Distribution Agreement

In presenting this thesis as a partial fulfillment of the requirements for a degree from Emory University, I hereby grant to Emory University and its agents the non-exclusive license to archive, make accessible, and display my thesis in whole or in part in all forms of media, now or hereafter now, including display on the World Wide Web. I understand that I may select some access restrictions as part of the online submission of this thesis. I retain all ownership rights to the copyright of the thesis. I also retain the right to use in future works (such as articles or books) all or part of this thesis.

Veronica Vazquez Olivieri March 27, 2019

Effects of Primed Speaker Knowledge on Speech Perception

By

Veronica Vazquez Olivieri

Lynne C. Nygaard, Ph.D.
Adviser

Psychology Department

Lynne C. Nygaard, Ph.D.
Adviser

Hillary R. Rodman, Ph.D.
Committee Member

Donald Tuten, Ph.D.
Committee Member

2019

Effects of Primed Speaker Knowledge on Speech Perception

By

Veronica Vazquez Olivieri

Lynne C. Nygaard, Ph.D.
Adviser

An abstract of
a thesis submitted to the Faculty of Emory College of Arts and Sciences
of Emory University in partial fulfillment
of the requirements of the degree of
Bachelor of Arts with Honors

Psychology Department

2019

Abstract

Effects of Primed Speaker Knowledge on Speech Perception

By Veronica Vazquez Olivieri

Human speech changes significantly depending on the context in which it is produced. Variation in speakers' utterances is often systematic and stems from a variety of sources such as a speaker's age, gender, regional origin, and individual identity. This variability is not discarded, but rather rapidly encoded and integrated with linguistic structure during spoken language processing. In two experiments, the present study examines the extent to which social cues about the speaker, such as speakers' age, can shift the perception of the acoustic-phonetic features of speech (e.g., voice-onset time, VOT). Using vowel length as a proxy for speaking rate, the first experiment presented participants with either shortened or lengthened vowel lengths in VOT continua during a categorization task. Findings from the first experiment showed that categorization of speech sounds can be changed by altering the acoustic characteristics of the signal.

The second experiment investigated the extent to which knowledge of a speaker's age can shift the perception of VOT. Listeners engaged in a task identical to that in Experiment 1 except that they were primed with visual information about the speaker's age (a photograph of an old or young speaker). Results show different categorization performance as a function of age prime for tokens across continua differing in VOT. Listeners appeared to be tracking and encoding the distribution of features associated with a particular group of speakers rather than just inferring general characteristics of speech associated with that group of speakers. These findings demonstrate that priming listeners with knowledge of a speaker's age can induce similar patterns to those seen when the speech signal is altered. Speech perception appears to entail both

bottom-up and top-down processes such that listeners employ their knowledge of group-specific acoustic-phonetic features of speakers to adjust their perception of speech.

Keywords: VOT, speaker Age, Vowel Length, Priming

Effects of Primed Speaker Knowledge on Speech Perception

By

Veronica Vazquez Olivieri

Lynne C. Nygaard, Ph.D.
Adviser

A thesis submitted to the Faculty of Emory College of Arts and Sciences
of Emory University in partial fulfillment
of the requirements of the degree of
Bachelor of Arts with Honors

Psychology Department

2019

Acknowledgements

I would like to thank Dr. Lynne Nygaard and Dr. Christina Tzeng for their invaluable mentorship and guidance throughout the year. I will be forever grateful to have had the opportunity to work with such admirable role models. I would also like to extend my gratitude to Emory's Undergraduate Research Program (URP) for funding my research and to the Speech and Language Perception Lab for their thoughtful insights.

Table of Contents

Effects of Primed Speaker Knowledge on Speech Perception

Speech is highly variable. As listeners, we must contend with the variation stemming from speakers with different individual vocal characteristics (Mitchel, Gerfen, & Weiss, 2016; Nygaard, Sommers, & Pisoni, 1994), from different regions (Hay & Drager, 2010; Labov, 1972, Niedzielski, 1999; Peterson & Barney, 1952), ages (Kleinschmidt, Weatherholtz, & Jaeger, 2018; Walker & Hay 2011), and genders (Johnson, Strand, & D'Imperio, 1999; Kleinschmidt, 2019) to understand the speech signal. In addition to variation associated with the speech of individual speakers, there is also systematic variation associated with groups of speakers such that listeners use different speaking styles to distinguish and identify one group of speakers from another. For example, a natural source of this variation is observed in the accentedness of speech by non-native speakers of a particular language (e.g., Sidaras, Alexander, & Nygaard, 2009). Thus, speakers are thought to adopt patterns of pronunciation that serve as markers of their individual identity or group membership (Johnson, Strand, & D'Imperio, 1999). In turn, listeners are sensitive to the variation in these patterns of pronunciation and may use this sensitivity to facilitate their perception of spoken language. The current study examines this sensitivity by exploring the extent to which social expectations related to a speaker's identity, namely speaker age, can influence the perception of spoken utterances.

There are two broad theoretical frameworks in the study of spoken language perception that approach the role of variability in different ways. One framework views variability as a problem and assumes that it must be ignored, overcome, or normalized to achieve understanding (e.g., Liberman, Cooper, Shankweiler, & Studdert-Kennedy, 1967; Miller & Volaitis, 1989; Pisoni, 1997). According to this perspective, during the perception of speech, the signal must be normalized and reduced to the invariant cues of the speech code, thereby removing variance

stemming from speaker differences in pronunciation, in order to understand the spoken message. According to these normalization approaches, listeners separate the linguistic units (e.g., phonemes, words) from information that is thought to be phonetically irrelevant for comprehension, such as cues to speaker identity (Heald & Nusbaum, 2014; Summerfield & Haggard, 1973). Ultimately, this tradition views the perception of social cues related to the speaker as being separate from the perception of the linguistic content of speech.

The second view considers that variability is a rich source of social information that directly affects how listeners understand spoken language (e.g., Goldinger, 1996). In this view, variation is not discarded but rather retained in memory, with representations of this variation activated during subsequent speech perception. This idea dates back to Firth (1964) who proposed that the patterns of speech production provide not only cues to linguistic structure in the speech signal, but also to social cues that convey socio-cultural and personal information about the speaker. Under Firth's assertion, individual differences in how speech is produced are central to how language works in society. For example, the fact that a speaker has a particular regional accent is not noise or something to be discarded, but rather serves a social function. More recent literature suggests that characteristics of the speech signal that provide information about speaker identity or group membership are used to make social inferences about the speaker and gauge the communicative setting (Kleinschmidt et al. 2018; Nygaard, 2005). Taken together, this framework proposes that variation in the speech signal is an important source of social information that is retained and used during the perception of spoken language.

Encoding and processing of indexical cues in speech

These views of speech perception suggest that the speech signal is highly complex

because it carries information not only about what is being said, but also how it is being said and by whom. Specifically, many have posited the idea that voice recognition and processing of linguistic content are not independent processes (Nygaard et al., 1994; Palmeri, Goldinger, & Pisoni, 1993). Talker-specific variability in surface characteristics is encoded during speech processing and is used early on in linguistic processing to interpret the meaning of particular words. For example, to examine the effects of speaker familiarity, Nygaard et al. (1994) trained two groups of subjects to recognize a set of voices over a nine-day period. At test, the participants transcribed novel words spoken either by the same speakers as heard during training or a new set of speakers. Results showed that test transcription was more accurate for utterances spoken by familiar rather than unfamiliar speakers. Thus, familiarity with speakers' voices was found to facilitate linguistic processing of novel words. These findings suggest that experience with and knowledge of specific speakers and their individual vocal characteristics are encoded and facilitate the processing of linguistic content.

Not only are surface characteristics encoded and used by listeners to facilitate speech processing, but they also provide information about the intended message the speaker is trying to convey. For example, emotional prosody, or the timing, rhythm, and intonation of speech can alter the interpretation of the spoken word. Kitayama and Howard (1994) found that emotional tone of voice influenced the interpretation of sentence-length utterances such that sentences that were presented in a sad tone of voice were more likely to be interpreted as having a sad meaning than sentences produced in a happy tone of voice. Apart from conveying emotional valence, prosody can also serve as a type of vocal gesture to reduce ambiguity when communicating referential information about objects (Tzeng, Duan, Namy, & Nygaard, 2017). Tzeng et al. (2017) found that when participants used novel words (e.g., “blicket”, “daxen”) to label a swatch

of color, they varied the production of their utterances depending on the brightness of the color. For example, a participant's pronunciation varied such that brighter shades were higher pitched, higher amplitude, and shorter in duration than those of darker shades. Thus, when the linguistic content is ambiguous, speakers change their prosody to convey referential information about the target referent. Tzeng et al. (2017) also found that listeners were sensitive to this information in a subsequent perception task, suggesting that prosodic modulation impacts listeners' comprehension. Taken together, these findings suggest that surface characteristics found in a speaker's prosody directly affect linguistic processing by influencing meaning-making mechanisms and facilitating a listener's interpretation of speech.

Listeners, in addition to encoding surface characteristics linked to specific speakers and their intentions, take into account the communicative context during language comprehension. In order to examine how knowledge of speaker identity affects spoken language processing, Van Berkum, Van Den Brink, Tesink, Kos, and Hagoort (2008) examined event-related brain responses for participants listening to sentences that contained inconsistencies between speaker identity and linguistic content. (e.g., an adult saying, "I cannot sleep without my *teddy bear* in my arms"). The authors found that a speaker's identity was taken into account 200-300 msec after the start of a spoken word and, importantly, that speaker inconsistencies elicited the same type of responses as semantic anomalies (e.g., "You wash your hands with *horse* and water"). These results provide neurophysiological evidence that listeners do not process speaker identity separately from linguistic content. This finding is important because it demonstrates that variability in the speech signal, stemming from a speaker's characteristics, is not discarded, but rather rapidly encoded and integrated with linguistic cues during spoken language comprehension.

Effects of social information about the speaker on spoken language processing

Not only do listeners integrate information about speakers with the linguistic input, but they are also particularly sensitive to individual speaking styles and use this sensitivity to create experience-based expectations (e.g., Johnson, Strand, & D'Imperio, 1999). Capitalizing on gender-based differences in the pronunciation of vowels, Johnson et al. (1999) found that listeners, when shown videos that presented the gender of the speaker, shifted how they categorized vowels as a function of the speaker's gender. Listeners who were told that a female speaker produced the stimuli were more likely to label it as *hood* rather than *hud* since the former pronunciation is produced more often by females than males, suggesting that listeners access gender expectations for what the speaker should sound like during speech perception.

Listeners are also sensitive to socially conditioned variation associated with regional differences in spoken language (e.g., Hay & Drager, 2010; Niedzielski, 1999). Hay and Drager (2010) found that, when primed with concepts stereotypically associated with either Australian (koala) or New Zealand (kiwi) culture, listeners' perception of vowels shifted in accordance with the primed culture, suggesting that knowledge about the speaker's social characteristics can bias speech perception. Similarly, Niedzielski (1999) found that, when primed with information about geographical regions (e.g., Canada or Michigan), listeners were more likely to report hearing speech sounds typically associated with each regional dialect such that they heard more raised vowels in the Canadian condition (e.g., "aboot the hoose") and lower vowels in the Michigan condition (e.g., "about the house"). Taken together, these findings suggest that stored social information of a speaker's dialect can affect listener's perception of speech at the acoustic-phonetic level since listeners report hearing pronunciations that are associated with the speaker's regional background.

Furthermore, listeners' stereotypical expectations might bias them to hear speech as more or less intelligible. Rubin (1992) found that listeners who were told that a recorded lecture came from an Asian instructor, as opposed to a Caucasian one, judged the speech to be more incomprehensible and more strongly accented. Notably, listeners in both the Caucasian and Asian condition heard the same recording, suggesting that social characteristics associated with a speaker or group of speakers are salient properties that restructure our expectations and directly affect how we judge properties of speech. These findings indicate that social information about the speaker may lead to a misalignment between what is expected and actually presented in the speech signal, ultimately changing the nature of linguistic processing (Babel & Russell, 2015).

Not only is speaker race a predominant social characteristic that influences a listener's expectation when processing the speech signal, but speaker age can also shift the perception of vowels. In particular, Hay, Warren, and Drager (2006) presented listeners across conditions with the same auditory stimuli containing four different voices, but depending on condition, were paired with photos of speakers that varied as a function of age and social class. Of particular relevance to the current study is the extent to which participants' responses varied as a function of age. Specifically, the researchers found that the presence of visual stimuli significantly predicted whether participants were more likely to perceptually distinguish phonemes based of visual age of the speaker. For example, when listeners believed they were listening to an older speaker, they paid more attention to the stimuli and heard the stimuli more accurately relative to when they believed they were listening to a younger speaker. Thus, a visual representation of age seems to regulate attention to the speech signal, in turn influencing speech categorization behavior. This finding suggests that visual stimuli that provide information about the speaker, such as age and social class, can activate stereotypes that are subsequently used to allocate

attentional resources for the understanding and categorization of the incoming speech signal.

The above findings align with theoretical views that support the influence of knowledge about speakers creating expectations that affect linguistic processing. For example, in an *exemplar model of speech production and perception,* detailed properties of incoming acoustic signals are stored in listeners' linguistic representations (Goldinger, 1996). Goldinger (1996) argues that listeners form exemplars by encoding highly detailed representations of the speaker's utterances. During speech perception, sets of exemplars are activated based on how similar the incoming speech signal is to those exemplars. As a result, incoming voice information that matches our stored exemplars may result in faster processing times (Palmeri, Goldinger, & Pisoni, 1993). Sumner (2015) extends these models by making specific predictions about the encoding of indexical information, or information about the speaker's social, physical, and psychological characteristics. For example, Sumner (2015) describes a process called *social weighting* by which we access social representations early in speech processing, allowing potential biases to influence the activation of linguistic representations and the allocation of cognitive resources for understanding spoken language. According to Sumner (2015), we employ voice cues to adjust our subsequent perception in the most contextually relevant manner.

Kleinschmidt, Weatherholtz, and Jaeger (2018) propose the *ideal adapter model* in which individuals are continually tracking statistical input about speaker and group-specific acoustic distributions to make predictions about what they will hear. Specifically, the model provides an explanation for both the ability to learn novel instances of speaker-specific distributions and to draw upon stored implicit knowledge of speaker and group-specific statistics. Rather than just acknowledging that speech perception is sensitive to differences between speakers, this model proposes a way to quantify the extent to which socio-indexical features are informative and

helpful for organizing the variation in language that listeners must contend with. According to this model, listeners learn an internal model of speaker variability, that includes socio-indexical features, and employ it to make inferences about an unfamiliar speaker's utterances. As a result, listeners are able to infer an unfamiliar speaker's age, sex, and regional origin based on the stored knowledge of speaker cue distributions.

Speech rate, voice onset time, and the age of the speaker

Acoustic characteristics such as phonemes and vowels vary as a function of social variables and are interpreted with respect to a specific social context. Additional examples of these acoustic characteristics are speaking rate and voice-onset-time (VOT). Sidaras (2011) found that speech rate is a salient property that listeners use to make perceptual judgements about the age of a speaker such that listeners will rate a slower-rate utterance as spoken by an older rather than younger individual. Sidaras (2011) also presented evidence that primed with a stereotype of an "old" or "young" person, listeners will create an expectation of the speaker and accommodate to the speaker's expected speaking rate (e.g., speak at a slower speaking rate after hearing older speakers). This finding suggests that vocal accommodation, defined as the way listeners adjust their speech production to another speaker's speech, can be altered as a function of the expected social characteristics of the speaker. Ultimately, Sidaras (2011) showed that knowledge about the speaker activates social expectations and stereotypes that affect not only how listeners perceive what they hear, but also how they then produce speech themselves.

In addition to influencing the perceptual judgement of age, speaking rate changes acoustic realization of a host of different acoustic properties (Allen & Miller, 2001) and can influence the categorization of speech sounds that depend on timing information. One such rate-

dependent property is voice onset time or VOT, defined as the timing between the release of the consonant and the onset of voicing. For example, the primary distinction between the voiced consonant /b/ and its voiceless counterpart /p/ lies in differences in VOT, as /p/ has a longer VOT. When producing a /p/ sound, vocal cord vibrations begin after a short delay relative to the pronunciation of /b/ during which vocal cord vibration occurs at approximately the same time as the release of the consonant. VOT has been shown to vary systematically with speaking rate such that as speaking rate slows down (and syllables become longer), the VOT boundary shifts systematically to longer VOTs (Miller & Volaitis, 1989; Theodore, Miller, & DeSteno, 2009). This systematic variation is important because it constitutes a source of stability at the individual speaker level that listeners can use to organize their representations of speech sound categories.

VOT also systematically varies with vowel length (VL), defined as the perceived duration of the vowel sound. Because vowel lengths shorten and lengthen systematically with the rate of speech, it has been employed as an indicator of speaking rate. Changes in vowel length (VL) have been found to affect the categorization and the category structure of voiced and voiceless sounds (e.g., /b/ and /p/ categories; Toscano & McMurray, 2015). For instance, at slow rates and longer vowel lengths, listeners shift their VOT boundaries toward longer VOTs and identify more sounds as voiced (e.g., more /b/ sounds) (Miller, Green, & Reeves, 1986; Toscano & McMurray, 2015). Conversely, at fast rates and shorter vowel lengths, listeners identify more sounds as voiceless (e.g., more /p/ responses) and perceive shorter relative VOTs (Miller, Green, & Reeves, 1986; Toscano & McMurray, 2015).

Not only does VOT vary as a function of speaking rate, but it also varies as a function of speaker (Allen, Miller, & DeSteno, 2003) and coarticulatory context (Nearey & Rochet, 1994). In their analysis, Kleinschmidt (2019) report that when comparing gender, dialect, and speaker

identity, speaker identity resulted in the most informative variable for VOT distributions. Additionally, given the systematic variation between speakers, listeners can then make inferences about an unfamiliar speaker's age (e.g., Kleinschmidt, Weatherholtz & Jaeger, 2018). Kleinschmidt et al. (2018) find that it is possible to infer an unfamiliar speaker's age based on their VOT distributions because voiceless stops such as /p/ have primarily been found to have shorter VOTs when produced by older speakers, despite a slower speaking rate. This is thought to occur because, with age, physiological and anatomic changes affect an individual's production of speech (Benjamin, 1982). These findings suggest that VOT is an informative variable that varies a function of speaker identity, specifically with age of the speaker.

Lastly, listeners are sensitive to systematic variation in VOT associated with particular speakers or types of speakers. Zhang and Holt (2018) found that information about the speaker can lead to changes in listeners' categorization of speech sounds varying in VOT. For instance, when presented with visual and auditory speech stimuli (e.g., audiovisual presentations of speech tokens spoken by a female speaker) participants changed the way they categorized words along a /b/ - /p/ continuum. Consequently, Zhang and Holt (2018) argue that statistical learning can lead to *adaptive plasticity* or *perceptual learning* as a result of exposure to speaker-specific acoustic realizations of speech. If there is lexical information that suggests that the observed ambiguity is due to a characteristic of the speaker producing it, the system will recognize that it is informative and helpful and will adapt accordingly. In other words, the authors suggest that listeners are not only sensitive to and track speaker-specific patterns, but that they also make rapid adjustments by reweighting their perception. Taken together, this suggests that listeners are tracking speaker-specific patterns of VOT and employing them while perceiving speech and categorizing stimuli.

The present study

The present study includes two experiments that explore how vowel length and knowledge of speaker age shift the categorization of speech sounds. VOT is an interesting variable to use to examine the role of social expectations in relation to speech perception because it is well-established that listeners are sensitive to variation in speaking rate and consequently adjust the VOT boundary for categories of sounds varying as a function of speaking rate (Theodore, Miller, & DeSteno, 2009; Toscano & McMurray, 2015). Age is an interesting social factor because it encompasses robust social stereotypes that may influence the perception of speaking rate and vowel length. Given that there is little work examining how linguistic cues such as VOT are influenced by how listeners link systematic variation with particular groups and expectations that are accessed during early stages of language processing, variation in VOT provides an avenue for exploring the idea of how social information might affect acoustic-phonetic realization and perception.

The primary goal of the current study is to clarify how knowledge about the speaker can affect listeners' perception and categorization of speech sounds. It is known that speaking rate varies with age (Torre & Barlow, 2009) and listeners assume that older people exhibit slower speaking rate patterns (Sidaras, 2011). Additionally, it has also been shown that VOT varies systematically with speaking rate, with slower speaking rates associated with longer VOTs. Evidence also suggests that listeners can estimate speaking rate from vowel length (VL) and that VOT shifts as a function of VL (Toscano & McMurray, 2015; Theodore et al., 2009). However, it remains unclear how knowledge about the speaker's age might affect how participants listen to VOT.

Experiment 1 examines how changes in vowel length affect the perception of VOT tokens and serves to confirm that the perception of VOT can be shifted in preparation for

assessing the expectations of speaker knowledge in Experiment 2. Experiment 2 examines how knowledge of the speaker's age can influence the perception of VOT in the speech signal. A possible prediction regarding the way knowledge of the speaker will influence speech perception in the current study is that listeners will form an expectation based on their *stereotype* of an older speaker's speech having a slow speaking rate (Sidaras, 2011). As Sumner (2015) would argue, listeners may access representations consistent with their stereotype and will perceive what they *expect* to, even if it does not align with what they heard or the actual distribution of values associated with a particular group of speakers. As a result, listeners may perceive longer VOTs because VOT typically increases with slower speaking rate.

Experiment 1

The purpose of Experiment 1 was to replicate findings that indicate that VOT perception systematically varies with speaking rate, whereby slower speaking rates result in listeners' perception of longer VOTs (Toscano & McMurray, 2015). In line with previous studies (e.g., (Theodore, Miller, & DeSteno, 2009), vowel length was used as a proxy for speaking rate. We predicted that listeners' perception of VOT would be dependent on vowel length. In particular, longer vowel lengths would bias listeners to hear longer relative VOTs (more /b/ responses). Conversely, short vowel lengths (more /p/ responses).

Methods

Participants

Eighteen adults were recruited to participate in a 30-minute experiment. All participants (14 females, 4 males) were native speakers of American English, had no history of speech or

hearing disorders, and were between the ages of 18-30 years old (M = 21.67). Written informed consent was collected from all participants under a research protocol that had been approved by the Emory University Institutional Review Board. Participants were compensated $10 for their participation in the study.

Stimuli

To examine how vowel length influences the perception and categorization of words varying in VOT, four nine-step VOT continua were created with /b/ (voiced) and /p/ (voiceless) as endpoints, one for each set of /b/-/p/ minimal pair words (bath-path, beak-peak, bike-pike, buck-puck). Words were recorded by a male speaker of an ambiguous (perceived as sounding neither young nor old) age (Sidaras, 2011). Using Praat (Boersma & Weenink, 2018), a sound analysis software program, referents varying along a nine-step VOT continua were created by cross-splicing the voiced and voiceless tokens with the original voiced token serving as the /b/ endpoint (Toscano & McMurray, 2015). Each VOT step was created by removing a period of the onset equal to the duration for a specific VOT step and splicing it with the same amount of voiceless aspiration onto the voiced token (e.g., a 5-ms voiced segment would be removed from the voiced token, and 5-ms of the voiceless token would be added to the voiced token create the 5-ms VOT step).[1] The same splicing process was repeated to create each VOT step token across the four minimal word pairs. As a result, each VOT continuum consisted of endpoint stimuli at 0 VOT and 40-ms VOT, yielding nine tokens. Given that there were four minimal word pairs, a total of 36 tokens were created.

[1] Each token was marked from the onset of the release burst to the corresponding length of any given VOT token (e.g., 5-ms VOT step) for splicing at zero crossings.

To manipulate speaking rate, the vowel length duration (VL) of tokens in each continuum was either lengthened (long VL) or shortened (short VL) using the pitch-synchronous overlap-add method in Praat (Toscano & McMurray, 2015). Following this method, the onset and the offset of each vowel were identified by measuring the release burst to the offset of voicing. Once identified, this area was either increased or decreased by 40% of its original duration (Toscano & McMurray, 2012; 2015). Mean vowel duration for the short VL condition was 88ms and 203ms for the long VL condition.

Procedure

Participants completed a 20-minute computerized experimental task in a sound attenuated room in the Speech and Language Perception Lab of Emory University. Before beginning the experiment, participants were informed that they would be hearing words over headphones and categorizing speech sounds. Participants were randomly assigned to one of two conditions: short VL or long VL. Participants in each condition heard each of the 36 tokens with either shortened or lengthened vowels (4 word pairs x 9 continuum steps) repeated 9 times, totaling to 324 trials, with trial order randomized. Each trial began with a fixation cross (500ms) appearing on the computer screen, followed by presentation of the stimulus sound. Participants were encouraged to make their responses as quickly as possible without compromising accuracy. Participants were told that their task was to categorize each stimulus as beginning with a /b/ or /p/ by pressing one of two buttons, one corresponding to /b/ and the other to /p/ on a response box. Stimuli were presented binaurally via Beyerdynamic DT100 headphones on a Dell PC computer using Eprime 2.0 (Schneider, Eschman, & Zuccolotto, 2002) stimulus presentation software.

Results and Discussion

In order to examine categorization behavior across the four continua, a repeated-measures analysis of variance (ANOVA) examining vowel length (long vs. short) as a between-subjects factor and VOT (nine steps) as a within-subjects factor assessed the extent to which vowel length shifted VOT perception. The proportion of /b/ responses was calculated for each step of the four continua for both vowel lengths. A significant main effect of VOT step was found, Greenhouse-Geisser corrected F (2.84, 45.5) = 330.74, *partial* η^2 = 0.95, $p < 0.01$, suggesting that the proportion of /b/ responses varied as a function of token step across the VOT continua. The main effect of VL was non-significant, F (1, 16) = 1.30, *partial* η^2 = 0.08, $p > 0.05$. However, the interaction effect between VOT step and VL was significant, Greenhouse-Geisser corrected F (2.84, 45.5) = 2.38, *partial* η^2 = 0.13, $p < 0.05$, suggesting that the effect of VL on proportion /b/ responses varied across the different VOT steps in the continua. As seen in Figure 1, and consistent with Toscano and McMurray (2015), participants reported more voiced (/b/) responses for long VL and more voiceless (/p/) responses for short VL for the tokens in the middle of the continua.

Nine independent sample t-tests were conducted to assess the effect of VL on proportion /b/ responses for each continuum step. There was a significant difference in the average proportion /b/ responses at step 4 between the long VL (M = 0.85, SD = 0.19) and short VL conditions (M = 0.64, SD = 0.17); $t(16) = 2.43$, $p = 0.027$. This suggests that across VL conditions, participants significantly differed in how they categorized step 4. Those who heard long VL categorized this token reliably more often as beginning with a /b/ versus /p/ sound across the four word-pair continua. No other significant differences were found between VL condition for the other eight VOT steps.

These findings support our prediction that a listener's perception of VOT would be dependent on vowel length whereby longer VL would lead to the perception of longer relative VOTs (more /b/ responses) and short VL would lead to more voiceless responses (more /p/ responses). This pattern of results provides evidence that categorization of the tokens among the four-minimal pair VOT continua can be shifted. The significant interaction between VOT step and VL is driven primarily by responses to continuum step 4. However, this is not unexpected, as VL is most likely to exert an effect on categorization when the token is most ambiguous with respect to VOT. That VL biased VOT perception in the predicted manner justifies our conceptualization of VL as a proxy for speech rate, and importantly, as a means by which listeners' perception of VOT can vary contextually. Further, the findings suggest that altering the acoustic characteristics of the signal can change the perception and categorization of VOT.

Experiment 2

Given the findings in Experiment 1 on how vowel length shifts the perception of VOT, Experiment 2 investigated to what extent knowledge about the speaker affected perception of speech. Specifically, we examined if knowledge of the speaker's age could shift the perception and categorization of VOT. Participants were primed with young or old stereotypes (in the form of pictures) and were asked to categorize an age-ambiguous speaker's words as beginning with a /b/ or a /p/. Having an age-ambiguous speaker allowed us to examine to what extent listeners when primed with an age stereotype used their expectation about a speaker during the perception of speech. Participants heard the same auditory stimuli in two conditions, one in which participants saw a photograph of a young speaker, and one in which they saw a photograph of an old speaker.

We predicted that listeners would generate predictions, based on expectations of how an

older or younger speaker should sound, and perception would be based on those expectations even if those biases did not align with what they heard (Sumner, 2015). As a result, this process would lead to differences in categorization of speech sounds depending on age prime. For example, if listeners think they are listening to a young speaker, they will expect and perceive a faster speaking rate and shift their VOT boundaries toward shorter VOTs. Hence, listeners primed with a young speaker stereotype will categorize more words as beginning with /p/ since more voiceless sounds are identified at fast rates (Toscano & McMurray, 2015). However, if they think the speaker is older, they will perceive a slower speaking rate and shift their VOT boundaries toward longer VOTs, ultimately categorizing more words as starting with /b/ since more voiced sounds are identified at slow rates (Toscano & McMurray, 2015). Thus, we expected that listeners would use their expectations of a young person speaking faster and an old person speaking slower to guide their perception and categorization behavior.

Methods

Participants

Thirty-four adults were recruited to participate in a 30-minute computerized experiment. All participants (24 females, 10 males) were native speakers of American English, had no history of speech or hearing disorders, and were between the ages of 18-30 years old (M = 19.03). Data from one participant was excluded due to failure to meet the inclusion criteria of being a native English speaker. No additional data were excluded. Written informed consent was acquired from all participants under a research protocol that had been approved by the Emory University Institutional Review Board. Participants were compensated with $10 or received course credit for their participation.

Stimuli

The nine-VOT steps of the four sets of /b/-/p/ minimal pair words (bath-path, beak-peak, bike-pike, buck-puck) were presented to the participants. Unlike in Experiment 1, here listeners only heard tokens including the unaltered vowel lengths. Given that Experiment 1 confirmed that vowel length could shift categorization behavior, here we used the unaltered tokens to examine if priming of age could act in a similar way to shift categorization responses. In order to prime implicit stereotypes of age, a set of two pictures of males varying in age was presented to participants (Appendix A). These photographs were used in a previous study (Sidaras, 2011) to reliably correspond to an "old" and "young" stereotype and were originally taken from a lifespan database of facial stimuli of neutral expressions produced by Minear and Park (2004).

Procedure

Participants completed the same procedure as in Experiment 1 with the exception that the listeners in this study were presented a picture of one of the two speakers on each trial. Participants were randomly assigned to one of two conditions, old or young speaker. Participants in the two conditions heard the same stimuli. However, the speaker's expected age was manipulated across conditions such that half of the participants were shown a photograph of a young speaker, and the other half, an old speaker. During the task, the photograph of the speaker was presented at the start of each trial before the sound onset and remained on the screen until the participant made a response. On each trial, participants heard a given token over the headphones and were asked to categorize it as beginning with a /b/ or /p/.

Results and Discussion

In order to examine categorization behavior across the four continua, a repeated-measures analysis of variance (ANOVA) including Age (old vs. young) as a between-subjects factor and

VOT (nine steps) as a within-subjects factor assessed the extent to which knowledge about the age of the speaker shifted VOT perception. The proportion of /b/ responses was calculated for each step of the four continua for both Age primes. A significant main effect of VOT step was found, Greenhouse-Geisser corrected F (2.49, 77.10) = 481.62, *partial* η^2 = 0.94, $p < 0.01$, indicating that the proportion of /b/ responses varied as a function of token step across the VOT continua. A significant main effect of Age was also found, F (1, 31) = 6.39, *partial* η^2 = 0.17, $p < 0.05$, suggesting that priming the participants with a picture of an old or young male led to significantly different categorization of the VOT tokens. The interaction between VOT step and Age was significant, Greenhouse-Geisser corrected F (2.49, 77.10) = 4.49, *partial* η^2 = 0.13, $p < 0.01$, suggesting that the effect of Age on proportion /b/ responses varied across the different VOT steps in the continua. As seen in Figure 2, there were more voiced (/b/) responses when listeners were shown a picture of a young speaker, whereas there were more voiceless (/p/) responses when shown a picture of an older speaker. Further, comparing the VOT category boundaries, we see that presentation of the young picture shifted responses to the right, toward longer VOTs, and the old picture produced a shift to the left, toward shorter VOTs.

Nine independent sample t-tests were conducted to assess the effect of Age prime proportion /b/ responses for each continuum step. There was a significant difference in the average proportion of /b/ responses for VOT step 4 (old prime, M = 0.78, SD = 0.17; young prime, M = 0.92, SD = 0.12; $t(31) = -2.63$, $p = 0.013$) and step 5 (old prime, M = 0.42, SD = 0.25; young prime, M = 0.66, SD = 0.24; $t(31) = -2.90$, $p = 0.007$). No other t-tests significantly differed between Age conditions.

Taken together, these data provide evidence that a speaker's age can shift the categorization of speech sounds. Participants who viewed the photograph of the young speaker

shifted their responses toward voiced (more /b/ responses) and perceived longer VOTs whereas participants exposed to the old picture shifted towards the voiceless boundary (more /p/ responses) and perceived shorter VOTs. The findings suggest that age serves as a salient social factor that listeners use to categorize a speaker's utterance. However, the current findings do not align with the hypothesis that participants primed with a young picture would respond with more /p/ categorizations due to the expected faster speaking rate of the speaker. That is, those primed with the old picture did not shift their categorization toward the voiced boundary and provide more /b/ responses due to the slower expected speaking rate of the speaker. Instead, the findings indicate that listeners are not using an expected stereotype of an older speaker speaking slower and a younger speaker speaking faster to adjust their perception of speech.

General Discussion

The purpose of the present research was to investigate the extent to which social information about the speaker influences the perception of speech. Specifically, we examined if knowledge of a speaker's age would influence the perception of VOT. Two experiments were conducted that employed a categorization task with participants labeling tokens along nine-step VOT continua as either beginning with a /b/ or a /p/. Experiment 1 served to replicate findings that perception of VOT systematically varies with vowel length, whereby longer vowel lengths result in listeners' perception of longer VOTs and more voiced responses (Toscano & McMurray, 2015). As predicted, results showed that categorization of perceived VOT was influenced by the vowel length presented, whereby longer vowel lengths resulted in more voiced responses and, conversely, shorter vowel lengths resulted in more voiceless responses.

Experiment 2 examined the extent to which knowledge of a speaker's age can shift the categorization of age-ambiguous speech. Participants viewed a photograph of either an old or a

young speaker prior to categorizing the tokens. Listeners are not just using their global expectation of a younger speaker speaking faster and an older speaker speaking slower, rather it appears that they are tracking distributions of specific acoustic-phonetic features associated with groups of speakers (Kleinschmidt, 2019). Our findings are consistent with evidence that older speakers, despite having a slower speaking rate, produce shorter VOTs (Kleinschmidt, 2018; Torre & Barlow, 2009). Thus, the observed pattern of results suggests that listeners are encoding and learning an internal model of speaker variability that registers speaker and speaker group-based categories. Further, the findings demonstrate that social knowledge of the speaker can change listeners' categorization of speech sounds. Taken together, these results indicate that a linguistic cue such as vowel length and a social cue such as a picture of a speaker are sufficiently salient to shift a listener's categorization boundaries of VOT and change the way they perceive and categorize speech.

The results of Experiment 1 provide evidence that the VOT continua used in this experiment were orderly since systematic shifts in the perception of VOT that are expected with changes in vowel length were observed. Differences in VOT across the steps of the continua (0-40ms) significantly predicted categorization of the tokens as beginning with a /b/ or a /p/. This finding is well-established and consistent with the literature on perceptual categorization of voicing continua that shows how changes in VOT lead to shifts in the categorical boundaries between voiced and voiceless phonemes (Lisker & Abramson, 1964; Miller & Volaitis, 1989) Further, we found that changes in vowel length influenced categorization of tokens across the steps in the VOT continua. This result is consistent with past research that reports the effect of vowel length on the categorization of VOT (McMurray, Clayards, et al., 2008; Toscano & McMurray, 2012, 2015). According to this effect, vowel length interacts with VOT to shift

listeners' segmental category boundaries. Our results align with this literature since we found that longer vowel lengths led to the perception of more voiced tokens and shorter vowel lengths to the perception of voiceless tokens. Since we used vowel length as a proxy for speaking rate, it is pertinent to note that findings in the speaking rate literature also note systematic shifts in VOT perception as a result of changes in speaking rate, whereby slower speaking rates result in longer VOTs (Theodore, Miller, & DeSteno, 2009).

The results from Experiment 2 demonstrate that a visual cue to speaker age can change speech categorization behavior. Critically, we found that priming the participants with a picture of an old or young male led to significantly different categorization of VOT across the continua. Our original predictions were based on past research indicating that more voiceless sounds are identified at fast rates compared to more voiced sounds identified at slow rates (Toscano & McMurray, 2015). As a result, we hypothesized that listeners who viewed pictures of a younger or older speaker would be primed to access speech patterns that might be stereotypical of younger or older speakers generally, such as faster speech rate associated with younger speakers leading to the categorization of more voiceless sounds, and conversely, when primed with an older picture, slower rates associated with older speakers leading to the categorization of more voiced sounds. However, this was not the pattern seen. Interestingly, we found that participants primed with a young picture shifted their VOT boundaries toward longer VOTs and categorized more sounds as voiced whereas those primed with an old picture shifted toward shorter VOTs and reported more sounds as voiceless. As a result, what we see may be consistent with observations that in fact older adults, despite having a slower speaking rate, shift their VOT boundaries and produce shorter VOTs (Kleinschmidt et al., 2018; Torre & Barlow, 2009)

One possible explanation for the observed direction of the category boundary shifts as a result of the age prime is that listeners are tracking speaker variability and using their internal models about the specific acoustic-phonetic characteristics of old and young speakers to categorize what was actually present in the acoustic speech signal. This suggests that listeners are tracking and registering variation in the speech associated with groups of particular people and using it during subsequent speech processing. This idea is complimentary to that of Kleinschmidt et al.'s (2018) *ideal adapter model* that describes how listeners continually track statistical input about group-specific acoustic distributions to make subsequent predictions.

Kleinschmidt et al. (2018) argue that listeners interpret acoustic-phonetic cues differently based on associations with speaker's apparent socio-indexical features, which are used and stored as implicit knowledge. Thus, when considering how knowledge of the speaker influenced speech perception in the current study, one possible explanation consistent with our findings and Kleinschmidt et al.'s (2018) model is that older speakers, despite their slower speaking rate, produce shorter VOTs (Kleinschmidt et al., 2018; Torre & Barlow, 2009). That the pattern of our findings is consistent with this explanation indicates that listeners are not just using an *expectation* of slower or faster speaking rate when primed with an older or younger speaker, but rather they appear to be registering and encoding *distributions* of variation associated with speakers and groups of speakers such that they are developing internal models of speaker-specific categories.

It is tempting to definitively conclude that listeners are learning an association between Age and VOT, however findings exploring the relation of VOT and Age are mixed in that VOT has been found to be more variable with increasing age (Sweeting & Baken, 1982) and either shorter or not significantly different (Torre & Barlow, 2009). Thus, we cannot know for certain

what internal model of speaker-distributions listeners are employing during speech perception. However, our findings suggest that at some level, listeners are in fact tracking acoustic-phonetic variation in VOT associated with speakers and that this is contributing to classes of linguistic representations being employed during speech perception.

Supporting this possible explanation, previous research suggests that individual speakers differ in their acoustic-phonetic properties of their speech, and listeners use these differences to recognize both a speaker's voice and words during speech perception. In particular, Allen, Miller, and DeSteno (2003) found that individual speakers differ from one another in their VOT productions. Not only do speakers exhibit differences in VOT productions, but listeners are also sensitive to this variation and encode the association while processing the signal. For example, Allen and Miller (2004) found that listeners were sensitive to speaker differences in VOT and that they used this variation to identify the speaker in subsequent tasks. After training listeners on the speech from two speakers, the VOT of each speaker was manipulated so that one had short VOTs and the other long VOTs. The authors found that when presented with the manipulated tokens at test, the listeners could select the variant of speech consistent with the speaker exposed to during their training, suggesting that they learned to associate a speaker with a characteristic pattern of VOT. Notably, the registration of variation associated with a specific speaker created during training generalized to novel words, suggesting that sensitivity to differences in VOT may contribute in part to the benefit of speaker experience. Taken together, these findings indicate not only that speakers differ in their VOT productions, but that listeners are also sensitive to these differences and use it to identify speakers.

Our findings support theories that argue for an association between speech and social cues in that speaker identity and variation associated with the speech of individual speakers and

groups of speakers is not discarded, rather it is encoded and used during speech perception. Processing of phonetic content and variation associated with speakers are not independent processes. We encode socio-acoustic properties that influence spoken word recognition and meaning making (Van Berkum et al., 2008). Further, detailed properties of incoming acoustic signals are extracted and stored in listener's representations to aid subsequent speech perception (Goldinger, 1996). Sumner (2015) posits that listeners readily map sound patterns in speech to our stored social representations of different groups and individuals to determine our allocation of cognitive resources during speech perception. As a result, our biases can influence the way we attend and process speech. The current findings provide evidence for these classes of theories since it suggests that the photograph of the speaker served as a cue that activated listeners' stored representations associated with a given speaker, young or old, that impacted the way they perceived and categorized speech. Thus, differing weights of saliency of an individual's exemplar model might influence the activation and allocation of cognitive resources and generation of expectations related to registration of particular kinds of variation.

The weighting of social categories in generating predictions about speech input varies as a function of the listeners' previous experiences. Some listeners may have greater exposure to the varying relationships between variants produced and speaker age associated with different individuals, resulting in differing saliencies of an age stereotype (Drager, 2011). For example, some individuals may have many experiences with an older speaker that has allowed for encoding of that speaker's voice and the relation to a given age group, therefore creating instances that are encoded into memory. As a result, when providing cues similar to an individual's experience, unconscious expectations about the speaker, based on their exemplar memory, may influence the perception of the speech signal. This suggests that, compared to a

less salient relation between speech sound and speaker group, salient social information that activates similar encoded instances in memory influences the perception more strongly. Listeners that have encountered more variation in speaker utterances are likely to differentially attend to the speech signal and place more attention on atypical, socially salient tokens, thereby inducing greater encoding (Sumner, 2014). One factor that is thought to contribute to the saliency of atypical tokens is the level of surprisal that a listener experiences when prior expectations based on linguistic and social contexts mismatch with the auditory input (Jaeger & Weatherholtz, 2016). In other words, atypical tokens represent novel variants that, when encountered, "stand out" to the listener. Consequently, high frequency of exposure or perceived informativeness of this atypical variant results in an increased likelihood that the listener will associate the variant with, for instance, a particular social group. Taken together, this suggests that greater exposure to variation in any given category will likely result in more salient relations between speech sound categories and speaker groups that are employed to facilitate speech perception.

This proposal is one that has been similarly posited by Johnson (2005) where he argues that based on similarity between the speech sounds and existing representations of exemplars, the strength of connection between exemplars and cognitive categories may explain the categorization of stimuli associated with socio-linguistic variation. Depending on the extent of similarity between the spoken utterance and the stored exemplar, there will be differing degrees of expectation or category activation used to categorize speech. For example, our finding that Age influenced perception of VOT and categorization behavior indicates that listeners are tracking variation and are employing their stored exemplars when categorizing the presented utterances. Hence, this suggests that listeners are encoding sociolinguistic acoustic-phonetic

variation associated with speakers and, depending on the similarity of the given cue to that stored exemplar, will produce differences in categorization behavior.

The finding that introducing a photograph of a speaker biased categorization behavior suggests visual facial information is an important source of information in spoken language processing. Heald and Nusbaum (2014) argue that presence of a visual face provides a reliable cue about the speaker identity and the social category membership. To test the effect of seeing a speaker's face on word recognition performance, Heald and Nusbaum (2014) presented participants with either audio-visual or audio-only versions of the same recordings of the words. The authors found that listeners who had been exposed to both the audio and a visual picture of a speaker's face were faster at recognizing the speech during a speeded word recognition task than those only exposed to the audio. This result is consistent with the idea that seeing a person speak provides an additional source of information about the speaker that is used during linguistic processing (e.g., Johnson, Strand, & D'Imperio, 1999). Furthermore, Zhang and Holt (2018) found that visual information conveying speaker identity was sufficient for listeners to track distinct acoustic regularities of the speech signal associated with speakers. For example, by using the same auditory stimuli and pairing it with either a male or a female visual face, the authors found that the visually cued attributes of the speakers were enough for listeners to use their knowledge of speech distributions associated with a speaker and shift their perception of linguistic characteristics. Thus, in relation to the present experiment, these findings suggest that the presence of the speaker's face acted as a robust social cue that listeners used to interpret and categorize the VOT tokens.

The extent to which listeners create expectations based on social cues can result in different behaviors. Sidaras (2011) argues that social variables interacted with lower-level

perceptual motor links in speech and led listeners to accommodate by adjusting their speech production based on the expectation rather than what was perceptually present. For example, participants primed with an old stereotype accommodated by slowing their shadowing responses while those primed with a young stereotype sped up their shadowing responses. Of particular relevance to the current study is that Sidaras (2011), in addition to a written description of the speaker, used the same picture stimuli as that employed in Experiment 2. Thus, for Sidaras (2011), the presented social cues elicited expectations of young and old speech that influenced a participant's general motor and accommodation behavior. With respect to our findings, we see how the same visual stimuli led listeners to use their expectations of the speaker to process the perception of speech and shift their categorization behavior. Taken together, these findings suggest that social cues are one way by which we can influence the link between perception of speech and motor and accommodation behaviors in speech communication.

Given that we integrate both sensory and prior knowledge to understand the speech signal, speech perception can be thought to include both bottom-up and top-down processes (Remez, Rubin, Pisoni, & Carrell, 1981). Since prior knowledge has been found to facilitate speech perception primarily through top-down modulation (Sohoglu, Peelle, Carlyon, & Davis, 2012), we believe that the visual picture of the speaker served as a social clue that influenced a top-down process. The obtained findings do not align with the possibility that listeners rely solely on sensory information during the perception of speech, as would be expected in a primarily bottom-up process. Listeners in both conditions heard the same acoustic speech signal, the only difference was the presentation of either a young or an old picture. Thus, it seems that perceptual experience is based on something in addition to information present in the speech

signal. A listener's prior knowledge and expectation of what a speaker sounds like is salient enough to produce perceptual biases that result in shifts in categorization.

As mentioned previously, in supporting of this account, previous studies have reported that the presentation of social information influences speaker expectations and affects a listener's perception of various properties of speech (e.g, Johnson, Strand, & D'Imperio, 1999; Hay & Drager, 2010; Niedzielski, 1999; Rubin, 1992). For example, Johnson, Strand, and D'Imperio (1999) found that presenting videos including information about the gender of the speaker led listeners to access expectations consistent with how men and women speak and caused shifts in how they categorized vowels as a function of the speaker's gender. In a similar way, the current study finds that visual age of the speaker influences how participants categorize VOT. Thus, our findings indicate that information of the speaker affects top-down processes that lead to shifts in the perception of speech.

The current findings can be extended in many ways. Given how knowledge and exposure to speakers impact our social representations that are employed during speech perception, conducting this experiment with listeners of an older age would provide an additional account of how our representations are influencing our perception. Particularly, it would help clarify the extent to which listeners are in fact employing robust age group based representations since we would expect that with greater age and possibly more experience with a range of speakers, listeners would have more salient relationships between linguistic variants and age (Drager, 2011).

Without systematically controlling for listeners' previous exposure to stereotypically old versus young speech, we were unable to explicitly assess the type of expectations listeners were generating from the visual information of the speaker. Thus, another way to extend the current

findings is by creating an exposure phase where we explicitly manipulate what speaker or speaker group representations are being formed. In this way, we would be better able to understand the association between specific types of expectations and changes in speech perception. Although the two pictures presented in the current study represented talkers in two different age groups, the two speakers might have differed on other salient characteristics, such as approachability, that may have affected listeners' categorization behavior. Future work should also rule out the possibility that merely presenting any visual cue would influence speech categorization, for example, by presenting listeners with visual representations of inanimate objects (e.g., shapes) rather than pictures of the speakers. Given how age was seen to be a robust social cue that listeners take into account when categorizing speech, future research should continue to examine what other social cues or speaker group stereotypes create similar effects.

Finally, future studies should also examine the role of speaker familiarity and exposure to different individual and groups of speakers varying in VOT. Given that past research suggests that listeners are able to identify individuals differing from one another based on their VOTs (Allen & Miller, 2004) and that exposure to variation in speaker utterances may lead to more salient relationships and differences in category activation (Johnson, 2005; Sumner, 2014) would exposure to multiple speakers affect the extent to which listeners track acoustic-phonetic variation? Further, would this provide a benefit in perceptual fluency during speech processing? The current study suggests yes, but more research is required in this area.

Overall, in addition to finding that categorization of VOT can be changed by modifying the acoustic characteristics of the speech signal, this study suggests that knowledge of a speaker's age can influence the perception of VOT. More broadly, the current results suggest that listeners are registering specific acoustic-phonetic variation associated with speakers and

encoding this variation along with sum total of their experience with a particular group of individuals. Broadly, this research tells us that knowledge of social characteristics of a speaker can alter a listener's judgment and perception of speech. As listeners, we are exposed to social characteristics of speakers that are rapidly encoded and integrated when we process the speech signal. By understanding how indexical information, such as a speaker's individual identity, regional origin, and social characteristics, can alter a listener's judgement of speech, we can address the type of representations listeners are using when interacting with others.

The current study contributes to the existing literature by providing additional evidence that listeners encode speaker-related differences in speech in a way that directly affects spoken language processing. More specifically, the current findings offer novel evidence that listeners are tracking and encoding differences in VOT associated with groups of individuals differing in age, and using them to form expectations of the speaker to facilitate speech processing. These findings encourage future research in several directions to more clearly characterize the extent to which social variables and expectations of a speaker can influence our perception of speech.

References

Allen, J. S., & Miller, J. L. (2001). Contextual influences on the internal structure of phonetic categories: A distinction between lexical status and speaking rate. *Perception & Psychophysics, 63*(5), 798-810. doi:10.3758/bf03194439

Allen, J. S., Miller, J. L., & DeSteno, D. (2003). Individual talker differences in voice-onset-time. *Journal of the Acoustical Society of America*,*113*, 544-552. https://doi.org/10.1121/1.1528172

Allen, J. S., & Miller, J. L. (2004). Listener sensitivity to individual talker differences in voice-onset-time. *Journal of the Acoustical Society of America*, *115*, 3171-3183. https://doi.org/10.1121/1.1701898

Babel, M., & Russell, J. (2015). Expectations and speech intelligibility. *The Journal of the Acoustical Society of America*, *137*(5), 2823-2833.

Benjamin, B. J. (1982). Phonological performance in gerontological speech. *Journal of Psycholinguistic Research, 11*, 159–167.

Boersma P., & Weenink D. (2018). Praat: Doing phonetics by computer (Version 6.0.43) [Computer program]. Retrieved November 12, 2018, from http://www.praat.org

Drager, K. (2011). Speaker Age and Vowel Perception. *Language and Speech*, *54*(1), 99–121. https://doi.org/10.1177/0023830910388017

Firth, J. R. (1964). *The tongues of men.* London: Oxford University Press.

Goldinger, S. D. (1996). Words and voices: Episodic traces in spoken word identification and recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 22,* 1166–1183.

Hay, J., Warren, P., & Drager, K. (2006). Factors influencing speech perception in the context of a merger-in-progress. *Journal of Phonetics, 34*(4), 458-484. https://doi.org/10.1016/j.wocn.2005.10.001.

Hay, J., & Drager, K. (2010). Stuffed toys and speech perception. *Linguistics*, 48(4), 865–892. https://doi.org/ 10.1515/ling.2010.027.

Heald, S. L. M., & Nusbaum, H. C. (2014). Talker variability in audio-visual speech perception. *Frontiers in Systems Neuroscience,* Volume 5*,* doi:10.3389/fpsyg.2014.00698

Jaeger, T. F., & Weatherholtz, K. (2016). What the heck is salience? How predictive language processing contributes to sociolinguistic perception. *Frontiers in Psychology, 7,* 1–5. https://doi.org/10.3389/ fpsyg.2016.01115.

Johnson, K., Strand, E.A. & D'Imperio, M. (1999). Auditory-visual integration of talker gender in vowel perception. *Journal of Phonetics*. 27, 359-384.

Johnson, K. (2005). Speaker normalization in speech perception. In Pisoni, D.B. & Remez, R. (eds) *The Handbook of Speech Perception.* Oxford: Blackwell Publishers (pp. 363-389).

Kitayama, S., & Howard, S. (1994). Affective regulation of perception and comprehension: Amplification and semantic priming. In P. M. Niedenthal & S. Kitayama (Eds.), *The heart's eye: Emotional influences in perception and attention* (pp. 41-65). San Diego, CA, US: Academic Press. http://dx.doi.org/10.1016/B978-0-12-410560-7.50009-0

Kleinschmidt, D. F., Weatherholtz, K., & Jaeger, T. F. (2018). Sociolinguistic Perception as Inference Under Uncertainty. *Topics in Cognitive Science*. doi:10.1111/tops.12331

Kleinschmidt, D. F. (2019). Structure in talker variability: How much is there and how much can it help?, *Language, Cognition and Neuroscience, 34:1,* 43-68, DOI: 10.1080/23273798.2018.1500698

Labov, W. (1972). *Sociolinguistic patterns*. Philadelphia: University of Pennsylvania Press.

Liberman, A., Cooper, F., Shankweiler, & M. Studdert-Kennedy (1967). Perception of the speech code. *Psychological Review, 74*, 431-461.

Lisker, L., & Abramson, A. S. (1964). A cross-language study of voicing in initial stops: Acoustical measurements,'' *Word, 20*, 384–422.

McMurray, B., Clayards, M. A., Tanenhaus, M. K., & Aslin, R. N. (2008). Tracking the time course of phonetic cue integration during spoken word recognition. *Psychonomic Bulletin and Review*, *15*, 1064–1071. doi:10.3758/PBR.15.6.1064

Miller, J. L., Green, K. P., & Reeves, A. (1986). Speaking rate and segments: A look at the relation between speech production and speech perception for the voicing contrast. *Phonetica, 43*, 106-115

Miller, J. L., & Volaitis, L. E. (1989). Effect of speaking rate on the perceptual structure of a phonetic category. *Perception & Psychophysics, 46*, 505–512.

Minear, M., & Park, D.C. (2004). A lifespan database of adult facial stimuli. *Behavior Research Methods, Instruments, & Computers, 36*, 630–633.

Mitchel, A. D., Gerfen, C., & Weiss, D. J. (2016). Audiovisual perceptual learning with multiple speakers. *Journal of Phonetics, 56*, 66–74. doi:10.1016/j.wocn.2016.02.003

Nearey, T. M., & Rochet, B. L. (1994). Effects of place of articulation and vowel context on VOT production and perception for French and English stops. *Journal of the International Phonetic Association, 24*(1), 1–18. doi:10.10 17/S0025100300004965

Niedzielski, N. (1999). The effect of social information on the perception of sociolinguistic

variables. *Journal of Language and Social Psychology, 18*(1), 62–85. https://doi.org/10.1177/0261927X99018001005.

Nygaard, L. C., Sommers, M. S., & Pisoni, D. B. (1994). Speech perception as a talker contingent process. *Psychological Science, 5*(1), 42–46. https://doi.org/10.1111/j.1467 9280.1994.tb00612.x.

Nygaard, L.C. Perceptual Integration of Linguistic and Non-Lingusitic Properties of Speech. (2005). In: Pisoni DB, Remez RE, editors. *The Handbook of Speech Perception.* Blackwell Publishing; Oxford

Palmeri, T. J., Goldinger, S. D., & Pisoni, D. B. (1993). Episodic encoding of voice attributes and recognition memory for spoken words. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 19*(2), 309-328. http://dx.doi.org/10.1037/0278- 7393.19.2.309

Peterson, G. E., & Barney, H. L. (1952). Control methods used in a study of the vowels. *Journal of the Acoustical Society of America, 24,* 175-184. http://dx.doi.org/10.1121/1.1906875

Pisoni, D. B. (1997). Some thoughts on "normalization" in speech perception. In K. Johnson & J.W. Mullennix (Eds.), *Talker variability in speech processing* (pp. 9- 32). San Diego: Academic Press.

Remez, R. E., Rubin, P. E., Pisoni, D. B., & Carrell, T. D. (1981). Speech perception without traditional speech cues. *Science, 212,* 947–950.

Rubin, D. L. (1992). Non-language factors affecting undergraduates' judgments of nonnative English-speaking teaching assistants. *Research In Higher Education, 33* (4), 511-531.

Schneider, W., Eschman, A., & Zuccolotto, A. (2002). *E-Prime: User's guide*. Pittsburgh, PA: Psychology Software Incorporated.

Sidaras, S. K., Alexander, J. E., & Nygaard, L. C. (2009). Perceptual learning of systematic variation in Spanish-accented speech. *The Journal of the Acoustical Society of America*, *125*(5), 3306-3316.

Sidaras, SK. (2011). Hearing what you expect to hear: The interaction of social and cognitive mechanisms underlying vocal accommodation. (Doctoral dissertation). Emory University

Sohoglu, E., Peelle, J. E., Carlyon, R. P., & Davis, M. H. (2012). Predictive top-down integration of prior knowledge during speech perception. *Journal of Neuroscience*, *32*(25), 8443–8453. doi:10.1523/JNEUROSCI.5069-11.201

Summerfield, Q., & Haggard, M. P. (1973). Vocal tract normalization as demonstrated by reaction times. *Report of Speech Research in Progress, 2*, 12–23.

Sumner, M., Kim, S.K., King, E., & McGowan, K.B. (2014). The socially weighted encoding of spoken words: a dual-route approach to speech perception. *Front in Psychology. 4*, 1–13

Sumner, M. (2015). The social weight of spoken words. *Trends in Cognitive Sciences, 19*(5), 238–239.

Sweeting, P. M., & Baken, R. J. (1982). Voice onset time in a normal-aged population. *Journal of Speech and Hearing Research, 25,* 129–134.

Theodore, R. M., Miller, J. L., & DeSteno, D. (2009). Individual talker differences in voice onset-time: Contextual influences. *The Journal of the Acoustical Society of America,125*(6), 3974-3982. doi:10.1121/1.3106131

Torre, P., & Barlow, J. (2009). Age-related changes in acoustic characteristics of adult speech. *Journal of communication disorders. 42.* 324-33. 10.1016/j.jcomdis.2009.03.001.

Toscano, J. C., & McMurray, B. (2012). Cue-integration and context effects in speech: Evidence against speaking-rate normalization. *Attention, Perception, and Psychophysics, 74,* 1284-1301. doi:10.3758/s13414-012-0306-z

Toscano, J.C, & Mcmurray, Bob (2015). The time-course of speaking rate compensation: effects of sentential rate and vowel length on voicing judgments. *Language, Cognition and Neuroscience, 30*(5), 529-543, DOI: 10.1080/23273798.2014.946427

Tzeng, C.Y, Duan, J., Namy, L.L., & Nygaard, L.C. (2017). Prosody in speech as a source of referential information. *Language, Cognition and Neuroscience.* 1-15. doi:10.1080/23273798.2017.1391400.

Van Berkum, J. J., van den Brink, D., Tesink, C. M., Kos, M., & Hagoort, P. (2008). The neural integration of speaker and message. *Journal of Cognitive Neuroscience. 20.* 580–591. doi:10.1162/jocn.2008.20054

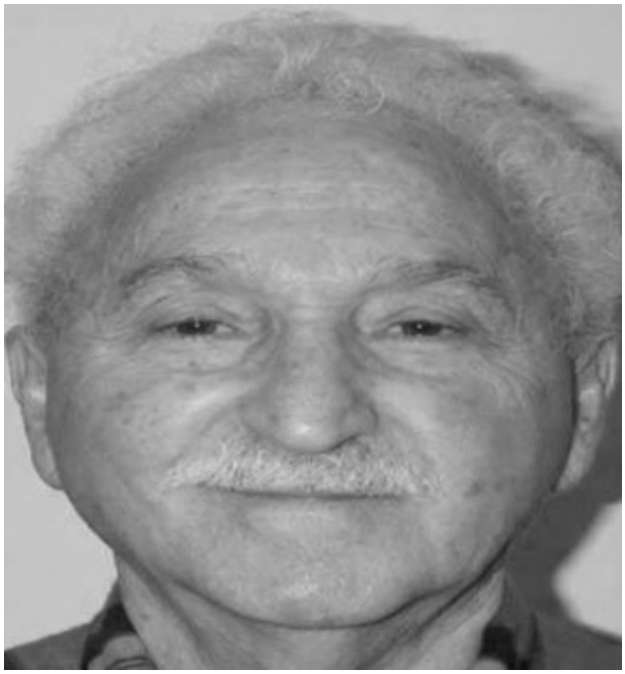
Walker, A. & Hay, J. (2011). Congruence between 'word age' and 'voice age' facilitates lexical access. *Laboratory Phonology*, *2*(1), pp. 219-237. doi:10.1515/labphon.2011.007

Zhang, X., & Holt, L. L. (2018). Simultaneous tracking of coevolving distributional regularities in speech. *Journal of Experimental Psychology: Human Perception and Performance, 44*(11), 1760-1779. http://dx.doi.org/10.1037/xhp0000569

Appendix A
Pictures of Old and Young Speakers for Experiment 2

Young Speaker

Old Speaker



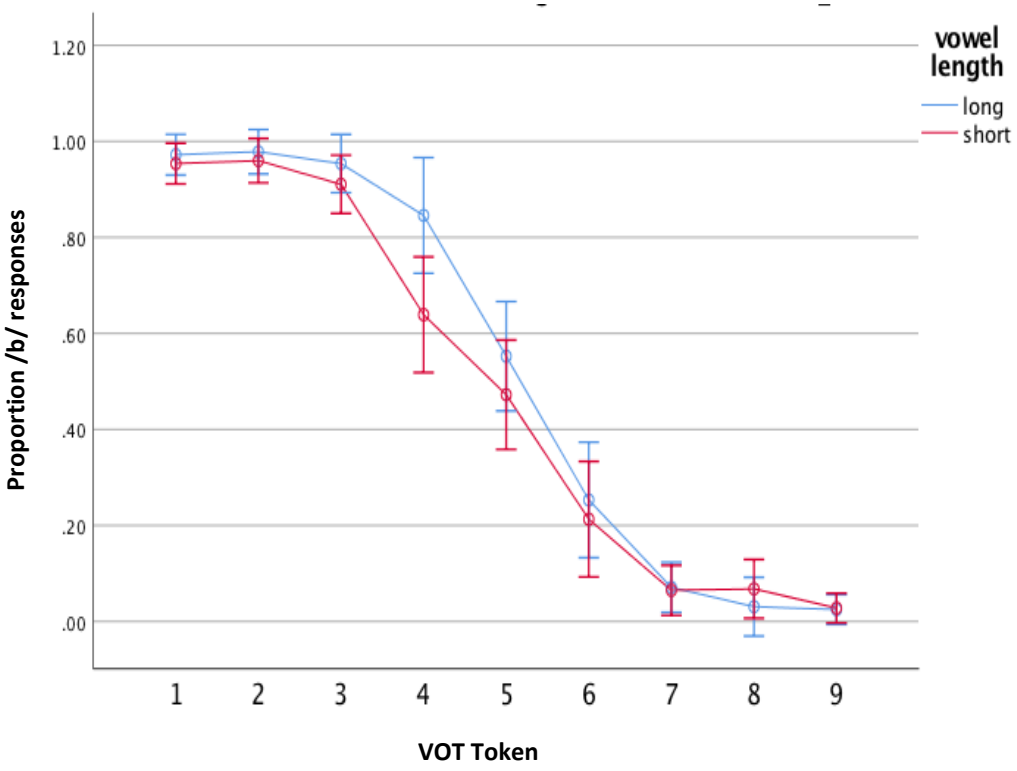


Figure 1. Responses across all continua as a function of vowel length and VOT. Error bars represent ± 1 standard error from the mean.

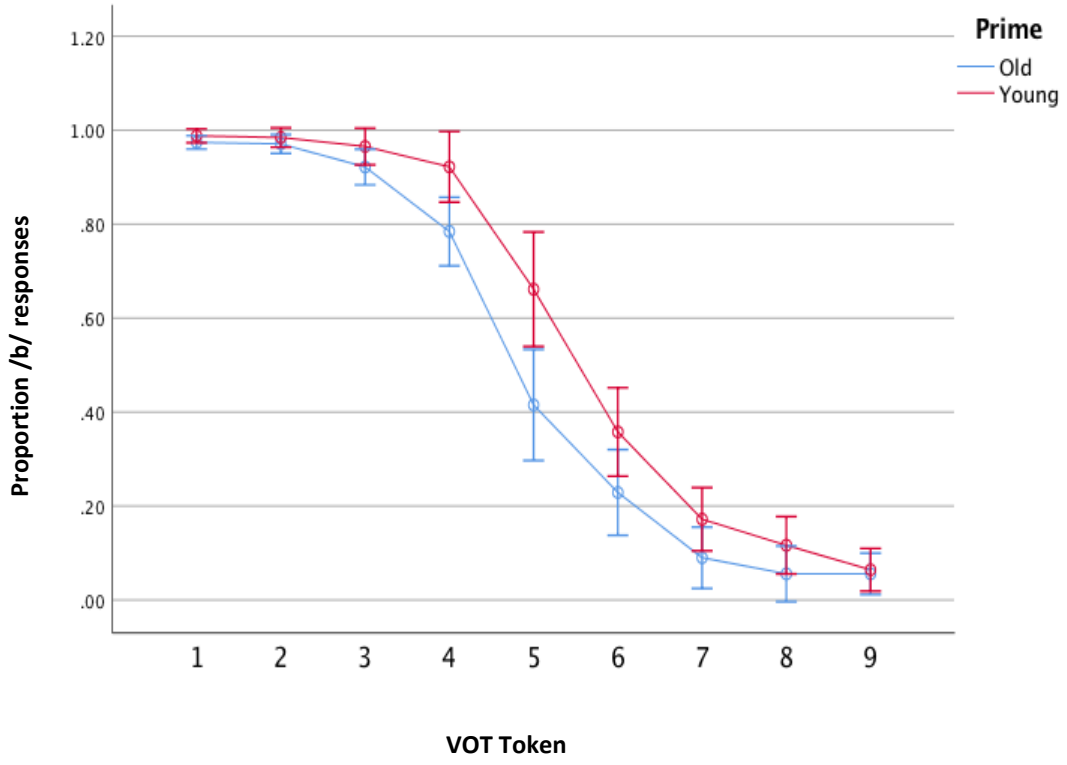


Figure 2. Responses across all continua as a function of Age and VOT. Error bars represent ± 1 standard error from the mean.