; Miller, Aibel, & Green, 1984). For example, in 1979 Miller and Liberman observed that the duration of the syllable determined if the participant perceived /b/ or /w/. This is because syllable duration influences the transition duration, which is used to perceptually differentiate between the consonant and the semivowel. The boundary in which perception changed from /b/ to /w/ experienced the largest shift in location within shorter syllables, and then these shifts become progressively smaller as

the syllable duration increases. Temporal cues such as this provide crucial information for identification of consonants.

Consonants are typically perceived categorically, meaning that consonants are perceived in an all-or-none manner as either voiced (i.e., vocal cords vibrate) or voiceless (i.e., no vibration of vocal cords) with no changes in quality within either category (Liberman, Delattre, & Cooper, 1958). For each place of articulation (i.e., location within the mouth that is used to produce consonants) the switch from voiced (e.g., /g/) to voiceless (e.g., /k/) occurs at a point along the voiced-voiceless continuum. This point is called the phonemic boundary, and it is located where identification of the consonant is at 50%, representing a shift from one category to the other (Pastore, 1990).

Categorical perception in consonants has been previously revealed to depend critically upon the rate of changes in the spectrum (i.e., amplitudes as a function of frequency; see Hall & Peck, 2016). A recent example (i.e., Hall & Peck, 2017) found that slowing the second format frequency transition, which provides perceptual information for the consonant, resulted in perception of more than two categories in the middle of the continuum. It revealed two distinct boundaries that differentiate between diphthongs (i.e., double vowels) and glides/semivowels (i.e., /w/), rather than the single boundary that was observed in the control condition. These boundaries shifted in relation to the location of the single boundary within the control condition when syllable duration increased. These kinds of observations were possible because slowing the transition revealed acoustic details that could not be perceived at faster rates. In contrast, the short syllable condition did not negatively impact categorical perception.

Speaking rate has shown to influence the production of phonemes by affecting voice onset time (VOT; e.g., see Eimas & Miller, 1982). VOT is the duration of aspiration noise prior

to the vibration of the vocal folds, also known as voicing. The duration of VOT is a temporal cue that allows the listener to perceptually differentiate between voiced and voiceless consonants (Liberman, Delattre, & Cooper, 1958). The VOT at which categorization changes marks the location of the phonemic boundary. This phonemic boundary shifts as a function of speaking rate, with it increasing as speaking rate slows, and decreasing as speaking rate increases (Miller & Volaitis, 1989; Volaitis & Miller, 1992). Therefore, the range of stimuli that are perceived as providing the best examples for the phonemic category shifts along the temporal continuum as well (Miller & Volaitis, 1989).

Methods of Manipulating Speaking Rate

Speaking rate is typically synthetically manipulated for perceptual studies, and there are several different ways to accomplish this. For example, Volaitis and Miller (1992) varied speaking rate across a continuum by manipulating the duration of the subsequent vowel from short (125 *ms*) to long (325 *ms*), representing a fast and slow rate, respectively. This method of compressing the vowel portion is based on the observation that vowel duration is most impacted by changes in speaking rate (Kuwaraba, 1996). Vowel compression has also previously been shown to be sufficient for altering phoneme categorization (Miller & Liberman, 1979; Miller, Aibel, & Green, 1984).

Studies on speech production have also found that consonant duration changes as a function of rate of speech. For example, when participants were asked to talk twice as fast, /g/ and /k/ consonant durations reduced by approximately 5 *ms* and 12 *ms* respectively (Kuwaraba, 1996). A nonlinear method of rate manipulation compresses or truncates separate portions (i.e., vowel and consonant) of an utterance differently based upon observations that different portions of the signal change by different degrees. The portions are then mended together to create one

cohesive, shorter word that imitates a fast rate of speech (Olson & Berry, 1982). Compressing the consonant and vowel portions (CV compression) of the utterance is a method that most closely models what occurs naturally, and therefore could act as a more valuable method of rate manipulation for basic research on language and speech.

Total compression is a method of rate manipulation that is easily accomplished and has been used for several decades for word-level perceptual research and numerous computer-speech applications (Olson & Berry, 1982). This method involves simply time-compressing the entire phoneme by increasing the overall playback speed. Although this does not mimic what occurs naturally, it is still used in speech perception research because of its simplicity and efficiency (e.g., see Dilley & Pitt, 2010 for a recent example). Total compression also has been shown to be effective at maintaining intelligibility of synthetically generated words at extremely fast rates. In addition, humans report higher intelligibility of words that were time compressed samples of normal speech compared to words naturally produced at fast rates (Valentini-Botinhao, Toman, Pucher, Schabus, & Yamagishi, 2015).

While the results of Volaitis and Miller (1992) are useful for demonstrating an impact of speaking rate on phoneme perception, there are several remaining questions that need to be addressed if we are to gain a full understanding of rate's influence. For example, humans are capable of producing consonant sounds as short as 15 *ms* when asked to talk twice as fast (Kuwaraba, 1996). It is clear that the range of possible rates well exceeds those that were initially tested in the Voliatis and Miller (1992) study, which used a 45 *ms* consonant duration throughout all their continua. It is important to test for any perceptual effects that this difference may have on VOT and voicing contrasts. Therefore, the Volaitis and Miller (1992) study should be

expanded upon, and potential further impacts of fast rates of speech on phoneme perception should be evaluated.

Additionally, the potential perceptual effects on categorical perception have been examined for only one method of manipulating speaking rate, Vowel compression [i.e., in the Volaitis and Miller (1992) study]. Thus, it is critical to determine the effects of other, alternative manipulations of speaking rate. That way, future research in speech perception could use the most valuable method of rate manipulation.

Hypotheses

This research project sought to address both of these issues. Perception of voicing contrasts (i.e., phoneme categorization and perceived goodness) were examined for continua reflecting normal, medium, and fast speaking rate speeds based upon those that have been naturally observed. Medium and fast speeds were generated using three methods of rate manipulation: Vowel compression, CV compression, and Total compression, all of which were evaluated separately for categorization performance. This should support a determination of whether one or more of these methods is particularly appropriate to impact speaking rate while not adversely affecting categorization performance.

Although previous literature provides some assumptions for the effectiveness of each method of rate manipulation (e.g., Volaitis & Miller, 1992; Dilley & Pitt, 2010; Olson & Berrt, 1982), the research question remained open ended as to which method(s), if any, are most appropriate for categorical perception studies. We expected the phonemic boundary to shift and categorical perception to be affected as speaking rate increases, consistent with past observations (e.g., Volaitis & Miller, 1992). The Vowel compression method has shown to increase rate without harming the classification of the consonant for moderately fast rates (Volaitis & Miller,

1992). However, compressing the vowel to extreme rates could still have had perceptual consequences on the consonant. This could have resulted in Vowel compression and CV compression having poor perception of the phoneme that is being synthetically manipulated by reducing the vowel duration too much. Shortening the vowel could prevent access to all the necessary information needed to categorically identify the preceding consonant (Miller & Liberman, 1979).

It is therefore possible that listeners may have had greater difficulty in classifying consonants as they are shortened along with the shortened vowel (i.e., during CV compression). Alternatively, it is possible that perception of brief consonants could have necessarily become simplified to purely binary decisions (i.e., /g/ or /k/, with no additional differentiation for any subtle within category differences). Compressing the consonant to extreme degrees could influence categorization performance, resulting in less variation across ratings of goodness, as well as strictly categorical responses. If slowing the duration of the consonant had perceptual consequences by adding two more categories (Hall & Peck, 2017), then we assumed that shortening it would have had the opposite result by limiting necessary perceptual information and resulting in ratings on extreme ends of the scale.

Total compression reflects adjustment of the durations of vowels and consonants without respect for how each uniquely changes in natural speech production. However, it has shown to be more intelligible than other nonlinear methods on word-level perception (Valentini-Botinhao et al., 2015), and therefore could be sufficient for phoneme-level perception. In contrast, it could produce the most adverse effects of the three methods as it does not imitate what occurs during speech production.

Method

Participants

We recruited JMU undergraduate students through the JMU Department of Psychology's participant pool, where students from introductory psychology courses participate in research in exchange for partial fulfillment of course requirements. Counterbalancing the three orders of speaking rate-related variables required us to recruit at least 18 participants. Data analyses were restricted to participants who were between the ages of 18 and 40 years old. This minimized any potential adverse perceptual consequences associated with high-frequency hearing loss as a result of aging (i.e., presbycusis). Additionally, all participants self-reported normal hearing capabilities. The participants were native speakers of American English, since the stimulus sample originated from a recording of an American English speaker. Furthermore, the use of an a priori performance criterion of 70% correct identification of the endpoints on the continuum at a normal speaking rate ensured that participants could reliably identify the phonemes.

Stimuli

Large differences in consonant duration between normal and fast speeds have been observed in the voicing contrasts for a velar place of articulation (/g/ and /k/) relative to other places of articulation, as have large differences in duration across varying rates for the vowel /a/ compared to other vowels (Kuwaraba, 1996). We therefore relied upon /ga/-/ka/ continua for this study. After all, it is these particular consonants and vowel that provide the widest of range of natural variability in speech production. As a result, they should have been most capable of revealing potential effects on perception due to a manipulation of speaking rate.

Continua were constructed from initial, digital recordings of a native English speaker producing /ka/ syllables at a normal speaking rate with a 44.1 kHz sampling rate (16-bit

resolution) using a dynamic microphone. We used a male speaker, which is common for speech recordings due to less variability in the frequency domain (Klatt & Klatt, 1990). We asked that the speaker produce multiple utterances of /ka/, from which the best instance was chosen. The utterance reflected the greatest clarity/comprehensibility while reflecting what was perceived to be a typical speaking rate (defined by our laboratory staff).

The consonant in the selected /ka/ production spoken at a normal speaking rate was edited to generate continua where VOT ranged from 1 to 61 *ms* , with several unequal increments in the beginning of the continua (1 *ms*, 6 *ms*, 11 *ms*) followed by increments of 10 *ms*, for a total of eight VOT values within each continuum. This was accomplished using Adobe *Audition* (version CS6) to edit the /ka/ sample by truncating or compressing the aspiration noise at the beginning of the consonant in step-wise fashion until VOT eventually reached 1 *ms*. The aspiration noise was identified by analyzing the spectrogram within the audio editing software.

The different methods of speaking rate manipulation were also implemented using Adobe *Audition*. The normal speed continuum remained unchanged as a baseline across all three rate manipulation methods. The Vowel compression manipulations were made by adjusting the duration of the vowel from its normal length (328 *ms*) to three-fourths for the medium speed (246 *ms*) and half for the fast speed (164 *ms*). The same was done for Total compression by truncating the normal length of the combined consonant and vowel (363 *ms*) to three-fourths and half of its duration. CV compression was accomplished by proportionally adjusting the consonant and vowel durations by following observations of naturally produced speech at varying speeds (e.g., see Kuwaraba, 1996). This study identified the average percentages for the consonant and vowel lengths for the velar place of articulation as well as the vowel /a/, and those percentages were used to differentially adjust the consonant and vowel proportions. At the time

of presentation, all stimuli were presented to listeners in a single-walled sound-attenuated chamber over Sennheiser HD 25-SP II headphones, with peak intensity not exceeding 80 dB[A].

Procedure

The participants were asked to classify each presented stimulus on a trial as accurately as possible (as /ga/ or /ka/). Each trial consisted of a randomly presented stimulus from the /ga/-/ka/ continua and the participants' responses were recorded. This action was self-paced, with a 500 *ms* inter-trial interval. On each trial the participants judged their perception of each stimulus by rating them on a six-point Likert scale with /ga/ and /ka/ on opposite ends of the scale (i.e. 1 = most "/ga/ like", 6 = most "/ka/ like").

Within each continuum, each of the eight stimuli (i.e., the eight changes in VOT values) were presented in random order 10 times. There were nine continua (3 rates x 3 methods of manipulating rate), with each continuum being evaluated within a separate block of trials. Using a Latin square design, the conditions were organized by presenting the three speeds of speaking rate one after another for each separate speaking rate manipulation. There was a total of 720 trials across blocks, which were completed within a single session of approximately 27 minutes. Rest breaks were provided in between each block of trials.

Since VOT continua are perceived categorically, phoneme categorization performance should be able to determine whether the phonemic boundary between /g/ and /k/ changes as a function of speaking rate. Previous studies of categorical perception have benefited from using a rating task, in which the participant has the option to identify changes in consonant quality across the continuum (e.g., see Conway & Haggard, 1971; Haggard, Summerfield, & Roberts, 1981; also see Hall & Peck, 2016, for a recent implementation of such a rating task). A rating scale with the two endpoints representing a true /ga/ or true /ka/ allowed the researchers to observe

phoneme identification (i.e., one through three as /ga/ and four through six as /ka/) as well as any subtle within-category differences.

Results

Data Analysis

The data were analyzed in two ways: ratings and binary responses. As previously stated, this choice was made in order to identify any shifts in categorization (using the binary data) as well as detecting any variability within these categories (using rating data). The /ga/-/ka/ ratings were translated into binary responses (i.e., /ga/ like response = 0, /ka/ like response = 1) by grouping responses on the left side of the continuum (1-3) as /ga/ and the right side (4-6) as /ka/. The aggregate binary data for the 10 instances of each stimulus represented a probability the participant would perceive /ka/. This allowed us to analyze the perceived quality of the stimuli but also categorization performance. The average ratings and average probabilities for each VOT were determined separately for each combination of speaking rate and rate manipulation.

Both the ratings and the binary data were separately analyzed using three corresponding 3x8 repeated measures ANOVAs with speed of speaking rate (normal, medium, fast) and VOT (1-61 ms in 8 steps) as factors. This method of data analysis was preferred, rather than running a three-way ANOVA, due to the fact that the normal continuum was presented three times so that it could be used to compare against the speeds of speaking rate within each method of manipulation. All post hoc pairwise comparisons of means were accomplished using a Bonforonni statistic. A Greenhouse Geiser was used in any case when the assumption of sphericity was violated.

The categorization probabilities also allowed us to find the true location of each participants' phonemic boundary for any given continuum through linear interpolation. The

effects of speaking rate on these determined boundary locations were analyzed separately for each method of speaking rate manipulation. This was accomplished using one-way repeated measures ANOVAs with speed of speaking rate as the factor. Results are summarized separately below for each method of rate manipulation, as well as for individual determinations of phonemic boundaries from the categorization data.

Consonant-Vowel Compression

Figure 1 displays the average ratings (left) and categorization performance (right) along with corresponding standard errors for each speaking rate. Each panel in the figure summarizes the data for a different method of speaking rate manipulation. Results from the CV manipulation are displayed in panel A.

As you would expect from VOT continua, rating performance (left side of panel A) was largely impacted by VOT. Specifically, there was a sharp shift in categorization near the middle of the continua. This tendency resulted in an observance of a main effect of VOT, $F(2.565, 43.613) = 354.860$, $p<.001$, $\eta^2 = .954$. Subsequent pairwise comparisons revealed that ratings became significantly more /ka/ like with increasing VOT for the following comparisons near the average phonemic boundaries displayed in the figure: 11 *ms* vs 21 *ms*, 21 vs 31, 31 vs 41, $p<.001$.

In contrast, speed did not have an overall impact on average ratings (see Figure 1, panel A, left). This tendency was confirmed by the lack of a main effect of speed, $F(1.438, 24.452) = 3.066$, $p=.079$, $\eta^2=.153$. However, this does not mean speed had no influence on performance. For example, in Figure 1 categorization as /ga/ for the lower VOT values was not as strong for fast and medium speeds, and it appears that there was a sharper shift in categorization at the normal speaking rate. This pattern of data contributed to a significant interaction between speed

and VOT, $F(4.232, 31.726) = 8.906$, $p<.001$, $\eta^2 = .344$. Subsequent pairwise comparisons revealed that stimuli at normal speeds received lower ratings relative to those at medium and fast speeds at both 1 *ms* and 6 *ms* VOTs, $p \leq .002$. There also was a significant shift towards more /ka/ like responses for stimuli with an 11 *ms* VOT presented at a fast rate compared to a medium rate, $p=.028$. Near the average boundary location at 21 *ms* VOT, the normal and fast speeds produced sharper shifts toward /ka/ responses compared to the medium speed, $p \leq .029$. Likewise, the normal speed resulted in more /ka/ like ratings for stimuli with 41 *ms* VOT relative to fast rates of speech, $p=.015$. A 61 *ms* VOT was rated as more /ka/ like at normal rates of speech compared to medium and fast rates of speech, $p \leq .001$. Thus, it appears that increases in speaking rate resulted in less steep functions.

Even though the slope of the functions from categorization performance (right hand side of panel A) were steeper than those obtained from the original ratings, the pattern of results was very similar. As expected there were sharp changes in the categorization performance as a function of VOT. This was confirmed by a main effect of VOT on categorization performance, $F(2.208, 37.541) = 436.917$, $p<.001$, $\eta^2 = .963$. Pairwise comparisons revealed increased probability of /ka/ responses as VOT increased for each of the following pairs: 1 *ms* vs 6 *ms*, 11 vs 21, 21 vs 31, 31 vs 41, $p \leq .006$.

Furthermore, there was not a difference in categorization performance as a function of speed, as indicated by an absence of a main effect, $F<1$. However, just as with the original ratings, categorization performance at certain VOTs was impacted by speed. This can be seen in panel A as a sharper shift in categorization at normal speed. This contributed to a significant interaction between speed and VOT, $F(4.069, 69.173) = 4.285$, $p=.004$, $\eta^2 = .20$. Pairwise comparisons further revealed that this effect was attributable to the fact that stimuli with a 6 *ms*

VOT were significantly more likely to be categorized as /ga/ at normal speed compared to fast speed, p=.016, and marginally more likely relative to medium speed, *p*=.052. Furthermore, stimuli near the average phonemic boundary at 21 *ms* VOT were more likely to be perceived as /ka/ at normal speeds relative to medium or fast speeds, $p \leq .018$.

Vowel Compression

Panel B in Figure 1 summarizes ratings and categorization performance for the Vowel manipulation of speaking rate. As with the CV manipulation, average ratings from the Vowel manipulation increased sharply as VOT increased to approach the average boundary locations toward the middle of the continua. This resulted in a main effect of VOT on rating performance, $F(1.833, 31.168) = 553.590$, $p<.001$, $\eta^2 = .970$. Pairwise comparisons confirmed statistically significant changes in ratings (i.e., more /ka/-like as VOT increased) for the following VOT pairings: 6 *ms* vs. 11 *ms*, 11 vs. 21, 21 vs. 31, 31 vs. 41, 41 vs. 51, $p \leq .001$.

In panel B of Figure 1 there is a noticeable shift in the slope of the average ratings function between medium and normal speeds. Medium speeds seemed to result in a shallower slope compared to normal speeds. This was confirmed with a significant main effect of speed, $F(2, 34) = 4.591$, $p=.017$, $\eta^2 = .213$. Pairwise comparisons further revealed a marginal tendency for stimuli at normal speeds to be rated as less /ka/ like than at the medium speeds, *p*=.06. More importantly, the shift in ratings due to speed depended upon VOT, as indicated by a significant interaction between speed and VOT, $F(5.051, 85.871) = 7.988$, $p<.001$, $\eta^2 = .320$. Further pairwise comparisons revealed that this shift was restricted to VOT values near the boundary location in the average functions, with increasing ratings at medium speeds relative to normal at 11 *ms* VOT, fast relative to other speeds at 21 *ms*, and fast relative to the other speeds, as well as medium relative to normal speed, at 31 *ms*, $p \leq .011$.

As expected, categorization resembled rating performance, despite a sharper shift in the slope of categorization functions towards the middle of the continua (see panel B of Figure 1). VOT had a large impact on categorization tendencies, and followed suit with the rating results by consequentially having a significant main effect for VOT, $F(2.491, 42.352) = 701.038$, $p<.001$, $\eta^2 = .976$. This was supported by pairwise comparisons revealing a tendency to categorize higher VOTs as /ka/ in the following values: 11 *ms* vs 21 *ms*, 21 vs 31, 31 vs 41, $p \leq .001$.

In contrast to the ratings results, overall levels of categorization performance were not significantly impacted by speed, $F(2, 34) = 1.076$, $p=.352$, η^2 .069. However, categorization did change as a function of speed for VOT values near the middle of the continua, contributing to a significant interaction between speed and VOT, $F(3.733, 63.458) = 4.452$, $p=.004$, $\eta^2 = .208$. Pairwise comparisons revealed a statistically significant shift toward more /ka/ categorizations for stimuli at medium and fast speeds at the 31 *ms* VOT, $p \leq .035$.

Total Compression

In comparison to the other rate manipulations, time compression of the entire signal (i.e., the Total rate manipulation) seemed to produce average rating and categorization functions with slightly steeper slopes (see panel C of Figure 1). For example, the average rating data (left side of panel C) reflects a noticeable change in performance from the middle of the VOT continua and upward. This tendency contributed to a main effect of VOT, $F(2.291, 38.95) = 433.465$, $p<.001$, $\eta^2 = .962$. Pairwise comparisons confirmed that there was an increase in /ka/-like ratings with increasing VOTs for the following adjacent pairs: 11 vs. 21, 21 vs. 31, 31 vs. 41, and 51 vs. 61 *ms*, $p \leq .023$.

Overall levels of rating performance in the Total compression condition did not change as a function of speaking rate alone, as indicated by the absence of a main effect of speed, $F(2, 34)$

= 2.561, p=.092, η^2 = .131. Speed did, however, impact ratings at certain VOTs. As can be seen in the left side of panel C, different average ratings across speeds were observed in the middle and at both ends of the continua. Such differences contributed to a significant interaction between speed and VOT, F(3.867, 65.74)=6.131, p<.001, η^2 =.265. Subsequent pairwise comparisons indicated that stimuli were rated higher (i.e., more /ka/ like) at fast speeds relative to normal speeds at the 1 *ms* VOT, and relative to all other speeds at the 6 *ms* VOT, $p\leq$.034. Furthermore, stimuli at the medium speeds were rated higher relative to all other speeds at the 31 *ms* VOT, stimuli at normal and medium speeds were rated higher at the 51 *ms* VOT, and stimuli at normal speeds were rated higher that all other speeds at the 61 *ms* VOT, $p\leq$.034.

The slopes of the average categorization functions for the Total compression conditions (displayed on the right side of panel C) were uniformly steep across speeds around the middle VOT values. These sharp shifts in categorization contributed to a significant main effect, F(1.887, 32.075) = 610.793, p<.001, η^2 =.973, and pairwise comparisons of means confirmed more /ka/ responses with increasing VOT for the following adjacent points along the continua: 11 vs. 21, 21 vs. 31, and 31 vs. 41 *ms*, $p\leq$.011. The nature of the obtained categorization functions did not significantly depend upon speed. Consistent with this assertion, neither a main effect of speed, F<1, nor an interaction between speed and VOT, F(3.904, 66.371) = 1.649, p=.174, η^2 =.088, were obtained.

Individual Phonemic Boundary Locations

Mean individual phonemic boundary locations, along with their corresponding standard errors, are summarized in Figure 2 for each combination of speed and speaking rate manipulation. As can be seen in the figure, boundary location did not appear to shift either a function of speaking rate within a given type of rate manipulation, or across different types of

rate manipulation. The lack of an effect of speaking rate within each method was confirmed by the absence of main effects of speed from the ANOVAs (for all methods of rate manipulation, $F<1$).

Discussion

Summary

Minimally, listeners in the current investigation generally appeared to perceive the continua in the CV manipulation more uniformly at normal speeds than at faster speeds. For example, participants were more consistent in rating lower VOT values as /ga/ like for the normal speeds compared to the medium and fast speeds. This could indicate that categorization became more difficult when both the consonant's formant transitions and the noise specifying the VOT are compressed to extreme degrees. Anecdotal evidence from the participants gave some indications that perception under such circumstances changed to different consonants, such as /va/ or /da/. This could explain why the ratings had a shallower slope at faster speeds given that no response categories were provided that matched these alternatives. It appears that limiting preceding information about VOT ahead of an already truncated set of consonantal transitions might have resulted in poor perception of the voiced consonant (/g/), which made it harder to categorize.

Similarly, the slopes of the functions for Vowel compression became shallower as speed increased. Stimuli with 31 *ms* VOTs were rated as more /ka/ like at faster speeds. The Vowel compression method seemed to have not had the same influence on short VOT values due to the fact that it only manipulates the steady-state portion of the syllable. Vowel formant frequencies do not change rapidly, so the vowel can afford to be compressed without drastically altering spectral information. This should allow for easier identification of /ka/ and /ga/. In addition,

ratings began to decrease around the boundary location as speaking rate increased. This, combined with the fact that the stimuli at normal speed were rated as less /ka/-like right after the boundary, contributed to the obtained interaction between speed and the nature of the speaking rate manipulation. However, unlike the CV compression manipulation, performance did not widely differ across speaking rates at lower VOT values. Rather, participants rated such stimuli consistently as /ga/-like. This makes sense in comparison to CV compression since Vowel compression did not interfere with any preceding information for the consonant's formant transitions.

Unlike the other speaking rate manipulations, there was not an interaction between speaking rate and VOT on categorization performance for Total compression. The slopes of the obtained functions (see panel C of Figure 1) are nearly identical around the middle of the continua, showing that speed did not impact boundary locations. However, the rating data did demonstrate slightly more shallow functions at faster speeds. This is more obviously seen in the voiced end of the continuum (i.e., /ga/), where the lower VOT values may have been less easily or less accurately categorized. This resembles the results of the CV compression, perhaps because both methods influence the consonant portion of the syllable. Some participants also mentioned that they perceived some consonants that were not represented on the response scale, such as /v/, /p/, and /b/, at faster speaking rates. Under such circumstances, some participants reported using a common strategy in which they rated syllables in the middle of the scale if they did not seem to hear a distinct /ka/ or /ga/, which would account for the observed elevated. average ratings for the /ga/ end of the continua at faster speeds. This could reflect an influence of more limited information being available about the consonant. After all, it is known that syllable length affects transition duration, which determines the perceptual differentiation between the

consonant and semivowel (Miller & Liberman, 1979). Conversely, previous studies have also reported the perception of alternate phonemes when consonantal formant transitions are significantly lengthened (e.g., see Hall & Peck, 2017).

Interpretation

So what do the obtained patterns of performance suggest about the utility of the three manipulations of speaking rate? The rating data suggests there is a difference in the ability to categorize the voiced consonant at varying speeds for Total and CV compression, but not Vowel compression. The Vowel compression manipulation preserved the voiced end of the continuum, as indicated by the fact that it was arguably easier to identify lower VOT values as /ga/. After all, the consonant's length at normal speed was only 35 *ms* long. Even just a few *ms* worth of information could make a substantial impact on the perception of the consonant. Therefore, truncating the consonant as well as trimming the VOT to minimal values (under Total and CV conditions) could adversely impact the perception of voiced material.

The boundary locations across the speaking rate manipulations did not shift as a result of the speed of speaking rate. This was not hypothesized, as we expected category boundaries to shift as speaking rate changed, similar to what has been seen in the literature (e.g., see Eimas & Miller, 1981; also see Volaitis & Miller, 1992). Although some individual participants' average boundary locations did decrease by as much as 5-7 *ms*, in VOT as speed increased, the majority of listeners did not experience a significant shift, or even experienced a slight shift in the opposite direction. Volaitis & Miller's (1992) study used a synthetic VOT continuum ranging from 10-120 *ms* for their short syllable (125 *ms* long) and 15-320 *ms* for their long syllable (325 *ms* long). It is unclear if the ratio of VOT length and subsequent syllable length ever occurs in nature, with VOT nearly matching the duration of the rest of the syllables at the end of the VOT

continua. In addition, the consonant duration that Voliatis and Miller used was 45 *ms*. We know, however, that humans are capable of producing consonants at much faster rates than that (e.g., see Kuwaraba, 1996). For this reason, the stimuli for the current investigation were instead created by adjusting the duration of a resynthesized, naturally spoken syllable using values that were within natural ranges of variation. This resulted in the consonant portion of the syllable to be approximately 35 *ms* at normal speed and only 15 *ms* at the fastest speed. In contrast to Volaitis and Miller, the average perceived boundary location in this study was already short to begin with at normal speeds. Thus, not only is there a difference in the naturally spoken versus synthesized duration of consonants, but it seems that the shorter consonants in this study leave little room for the boundary location to move.

The Total and Vowel compression methods could be argued to be the most appropriate when adjusting certain phonemes (at least those defined by voicing contrasts) to fast speaking rates. Total compression could be used as long as there is sufficient information remaining in the consonant to permit identification. This is fortunate since total compression is by far the easiest method of rate manipulation, and can be easily implemented by anyone who has audio editing software. Similarly, the Vowel compression is an appropriate method as well depending on the consonant in the syllable. These two manipulations may only be useful when applied to stop consonants with longer VOTs so that enough perceptual information is preserved in order to properly identify the consonant. It also is noteworthy that in some languages this method may not work, since certain consonants experience pre-voicing (i.e., negative VOT values, such as in Spanish; e.g., see Magloire & Green, 1999).

Limitations

One limitation of the current investigation that may further complicate a clear assessment of methods of rate manipulation pertains to the aforementioned reporting of perceptual changes to consonants into phonemes that were not included on the response scale (for the Total and CV compression manipulations). We did not provide instructions to direct listeners how to assign response categories to these stimuli if they heard consonants other than /ga/ or /ka/ (e.g., to use the voiced /g/ in instances when /b/, another voiced consonant was perceived). This could easily have impacted the data, especially for the lower VOT values towards what was supposed to be perceived as /ga/. In the future, such a possibility could be minimized by the inclusion of instructions for how to respond in these situations, or alternatively, by having multiple scales in order to separately respond to potentially new voiced-voiceless distinctions that could appear in certain conditions.

Implications and Future Research

Ultimately, all methods demonstrated reasonably sharp categorization and corresponding phonemic boundaries, and involved vowel compression. This suggests that vowel compression is sufficient as a manipulation of speaking rate, at least for this voicing contrast. This study provides some implications that Total compression could also be used in certain instances. Voiced stop consonants could cause problems with a Total compression manipulation if they are presented at high rates of speed due to the low VOT values. The duration of VOT in voiced stop consonants is already so short that compressing the entire syllable could limit too much perceptual information needed to correctly identify the consonant. However, some stop consonants with longer VOTs could be fine if manipulated using these methods since the data show consistent categorization tendencies towards the voiceless end of the continua.

Total compression is most often used in everyday products, mostly for text-to-speech systems such as eBooks or even voice automated personal assistants (e.g., Apple's Siri or Amazon's Alexa). It is fortunate that Total compression seems to work for certain consonants, as it is extremely easy to implement. Based upon the major findings from the project reported here, any individual or any company could feel justified to implement this method to increase speaking rate to high speeds without adversely affecting the perception of voicing contrasts. However, we should exercise caution to be sure not to generalize beyond the reported findings. Further research involving various other consonants and CV syllables is first required to determine whether these findings and interpretations extend to other contrasts. Only then will we have a more complete understanding of which manipulation is most effective in conveying an increase in speaking rate while posing a minimal threat to categorization.

References

Adobe Audition CS (Version 6) [Computer software]. (2013) Retrieved from http://www.adobe.come/products/audition

Borodin, Y., Bigham, J. P., Dausch, G., & Ramakrishnan, I. V. (2010). More than meets the eye. *Proceedings of the 2010 International Cross Disciplinary Conference on Web Accessibility (W4A) - W4A 10*. doi:10.1145/1805986.1806005.

Conway, D. A., & Haggard, M. P.(1971).New demonstrations of categorical perception. *Speech Synthesis and Perception* (Progress Report No. 5), 51–73. Cambridge: University of Cambridge, Psychology Laboratory.

Dilley, L. C., & Pitt, M. A. (2010). Altering context speech rate can cause words to appear or disappear. *Psychological Science,21*(11), 1664-1670. doi:10.1177/0956797610384743

Eimas, P. D., & Miller, J. L. (1981). *Perspectives on the study of speech*. Hillsdale, N.J.: Erlbaum.

Haggard, M. P., Summerfield, Q., & Roberts, M. (1981). Psychoacoustical and cultural determinants of phoneme boundaries: Evidence from trading F0 cues in voiced voiceless distinction. *Journal of Phonetics*, 9 (1), 49–62. doi: 1982-00897-001

Hall, M. D., & Peck, R. B. (2016). Categorical perception: effects of the extent and rate of spectral change. *Journal of Cognitive Psychology*, 29(1), 3-22. doi:10.1080/20445911.2016.1229704.

Klatt, D. H., & Klatt, L. C. (1990). Analysis, synthesis, and perception of voice quality variations among female and male talkers. *The Journal of the Acoustical Society of America, 87*(2), 820-857. doi:10.1121/1.398894.

Kuwabara, H. (1996). Acoustic properties of phonemes in continuous speech for different speaking rate. *Proceeding of Fourth International Conference on Spoken Language Processing. ICSLP 96*. doi:10.1109/icslp.1996.60730.

Liberman, A. M., Delattre, P. C., & Cooper, F. S. (1958). Some cues for the distinction between voiced and voiceless stops in initial position. *Language and Speech,1*(3), 153-167. doi:10.1177/00238309580010030.

Magloire, J., & Green, K. P. (1999). A Cross-Language Comparison of Speaking Rate Effects on the Production of Voice Onset Time in English and Spanish. *Phonetica,56*(3-4), 158-185. doi:10.1159/000028449.

Miller, J. L., Aibel, I. L., & Green, K. (1984). On the nature of rate-dependent processing during phonetic perception. *Perception & Psychophysics*, *35*(1), 5-15. doi:10.3758/bf03205919

Miller, J. L., & Liberman, A. M. (1979). Some effects of later-occurring information on the perception of stop consonant and semivowel. *Perception & Psychophysics*, 25, 457–465. doi:10.3758/BF0321382.

Miller, J. L., & Volaitis, L. E. (1989). Effect of speaking rate on the perceptual structure of a phonetic category. *Perception & Psychophysics, 46*(6), 505-512. doi:10.3758/bf03208147.

Olson, J. S., & Berry, L. H. (1982). *The State of the Art in Rate-Modified Speech: A Review of Contemporary Research.*

Pastore, R. E. (1990). Categorical perception: Some psychophysical models. In S. Harnad (Author), *Categorical Perception: The Groundwork of Cognition* (pp. 29-52). Cambridge: Cambridge University Press.

Valentini-Botinhao, C., Toman, M., Pucher, M., Schabus, D., & Yamagishi, J. (2015). Intelligibility of time-compressed synthetic speech: Compression method and speaking style. *Speech Communication*, *74*, 52-64.

Volaitis, L. E., & Miller, J. L. (1992). Phonetic prototypes: Influence of place of articulation and speaking rate on the internal structure of voicing categories. *The Journal of the Acoustical Society of America*, *92*(2), 723-735. doi:10.1121/1.403997

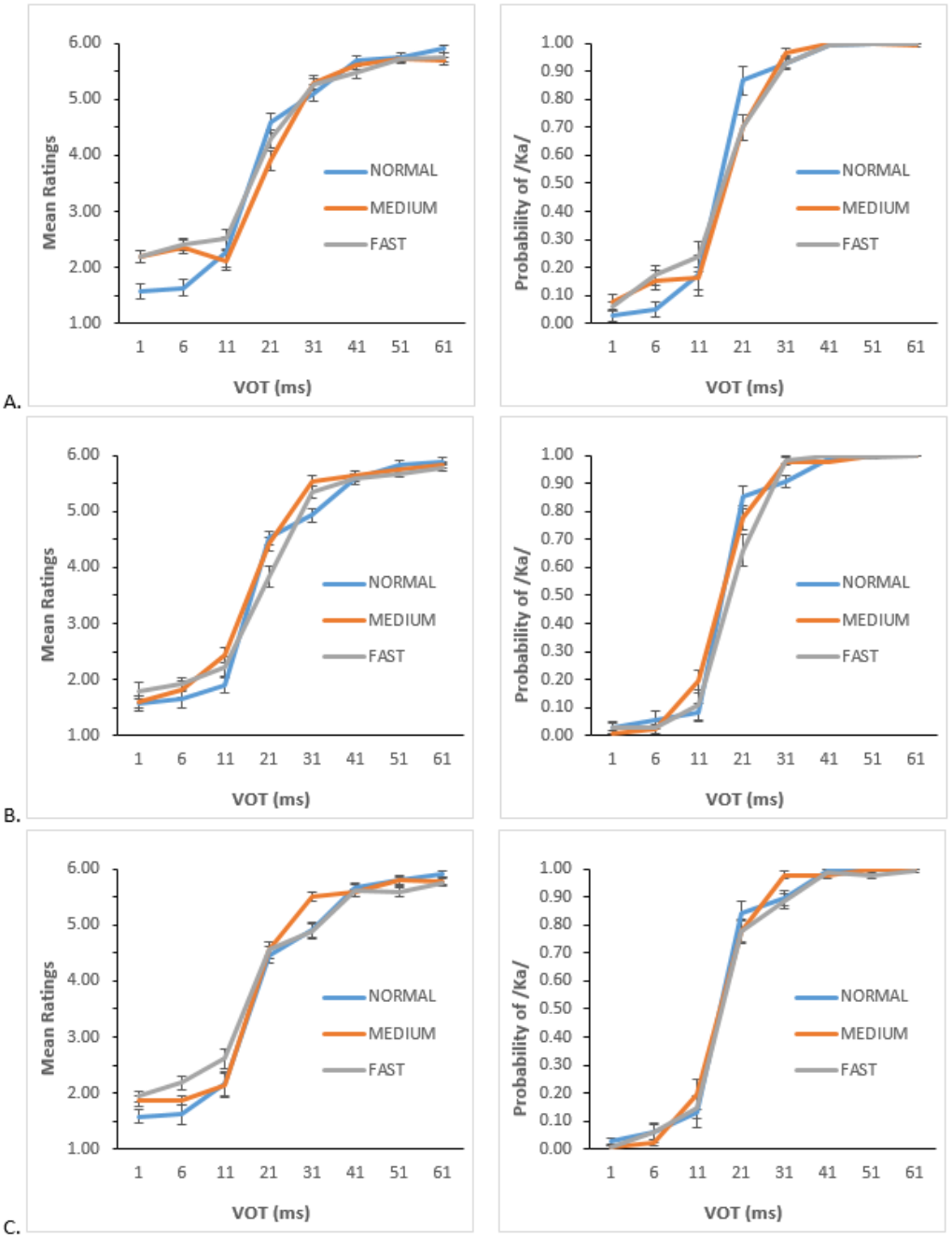

Figure 1. Mean rating performance and categorization performance with corresponding error bars for the three methods of speaking rate manipulation, including CV compression (panel A), Vowel compression (panel B), and Total compression, (panel C). Summaries of rating performance are displayed to the left, and summaries of categorization performance are displayed to the right.

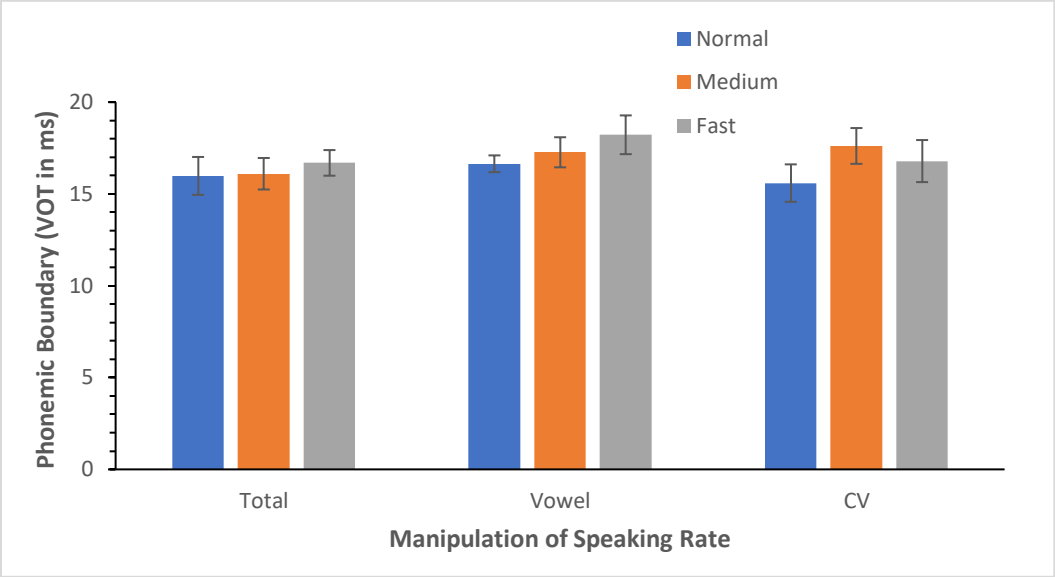

Figure 2. Mean boundary locations with corresponding error bars for each speed and method of speaking rate manipulation.