Speaker identity and spectral influences on word recognition

Santiago Barreda and Georgia Zellou

University of California, Davis
sbarreda@ucdavis.edu, gzellou@ucdavis.edu

ABSTRACT

Listeners are quicker responding to a word the second time it is heard, but this effect is reduced when the word is repeated by a different speaker. Is this reduction related to the auditory dissimilarity between different voices, or does it result from top-down effects associated with perceived speaker changes? To investigate this, listeners were presented with words differing in their pitch and/or apparent vocal-tract length, and performed a lexical-decision task where words were repeated at different delays between repetitions. Listeners also performed a voice-difference rating task using the same words and voices. At short delays, response-time patterns are better explained by perceived speaker-changes than by auditory dissimilarity. However, at longer delays between repetitions, response times are only affected by spectral mismatches. Results suggests that perceived speaker changes may influence the use of acoustic cues in word recognition, but only in the immediate vicinity of a perceived speaker change.

Keywords: word perception, speaker normalization, speaker perception

1. INTRODUCTION

Auditory repetition priming is the well-established phenomenon whereby reaction times to a heard item (the target) tend to be faster after prior exposure with the same item (the prime) [e.g., 2]. This is referred to as facilitation in word recognition, where greater facilitation indicates a larger difference between prime and target response times.

In prior studies, manipulating gross acoustic aspects of the prime-target relationship, in addition to varying delays between prime and target, has been used to assess which phonetic details in speech can influence lexical access immediately, and at longer temporal distances. For example, when a target is produced by a speaker of opposite gender from an immediately presented prime there is less facilitation compared to identical prime-target items [5]. However, with delays as short as 10 words, the negative effect of speaker changes on facilitation disappears [5, 8].

These prior findings raise questions about the relationship between acoustic detail and speaker identity perception in speech processing. For example, why (for immediately repeated words) is facilitation greater for repetitions from the same voice relative to repetitions from a different voice? One possibility is that degree of facilitation is positively related to the perceptual similarity between the prime and target voices. Since words produced by different speakers are likely to sound less similar than same-speaker words, there is likely to be less facilitation.

Another possibility is that top-down effects related to perceived speaker changes are influencing facilitation in lexical activation. For example, it is known that speaker normalization carries a cognitive cost that results in increased reaction time [1, 6, 7]. Furthermore, the estimation of apparent speaker characteristics after the detection of a speaker change may further increase reaction times after a detected speaker change. In either case, facilitation would be reduced as a result of perceived speaker changes, and by auditory differences only indirectly.

The current study aims to tease apart these two possibilities by systematically varying the acoustic characteristics of a voice to test whether facilitation patterns are best explained by gradient auditory similarity, or by the perception of speaker changes.

2. METHODS

Listeners first took part in an auditory lexical-decision task. Following this, listeners carried out a voice-difference rating task. In both tasks listeners were presented with words from a single male speaker that were modified to vary systematically in fundamental frequency (f0) and/or apparent vocal-tract length.

2.1. Participants

Thirty-seven native English speakers participated (aged 18-33; 28 female); all were undergraduates who received partial course credit for participation.

2.2. Stimuli

Stimuli consisted of 96 unique words and 96 unique nonwords produced by a male, native-speaker of English. All words were monosyllabic (the real words were high frequency content words). Recordings of these words were manipulated in the following manner using Praat [3].

First, the pitch of each word was smoothed so that it decreased linearly from 120 Hz to 100 Hz across the voiced portion of the word. The high f0 level was created by taking the words with smoothed f0 contours, and resynthesizing these words with an f0 that decreased linearly from 240 Hz to 200 Hz over the same span of the word. The result was two f0 levels for each word: one appropriate for an adult male (low f0) and another appropriate for an adult female (high f0).

Two copies were made of each word at the high and low f0 levels. One version was unchanged, so that it would imply a vocal-tract length (VTL) appropriate for an adult male (low VTL). For the second copy, the spectral envelope was shifted up by 15%, to a level roughly appropriate for an adult female (high VTL). This resulted in four versions of every stimulus word, with every word existing at high and low f0 and VTL levels.

2.3. Lexical decision task

Listeners heard a series of 408 items in a single block, presented with Eprime. For each trial, listeners were presented with an item auditorily over headphones, and were asked to indicate whether the word was a real word of English or not, as quickly and accurately as possible.

The list was composed of 96 real words 96 non-words, and 24 fillers. Each real word and nonword was represented twice in the list, once as prime and once as target, while each filler (all real words) was presented only once. Prime-target pairs were split into four voice contrast conditions based on the acoustic differences between them: same or different f0, crossed with same or different VTL. Twenty-four real words and 24 non-words were randomly selected for each contrast condition, for each listener.

Prime-target pairs were presented in one of two lag conditions: immediately after each other (“immediate priming”), or separated by 5 unrelated trials (“delayed priming”). Voice contrast conditions were balanced within lag levels, resulting in 12 prime-target pairs for each voice contrast condition, at each lag level.

2.4. Voice difference rating task

Only real words were used for this second task. Listeners were presented with pairs of words at the four different voice contrast conditions: same or different f0, crossed with same or different VTL. Twenty-four word pairs were randomly assigned to each voice contrast condition, for each listener.

For each trial, listeners heard a pair of words and were asked to indicate 1) whether they were pronounced by the same person or two different people, and 2) how different the voices sounded, using a continuous, sliding scale.

3. RESULTS

3.1. Lexical decision task

Facilitation was calculated by subtracting the response time to a target (in milliseconds), from the response time to its corresponding prime so that larger values indicate a relatively faster response for a target. Only correct responses to real word trials were included in this analysis. Because of large variation in both degree and range of facilitation between individual listeners, facilitation was standardized within-listener.

Facilitation was analysed in terms of the acoustic differences between the prime and target words, and the lag between prime and target. A three-way repeated-measures ANOVA was carried out on the average (standardized) facilitation according to the factors: VTL difference, f0 difference, and prime-lag condition (immediate or delayed priming). The result of this analysis is presented in Table 1.

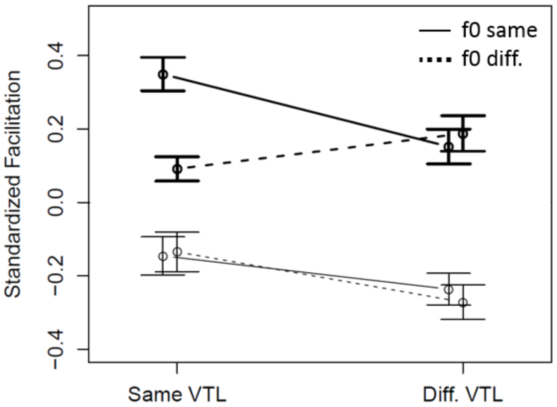
Table 1: Results of an ANOVA on facilitation with the factors: f0 (same or different), VTL (same or different) and Lag (immediate or delayed priming).

Effect	F (1,36)	p
f0	4.31	0.045*
VTL	4.41	0.043*
Lag	97.34	<0.001*
f0 x VTL	2.27	0.141
f0 x Lag	2.60	0.115
VTL x Lag	1.02	0.318
f0 x VTL x Lag	6.73	0.014*

The analysis presented in Table 1 indicates significant main effects for all three factors, however there is also a significant three-way interaction between the three factors. Fig. 1 displays this interaction. The three-way interaction between delay and VTL and f0 differences will be discussed in terms of the simple effects of VTL and f0 differences across prime-lag levels.

For delayed priming, there is only a significant effect for VTL differences [F(1,36) = 5.12, p = 0.036], and no other significant main effects or interactions. VTL differences between pairs resulted in less facilitation, and this difference was not significantly affected by the presence or absence of f0 differences.

Figure 1: Average facilitation, standardized within participant. More facilitation indicates a faster response to the second presentation of a word. Bold lines indicate immediate priming, standard lines delayed priming.



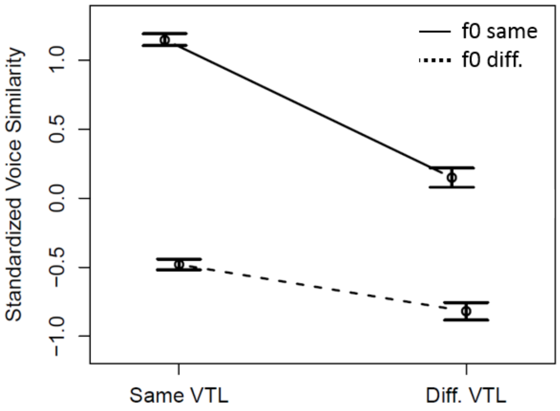
In contrast, for immediate priming there is a significant main effect for f0 difference [F(1,36) = 7.75, p = 0.008] and a significant interaction between f0 and VTL differences [F(1,36) = 10.16, p = 0.003], but no significant effect for VTL difference [F(1,36) = 1.02, p = 0.32]. The VTL x f0 interaction and the absence of a significant effect for VTL difference may be understood in terms of the changing effect for VTL differences across f0 difference levels. When voices had the same f0, there was a significant effect for VTL differences [F(1,36) = 7.38, p = 0.010], however this effect is no longer significant when pairs had different f0 [F(1,36) = 2.36 , p = 0.133].

3.2. Voice difference rating task

The voice-difference rating task consisted of a continuous voice-difference rating, and a binary response indicating whether the listener believed that the voices being compared were one or two different speakers. Continuous voice-difference responses were standardized within-listener to compensate for different uses of the continuous difference scale between participants, and reversed in sign so that they would correspond to voice similarity ratings.

To investigate the effects of gross acoustic differences on perceived voice-dissimilarity a two-way repeated measures ANOVA was carried out with two within-subjects factors: VTL difference (same or different) and f0 difference (same or different). This analysis revealed significant main effects for f0 difference [F(1,36) = 266.06, p < 0.001], VTL difference [F(1,36) = 138.80, p < 0.001], and a significant interaction between the two [F(1,36) = 43.38, p < 0.001]. Results are presented in Fig. 2.

Figure 2: Average voice similarity ratings across all listeners, standardized within participant. A larger value indicates more similar voices.



The interaction between f0 and VTL difference effects can be understood as the diminished effect for VTL differences in the presence of f0 differences, since f0 differences can result in the perception of very different voices even in the absence of VTL differences.

The percent of trials where a listener indicated hearing two different speakers was found for each voice contrast, for each listener. Results indicate that f0 differences drive the perception of speaker changes. In cases where voices had the same f0 listeners heard multiple speakers in 27% of cases, whereas when there were f0 differences between voices, listeners hears multiple speakers in 90% of cases.

4. DISCUSSION

Word recognition involves adapting to the characteristics of the speaker, a process known as normalization. In English, where changes in f0 level are not lexically contrastive, this process primarily involves adapting to the VTL of the speaker, since this will represent the primary source of spectral variation for repetitions of the same word between-speakers, within-dialect.

In light of the above, it is not surprising that VTL mismatches between prime and target result in decreases in facilitation in almost all cases in Fig. 1. Since word recognition will primarily be informed by spectral characteristics, facilitation is greatest in cases where the gross spectral characteristics (i.e., VTL) are most similar between prime and target. Furthermore, the similarity of effect for VTL at the two delay conditions indicates that this spectral information remains in memory, and is involved in the identification of future words.

Although the effect for VTL differences might be explainable in terms of voice similarity, other results are not easily explainable by such an account. First, there is no consistent effect for f0 on facilitation, and none at all with delayed priming. This is despite the fact that f0 is the strongest determiner of perceived voice similarity. Second, f0 differences result in a loss of the effect for VTL mismatches for immediate priming, resulting in the interaction seen in the top half of Fig. 1. If facilitation were related to voice similarity, the patterns in the top and bottom halves in Fig. 1 should both closely resemble the pattern in Fig. 2.

A possible explanation for these results may lie in top-down effects related to the detection of speaker changes. Facilitation may be thought of as resulting from the use of previously heard (extrinsic) information in vowel perception. It has been noted that this information is most useful in the absence of a change in speaker, which suggests the detection of speaker changes is inherently related to speech perception [6, 7]. For example, previous experiments have shown that listeners are slower to identify vowels when they are expecting speaker changes, and that perceptual errors in mixed-speaker lists are better explained by errors in detecting speaker change than by general acoustic dissimilarity [1, 7].

Mismatches in f0 between prime and target led to the perception of different speakers in a large majority of cases, and to the least similar-sounding voices. If these were as likely to result in perceived speaker changes in the lexical-decision task, this would result in an additional cognitive load related to the determination of apparent speaker characteristics and adaptation to the new speaker.

This account would explain the decrease in facilitation for f0 mismatches at immediate delay in cases where prime and target had the same VTL. This would also explain the lack of an effect for VTL mismatches at immediate priming when f0 differed between the voices. If decreases in facilitation related to f0 differences result from adaptation to the new listener, the VTL mismatch between prime and target will naturally no longer be an issue since VTL characteristics will be re-estimated for the new speaker.

This may also explain the lack of any effect whatsoever for f0 at delayed priming. If f0-related effects for immediate priming arise from a detected speaker change, then this would have no additional effect several stimuli later, since the relevant comparison is between the current stimulus and the previous one, and not between the current stimulus and all previous tokens.

It has been previously reported that that listeners are more likely to remember previously hearing a word if both instances are produced by the same speaker, and that they are explicitly able to recall if repetitions of words were produced by the same speaker, and at the same rate [4]. This indicates that information regarding f0, and the apparent speaker characteristics driven by these aspects of a voice, can be retained in memory. As a result, it cannot be said that f0 does not have an effect for facilitation at delayed priming because it is not remembered for a voice.

Instead, the lack of an effect for f0 at delayed priming suggests that spectral information remembered for previous voices may be disassociated from the indexical information related to that production, if only to facilitate speech perception. Furthermore, it suggests that lexical retrieval in English crucially depends on spectral information rather than information related to f0. Conversely, listeners rely on f0 to guide the perception of speaker changes, and may not be strongly influenced by spectral differences when it comes to lexical access.

5. REFERENCES

[1] Barreda, S. (2012). Vowel normalization and the perception of speaker changes: An exploration of the contextual tuning hypothesis. *The Journal of the Acoustical Society of America*, *132*(5), 3453-3464.

[2] Biederman, I., & Cooper, E. E. (1992). Size invariance in visual object priming. *Journal of Experimental Psychology: Human Perception and Performance*,*18*(1), 121.

[3] Boersma, P., & Weenink, D. (2009). Praat: doing phonetics by computer (Version 5.1. 05)[Computer program]. Retrieved May 1, 2009.

[4] Bradlow, A. R., Nygaard, L. C., & Pisoni, D. B. (1999). Effects of talker, rate, and amplitude variation on recognition memory for spoken words. Perception & psychophysics, 61(2), 206-219.

[5] Luce, P. A., & Lyons, E. A. (1998). Specificity of memory representations for spoken words. *Memory & Cognition*, *26*(4), 708-715.

[6] Magnuson, J. S., & Nusbaum, H. C. (2007). Acoustic differences, listener expectations, and the perceptual accommodation of talker variability. *Journal of Experimental Psychology: Human Perception and Performance*, *33*(2), 391.

[7] Nusbaum, H. C., and Morin, T. M. (1992). “Paying attention to differences among talkers,” in Speech Perception, Speech Production, and Linguistic Structure, edited by Y. Tohkura, Y. Sagisaka, and E. Vatikiotis-Bateson (OHM, Tokyo), pp. 113–134.

[8] Zellou, G., & Embick, D. (2014). Interaction of memory and specificity in auditory repetition priming. *The Journal of the Acoustical Society of America*,*135*(4), 2420-2420.