Perceptual adaptation to speaker characteristics: VOT boundaries in stop voicing categorization

Constance M. Clarke and Paul A. Luce

Department of Psychology
University at Buffalo, State University of New York
cclarke2@buffalo.edu

Abstract

Recent research suggests speech perception processes are flexible. For example, Norris et al. [Norris D, McQueen JM, & Cutler A (2003). Cognit. Psychol., 47, 204-238] demonstrated that listeners trained on stimuli containing ambiguous /s/-/f/ tokens subsequently showed an appropriate shift in their /s/-/f/ categorization boundaries based on the lexicality of the training stimuli. The present study extended this work by exploring changes in acoustic-phonetic criteria for stop category perception (voiced vs. voiceless) following brief exposure to sentence length stimuli. In a word-monitoring task, native English listeners heard sentences produced by a native English speaker in which all syllable-initial /t/ and /d/ segments were digitally modified to be atypical of native English pronunciation, specifically, to have short-lag /t/s and prevoiced /d/s. Categorization of /t/ and /d/ was tested using a 5-token voice onset time (VOT) continuum prior to exposure and again following 20, 40, and 60 sentences. As predicted, listeners' mean categorization boundary shifted to a shorter VOT after exposure to the modified speech. Results suggest listeners' perceptual criteria for stop consonants can be adjusted to better match speakers' productions. A follow-up experiment testing /g/-/k/ categorization with exposure to the same /t, d/-modified sentences did not show clear evidence for generalization to a different place of articulation.

1. Introduction

Although there is a great deal of consistency among native speakers of a language in the phonetic realization of speech sounds, subtle differences exist from speaker to speaker. For example, in addition to well-known cross-speaker variability in vowel formant frequencies [1], speakers differ systematically in production of frication noise [2] and stop consonant voice onset time [3]. Variability among talkers is even greater when different dialects, speech styles, speaking rates, and foreign accents are considered.

Given this variability in speech production, it is not surprising that recent research shows speech perception is flexible. For example, familiarity with a talker gained through lab training improves sentence transcription accuracy [4], and perceptual processing of non-native speech becomes more efficient following brief exposure [5]. However, it is not yet clear what specifically listeners learn about talkers' speech. A recent finding [6] suggests that acoustic-phonetic criteria for phoneme categorization is modifiable with experience. Dutch listeners performed a lexical decision task with words whose final fricative was replaced by an ambiguous sound between /s/ and /f/. When only /s/-final words contained the ambiguous sound, a subsequent /s/-/f/ categorization test showed more /s/ responses. The reverse was found for the /f/-final word condition. Moreover, this learning occurred with exposure to only 20 ambiguous tokens, pointing to malleable phonetic representations as one component of a flexible phonological processing system.

The present study examined whether there is similar flexibility in stop consonant categorization. We extended previous work by (a) embedding the training stimuli in sentence-length utterances, (b) examining the time course of perceptual change, and (c) testing whether adaptation to the voicing distinction at one place of articulation generalizes to another.

2. Experiment 1

In Experiment 1, listeners participated in one of two conditions: *Atypical* or *Typical*. In the Atypical condition, we exposed listeners to native speech containing alveolar stop consonants, /d/ and /t/, with voice onset times (VOTs) not typical of native English speech. VOT is the time between the stop release burst and the onset of glottal pulsing and is a strong cue to voicing in syllable-initial stops in English (and many other languages) [7]. Typical values for English /t/ are in the 30-100 ms (long-lag) range, and for /d/ are in the 0-30 ms (short-lag) range. In languages such as French and Spanish, /t/ is typically produced with short-lag VOTs, and /d/ with prevoicing (glottal pulsing beginning before the burst).

We chose VOT values for the atypical stop condition to approximate those that might be produced by a non-native speaker of English with a first language of Spanish or French, i.e., close to the short-lag range for /t/ and prevoiced for /d/. In the context of a word-monitoring task, native English listeners were exposed to /d/ and /t/ samples in a variety of words in sentence-length utterances. We predicted a shift in listeners' /d/-/t/ perceptual boundary to shorter VOTs following exposure. In addition, we tracked the time course of this shift by administering d/t categorization tests after 20, 40, and 60 sentences. Listeners in the Typical condition heard the same sentences, except they contained VOT values typical of a native speaker of English (i.e., short-lag /d/s and long-lag /t/s).

2.1. Method

2.1.1. Participants

Listeners were 116 University at Buffalo undergraduates given partial course credit for participation. All were monolingual native speakers of American English with no reported hearing disorders. The data of 14 participants were excluded for the following reasons: anomalous categorization boundary in one or more tests (4), technical problems (3), participant stopped responding (3), session interrupted (2), and high error rate in exposure task (2). The final set consisted of 102 listeners (60 females, 42 males). Half the

listeners participated in the Typical condition, and half in the Atypical condition.

2.1.2. Materials and stimuli

Stimuli included a 5-step /da/-/ta/ VOT continuum and 60 sentences for word monitoring. Sentences were a mix of Harvard Sentences [8] (e.g., The little tales they tell are false) and similar experimenter-created sentences (e.g., The last dive received a nine out of ten). The speaker was a male, aged 26, from Buffalo, New York, USA. Recording took place in a quiet room using a Sony ECM-MS907 microphone and a JVC XL-R5000 CD recorder (44.1 kHz, 16 bit).

Each syllable in the continuum was based on a naturally produced /da/ token, approximately 350 ms in duration. The /da/ burst and successively longer portions of the vowel were replaced by equivalent portions of a naturally produced /ta/ (see [9]). Each successive step corresponded to one additional glottal pulse of the /da/ vowel. The burst and aspiration amplitudes were covaried with VOT, with lower amplitudes for shorter VOT tokens. Based on pilot testing, five tokens were chosen that straddled the /d/-/t/ categorization boundary (in ms VOT): +12, +24, +34, +44, +56.

The word-monitoring sentences were designed to contain only alveolar stop consonants. (However, during Experiment 2, three /g/s and one /k/ were discovered; see section 3.1.2.) Each set of 20 sentences contained 20 /d/s and 20 /t/s in various sentence positions. (The only stops counted were syllable-initial singletons in strong syllables.) The speaker's mean VOT for the 60 /t/s was +53 ms (range: 22 to 90). The mean for those 36 /d/s with voicing lag was +11 ms (range: 6 to 23), and for the 24 /d/s with prevoicing was -47 ms (range: -18 to -68).

The sentences for the Typical condition were left as recorded. For the Atypical condition, the mean /t/ VOT was reduced to 30 ms (range: 22 to 35) by deleting aspiration from the ends of /t/ voiceless periods. Burst and aspiration amplitudes were also reduced by an amount commensurate to the amplitude reduction along the /da/-/ta/ series. For /d/ tokens, any voicing lag was deleted, and natural samples of prevoicing were added to all /d/s lacking it, resulting in a mean VOT of -44 ms (range: -18 to -68). Praat [10] and Peak (4.13, Berkeley Integrated Audio Software, Inc.) software packages were used for digital waveform editing.

Each sentence had a unique target word to monitor for. In each block of 20 sentences, 8 did not contain the target word, and 12 did (with 2 occurring at the beginning of the sentence, 4 in the middle, and 6 at the end). Target words varied in part of speech as well as in initial phoneme. Most were not /d/- or /t/-initial words.

2.1.3. Procedure

Stimuli were presented and responses recorded using PsyScope software [11] on an Apple iMac G4 computer running OS 9.2. Auditory stimuli were presented at a comfortable listening level over headphones.

The experiment took approximately 20 minutes and began with a categorization pre-test, followed by three word-monitoring blocks and three categorization post-tests, in alternating order. In each categorization test, the five syllables were presented six times each, for a total of 30 trials. Each participant heard the syllables in a different random order, with all five syllables presented once before the next repetition. Listeners were instructed to press one of two buttons labeled "TA" and "DA" as quickly and accurately as possible after hearing each syllable. Approximately half the participants in each condition pressed the button labeled "TA" with their dominant hand, and half with their nondominant hand.

Each word-monitoring block consisted of 20 different sentences presented in a different random order for each listener. Block order was counterbalanced across listeners. On each trial, a new target word was displayed orthographically on a computer screen until the participant pressed a button to begin the sentence. Participants were instructed to press the button again as soon as they heard the target word, and refrain if they did not hear it. After the sentence, "CORRECT!" or "**WRONG**" was displayed depending on the accuracy of detection.

2.2. Results and discussion

Listeners categorized each token six times per test; for the pre-test, however, responses to the first presentation of the continuum were discarded, leaving five responses per token. For each listener, the percent "D" response was determined for each token, and 50% categorization boundaries were calculated using linear interpolation.

Fig. 1 shows d/t boundaries as a function of categorization test. As predicted, boundaries shifted to a shorter VOT in the Atypical condition, but did not change in the Typical condition. A 2 (group) X 4 (test) X 3 (block order) mixed analysis of variance (ANOVA) resulted in a significant group X test interaction, $F(3, 288) = 3.05$, $p = .03$. Follow-up ANOVAs for each group showed no effect of test for the Typical group, $F(3, 144) < 1$, but a significant effect for the Atypical group, $F(3, 144) = 7.34$, $p < .001$, $\eta^2_{partial} = .13$. Planned comparisons showed a significant decrease by the first post-test, $p = .01$.

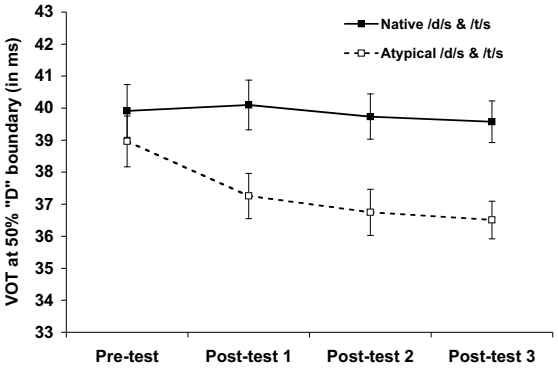

Figure 1: VOT boundaries for DA/TA categorization as a function of experience with English-typical or English-atypical /d/s and /t/s. Error bars represent standard error.

Listeners' VOT criteria for the /d/-/t/ voicing distinction changed depending on the speaker's phonetic realization of this distinction. Presumably the pre-test boundaries resulted from listeners' perceptual expectations based on experience with English. However, exposure to a speaker producing the distinction at a different point along the VOT dimension shifted the perceptual boundary in the direction of production. The effect was quite rapid, with a statistically significant

boundary decrease after only 20 /d/ and /t/ exemplars. In contrast, exposure to speech that matched perceptual expectations for the voicing distinction resulted in no boundary change from the pre-test value.

It should be noted that the total boundary change by the final post-test in the Atypical condition was only -2.5 ms VOT, a fairly small change (although a large effect statistically). This might be compared with the difference between the effective production boundaries in the typical and atypical stimuli, a 10 ms VOT difference at minimum. Although it is difficult to compare perceptual boundaries for isolated syllables and running speech, it is probably safe to say that perceptual criteria did not change to completely match the speech characteristics. An interesting question for future research is whether boundaries would continue to decrease with further exposure and ultimately match production values.

3. Experiment 2

The results of Experiment 1 indicate acoustic-phonetic criteria for stop categorization are changeable. Are the consequences of this learning restricted to /d/ and /t/, or might perception of stops at other places of articulation be affected? Experiment 2 tested whether experience with alveolar stops affects voicing categorization for velar stops. If listeners exposed to sentences containing only alveolar stops learn something about how the speaker makes the voicing distinction for stops in general, their /g/-/k/ boundary should also shift.

Previous research using other methodology suggests generalization across place of articulation might occur. In the selective adaptation literature, /p/ and /b/ were effective adaptors for /t/-/d/ as well as /p/-/b/ VOT series [12]. Further, listeners trained to discriminate [b]-[p]-[p^h] were subsequently also able to discriminate [d]-[t]-[t^h] without any additional training (and vice versa) [13]. These findings point to similarities in the voicing distinction among stops at different places of articulation that have consequences for perceptual categorization.

3.1. Method

3.1.1. Participants

Listeners were 92 University at Buffalo undergraduates given partial course credit for participation. All were monolingual native speakers of American English with no reported hearing disorders. Data from 2 participants were excluded for the following reasons: technical problems (1), and no categorization boundary in one test (1). The final set consisted of 90 listeners (49 females, 41 males). Half participated in the Typical condition, and half in the Atypical condition.

3.1.2. Materials, stimuli, and procedure

The word-monitoring sentences from Experiment 1 were used for the first 49 listeners. At that point, three /g/s and one /k/ were discovered in the sentence set. Of concern was a syllable-initial /g/ with a VOT of +25 ms, as this could provide negative evidence that the /g/-/k/ boundary should be similar to the /d/-/t/ boundary for the Atypical condition. The other three cases were not of concern: Two were word-final and unreleased; one was a prevoiced word-final /g/. For the remaining 41 listeners, the sentences containing the /g/s and /k/ were replaced. The resulting new sentence set contained 18 to 20 /d/s and /t/s in each block of 20. The new mean VOTs were the same or similar to the original set.

The /ga/-/ka/ VOT continuum was created with the same technique as the /da/-/ta/ continuum. The base /ga/ syllable was approximately 260 ms in duration. Based on pilot testing, five tokens were chosen that straddled the /g/-/k/ categorization boundary (in ms VOT): +13, +24, +33, +41, +51. All procedures were the same as for Experiment 1.

3.2. Results and discussion

Fig. 2 shows g/k boundaries as a function of categorization test. (Data for just the last 41 participants showed the same pattern of results, and the full participant set was used for all further analyses.) Contrary to expectation, boundaries for listeners in the Atypical condition did not decrease with exposure. In fact, both groups increased from pre-test to post-test 3, $F(3, 252) = 4.97$, $p = .002$, as shown by a 2 (group) X 4 (test) X 3 (block order) mixed ANOVA. There was no group X test interaction, $F(3, 252) = 1.23$, *ns*. Also unexpectedly, the Typical group's mean boundary was longer at pre-test than the Atypical group's, $F(1, 84) = 4.29$, $p = .04$. The source of this difference is unclear. However, overall mean reaction time (RT) for Atypical listeners ($M = 529$, $SD = 85.8$) was also significantly faster than for Typical listeners ($M = 581$, $SD = 88.4$), $F(1, 84) = 8.78$, $p = .004$. If the Atypical listeners tended to respond more quickly, the effective duration of the test syllables may have been shorter, and a shorter perceptual boundary could result. This is speculation, however, and it is also not clear why the groups differed in RT.

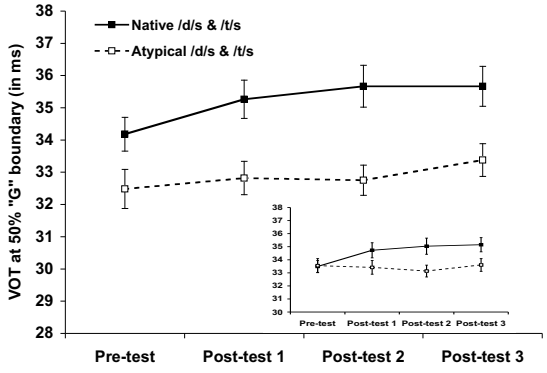

Figure 2: VOT boundaries for GA/KA categorization as a function of experience with English-typical or English-atypical /d/s and /t/s. Error bars represent standard error. Inset: groups matched on pre-test boundary and overall RT.

In case this group difference was obscuring the effects of interest, we matched the groups on mean pre-test boundary and overall mean RT (see Fig. 2 inset). After matching (*n*s = 34), the group X test interaction was statistically significant, $F(3, 186) = 2.74$, $p = .04$. Separate ANOVAs showed no effect for the Atypical group, $F(3, 93) < 1$, but a significant boundary increase for the Typical group, $F(3, 93) = 4.25$, $p = .007$, $\eta^2_{partial} = .12$. Planned comparisons showed a significant increase from pre-test to post-test 1, $p = .02$.

A possible explanation for this pattern of results is that something about the test series itself caused boundaries to increase with repeated testing and that exposure to the Atypical speech caused the opposite effect, with a net result of no change. To address this possibility, we added a No Exposure condition in which 53 participants performed the categorization tests and solved addition problems in silence in between tests. Their mean (standard deviation) boundaries were 32.27 (3.8), 32.18 (3.7), 33.38 (3.9), and 33.72 (3.4) for pre-test, post-test 1, post-test 2, and post-test 3, respectively. Similar to the Typical condition, boundaries increased significantly with repeated testing, $F(3, 153) = 7.28$, $p < .001$, but a significant increase from pre-test did not occur until post-test 2, $p = .01$.

Overall, the evidence for generalization to g/k categorization is equivocal at best. A differential effect of exposure condition was only statistically significant when groups were matched, and even then the effect was an increase in Typical boundaries, rather than a decrease in Atypical boundaries. Depending on interpretation, the increase in the No Experience group's boundary suggests that both the Typical and Atypical groups tended towards increasing boundaries, but that the effect of exposure to the atypical speech kept the Atypical group's boundary constant. However, this interpretation is questionable because the boundary change for the No Experience group was more gradual than that of the Typical group. Furthermore, the overall group differences remain unexplained.

4. General Discussion

The results of Experiment 1 extend previous findings [6] of perceptual learning for phonetic categories in two ways. A change in perceptual boundaries was shown (a) for stop voicing categorization, and (b) with exposure to exemplars embedded in continuous speech. Experiment 1 also replicated the speed of perceptual change. An effect was observed after exposure to only 20 exemplars of each category, within less than one minute of speech.

The extension of this finding to continuous speech is important because it is a step closer to a natural listening situation. It suggests that tuning of phonetic categories may underlie perceptual adaptation to a new speaker in the first moments of listening.

The results of Experiment 2 do not provide strong evidence that learning the voicing distinction for one place of articulation generalizes to a different place of articulation. However, the equivocal findings do not rule out generalization, and the question deserves further investigation, perhaps with a different speaker or phonetic contrast.

Recent studies demonstrating flexibility in speech processing challenge the notion of static linguistic representations and processes that has dominated the field of speech perception. Given the lack of invariance in speech, it is not surprising that the speech processing system must constantly adjust to maintain robust and accurate perception. The incorporation of such plasticity into models of speech perception will mark an important advance in the field.

5. Acknowledgements

This research was supported by NIDCD. Portions of this work were presented at the 148th meeting of the Acoustical Society of America, San Diego, CA, November 2004. The authors thank Jim Sawusch for helpful discussion and Tamar Izcovich, Amy Wu, and Janette Yung for assistance in data collection.

6. References

[1] Peterson, G. E., and Barney, H. L., "Control methods used in a study of the vowels," *J. Acoust. Soc. Amer.*, vol. 24, pp. 175-184, 1952.

[2] Newman, R. S., Clouse, S. A., and Burnham, J. L., "The perceptual consequences of within-talker variability in fricative production," *J. Acoust. Soc. Amer.*, vol. 109, no. 3, pp. 1181-1196, 2001.

[3] Allen, J. S., Miller, J. L., and DeSteno, D., "Individual talker differences in voice-onset-time," *J. Acoust. Soc. Amer.*, vol. 113, no. 1, pp. 544-552, 2003.

[4] Nygaard, L. C., and Pisoni, D. B., "Talker-specific learning in speech perception," *Percept. Psychophys.*, vol. 60, no. 3, pp. 355-376, 1998.

[5] Clarke, C. M., and Garrett, M. F., "Rapid adaptation to foreign-accented English," *J. Acoust. Soc. Amer.*, vol. 116, no. 6, pp. 3647-3658, 2004.

[6] Norris, D., McQueen, J. M., and Cutler, A. "Perceptual learning in speech," *Cognit. Psychol.*, vol. 47, no. 2, pp. 204-238, 2003.

[7] Lisker, L. and Abramson, A. S., "A cross-language study of voicing in initial stops: Acoustical measurements," *Word*, vol. 20, pp. 384-422, 1964.

[8] IEEE Subcommittee on Subjective Measurements, "IEEE recommended practices for speech quality measurements," *IEEE Transact. Audio Electroacoust.*, vol. 17, pp. 227-246, 1969.

[9] Newman, R. S., Sawusch, J. R., and Luce, P. A., "Lexical neighborhood effects in phonetic processing," *J. Exp. Psychol. Human Percept. Perform.*, vol. 23, no. 3, pp. 873-889, 1997.

[10] Boersma, P. and Weenink, D., "Praat: doing phonetics by computer," [Computer program], 2004 (Version 4.1.26), [Retrieved 2004 Feb 6], Available HTTP: http://www.praat.org/

[11] Cohen, J. D., MacWhinney, B., Flatt, M., and Provost, J., "PsyScope: A new graphic interactive environment for designing psychology experiments," *Behav. Res. Meth. Instrum. Comput.*, vol. 25, no. 2, pp. 257-271, 1993.

[12] Eimas, P. D., and Corbit, J. D., "Selective adaptation of linguistic feature detectors," *Cognit. Psychol.*, vol. 4, no. 1, pp. 99-109, 1973.

[13] McClaskey, C. L., Pisoni, D. B., and Carrell, T. D., "Transfer of training of a new linguistic contrast in voicing," *Percept. Psychophys.*, vol. 34, no. 4, pp. 323-330, 1983.