Differential cue weighting in perception and production of consonant voicing

Amanda A. Shultz
Linguistics Program, Purdue University, West Lafayette, Indiana 47907-2038
shultz1@purdue.edu

Alexander L. Francis[a)]
Department of Speech, Language, & Hearing Sciences and the Linguistics Program, Purdue University, West Lafayette, Indiana 47907-2038
francisa@purdue.edu

Fernando Llanos
School of Languages and Cultures, Purdue University, West Lafayette, Indiana 47907-2038
fllanos@purdue.edu

Abstract: This study examines English speakers' relative weighting of two voicing cues in production and perception. Participants repeated words differing in initial consonant voicing ([b] or [p]) and labeled synthesized tokens ranging between [ba] and [pa] orthogonally according to voice onset time (VOT) and onset *f*0. Discriminant function analysis and logistic regression were used to calculate individuals' relative weighting of each cue. Production results showed a significant negative correlation of VOT and onset *f*0, while perception results showed a trend toward a positive correlation. No significant correlations were found across perception and production, suggesting a complex relationship between the two domains.

PACS numbers: 43.70.Mn [AL]
Date Received: April 30, 2012 **Date Accepted:** June 25, 2012

1. Introduction

Theories of speech perception are often based on implicit assumptions about the relationship (or lack thereof) between mechanisms for perception and production of speech sounds, either for reasons intrinsic to the theory (Fowler, 1986; Liberman and Mattingly, 1985) or for developmental or other learning-related reasons (see Newman, 2003 for a brief review).

Early tests of these assumptions generally found little correlation between perception and production (see Newman, 2003 for a review), but recent research has been more promising. For example, Bradlow *et al.* (1997) showed that Japanese speakers' production of the English [r]-[l] contrast may improve following perceptual training even without any specific production training. Conversely, Shiller *et al.* (2009) showed that auditory feedback-driven changes to speech motor production routines were accompanied by a corresponding shift in the relevant perceptual category boundary. In studies of individual differences, Perkell *et al.* (2004) found that English speakers who were above the median in terms of vowel discrimination produced a vowel contrast more distinctively than did participants who scored at or below the median in perception, while Newman (2003) found that listeners who selected as a perceptual prototype a stop consonant with a longer voice onset time (VOT) also were likely to show longer VOTs in their productions of the same consonants.

[a)]Author to whom correspondence should be addressed.

The present study also addresses the possibility of a production-perception link in voicing, a contrast realized in terms of multiple acoustic cues (Lisker, 1986). However, instead of seeking a direct correlation of values of one cue across perception and production, the present study focuses on the relative weighting of two acoustic cues: VOT, the time between the release of a stop and the beginning of the vocal fold vibration, and onset *f*0, the fundamental frequency at the onset of phonation. In English, VOT has typically been found to serve as the primary cue to this contrast, such that voiced stops ([b], [d], [g]) have a shorter VOT than voiceless ones ([p], [t], [k]) (Lisker and Abramson, 1964), but onset *f*0 has also been identified as a significant secondary cue, such that voiced stops tend to begin with a lower *f*0 that rises into the following vowel, while voiceless stops begin with a higher *f*0 that falls into the subsequent vowel (Löfqvist *et al.*, 1989; Whalen *et al.*, 1993). Individual variability in the use of onset *f*0 as a voicing cue has been previously identified in Haggard *et al.* (1970) and there also appears to be some individual variability in production (Löfqvist *et al.*, 1995).

The purpose of this study is to explore the connections between production and perception by investigating the manner in which native English speakers' relative weighting of VOT corresponds to that of onset *f*0 in both a production task and a perception task. If mental representations for perception and production are identical or even tightly linked, then the individual weight that listeners give to specific cues in perception should correspond to the degree to which they use those same cues to instantiate a phonetic contrast in production.

2. Methods

2.1 Participants

A total of 32 native speakers of American English, 16 men and 16 women, were paid for their participation (mean age: 24.94 years, range: 20–32). All participants identified American English as their first language and none reported a history of speech or hearing disorder. Data from seven participants were excluded: Two failed a hearing screening (threshold >30 dB SPL at 0.5 kHz, or >25 dB SPL at 1, 2, 4, 6, or 8 kHz) and five failed to produce at least one short-lag exemplar of each [b]-initial word (producing only or mainly prevoiced tokens for [b]),[1] leaving data from 25 participants (11 men, 14 women) in the final analysis.

2.2 Speech production

Participants always completed the speech production task before the perception task so that insight from the perception task (which contrasted only [b]- and [p]-initial syllables) would not influence performance in the production task.

2.2.1 Speech materials

Eight target words were used (*bat*, *pat*, *bet*, *pet*, *beat*, *Pete*, *bit*, *pit*), comprising four monosyllabic, monomorphemic minimal pairs with CVC syllable structure, differing in initial stop consonant ([b] or [p]) and having a comparable frequency of occurrence in English and a high degree of familiarity (Washington University Database). Since vowels' pitch can be influenced by vowel height, four different vowels were chosen ([i], [I], [ɛ], and [æ]).

Sixteen additional monosyllabic words were included as foils (*fig*, *dig*, *fit*, *kit*, *heap*, *keep*, *feed*, *deed*, *fat cat*, *head*, *dead*, *hay*, *day*, *hot*, *cot*). All foils had item frequency and familiarity similar to that of the target words.

2.2.2 Recordings and acoustic measurements

The 24 words (8 targets, 16 foils) were presented one at a time to each participant on a computer monitor using E-Prime 1.2 (Schneider *et al.*, 2002), with a constant

presentation rate of 1 word per 2.5 s to control for influence of speaking rate on VOT (Kessinger and Blumstein, 1998). Participants were instructed to say the word aloud in a normal speaking voice as each word appeared. There were 5 blocks of 24 words (120 in all, with 20 beginning with [p] and 20 with [b]), randomized for each participant. All recordings were made using a Marantz solid state recorder (PMD660) at 44.1 kHz, 16 bit quantization, in.wav format with a unidirectional hypercardioid microphone (Audio-Technica D1000HE). Data from four words from four different participants were dropped due to mispronunciation during production.

VOT and onset *f*0 of the initial bilabial stops of the 40 target utterances were measured using Praat 5.1.25 (Boersma and Weenik, 2010). VOT was measured from the beginning of the burst to the onset of voicing as identified by hand from waveform and spectrogram displays (using Praat default settings). Onset *f*0 was measured at the first time interval at which the standard Praat pitch tracking algorithm was able to detect periodicity following the marked interval. Highly discrepant (i.e., suggesting pitch halving or doubling) *f*0 values were measured by hand by taking the inverse of the first discernable period of the vowel waveform.

For subsequent statistical analyses, frequency values for each subject were converted from hertz to semitones relative to that subject's mean onset *f*0 across all measured utterances using the semitone conversion equation provided in the Praat-internal users' manual (Boersma and Weenik, 2010), but here made relative to each individual's mean onset *f*0 instead of 100 Hz (1). This resulted in values expressing relative distance above/below the talker's mean onset *f*0 on a logarithmic scale. Values above zero represented higher-than-average onset frequencies, while negative values represented lower-than-average frequencies. This was done to facilitate combining scores across genders.

$$12 \ln (x \,/\, \text{individual mean onset } f0) \,/\, \ln 2. \quad (1)$$

As a measure of reliability, VOT and onset *f*0 values were re-measured and re-computed by another experimenter for 4 participants randomly selected from the total number of 32 participants (12.5% of the production data). All of the re-measured datasets were from participants included in the final analyses. The VOT measurements showed a strong significant correlation across the two experimenters, $r\ (157) = 0.97$, $p < 0.0001$, with a mean absolute difference in positive values of VOT (the values used in the main analysis of the present study), of ~2 ms. The onset *f*0 measurements also showed a strong significant correlation across the two experimenters, $r\ (157) = 0.99$, $p < 0.0001$, with an mean absolute difference of ~1.4 Hz.

2.3 Speech perception

2.3.1 Stimuli

Stimuli were created using the Klatt-style formant synthesizer (Klatt, 1980) implemented in Praat 5.2 (Boersma and Weenik, 2010). Tokens were designed to sound like a male talker, and varied along a two-dimensional continuum of VOT and onset *f*0. There were a total of 40 tokens, with 4 steps of onset *f*0 and 10 steps of VOT, ranging from [ba] to [pa].

The vowel had a duration of 315 ms. The amplitude of voicing rose from 45 dB at the start of the burst to 60 dB by 10 ms into the vowel of the prevoiced tokens, or from 0 dB at the start of the burst to 60 dB by 10 ms into the vowel for the positive VOT tokens. It then remained at 60 dB for 20 ms and then fell to 50 dB by the end of the syllable. Formant transitions for *F*1– *F*3 were 35 ms. *F*1 began at 220 Hz and rose to 710 Hz. *F*2 began at 900 Hz, rising to 1240 Hz, and *F*3 rose from 2000–2500 Hz. *F*4 and *F*5 were held constant at 3600 and 4500 Hz, respectively. Formant bandwidths were constant at the following values: *F*1: 50 Hz; *F*2: 70 Hz; *F*3: 110 Hz; *F*4: 170; *F*5: 250. For prevoiced tokens, the *f*0 and intensity of prevoicing was

held constant at 120 Hz and 45 dB, respectively. For all tokens, the burst had a duration of 4 ms with 4 frication formants, each with a frequency of 300 Hz and 100 Hz bandwidth. Onset *f*0 began at 90, 110, 130, or 150 Hz and converged to 120 Hz over the next 50 ms. Between the burst and the onset of the vowel, aspiration amplitude was zero for prevoiced tokens and linearly correlated with positive VOT values.

2.3.2 Task

Stimuli were presented using MATLAB 7.10 (MathWorks, 2010) via a Soundblaster Live! soundcard on a Dell Optiplex/Windows XP computer through headphones (Sennheiser HD 280 pro) at a comfortable listening level.

Participants heard one token at a time, using the MATLAB interface to click on one of two buttons labeled "BA" and "PA" to indicate what they heard. After each response, there was a 400 ms pause before the beginning of the next trial. On each trial the mouse pointer was re- centered between the two response buttons so as not to bias responses. Participants were not limited in response time, but were instructed to respond as quickly and accurately as possible. The left–right order of the response buttons was counterbalanced. There were 11 blocks of 40 tokens, for a total of 440 total trials.

3. Analysis and results

The relative weighting of the 2 acoustic cues in production was computed using discriminant analysis for each of the 25 participants, providing standardized canonical coefficients such that a larger coefficient denotes a stronger weighting of a variable. Results [Fig. 1(a)], showed a significant negative Pearson's product moment correlation, $r\ (23) = -0.42$, $p = 0.037$, between VOT and onset *f*0. Discriminant analysis was chosen over logistic regression, which would otherwise have afforded a more direct

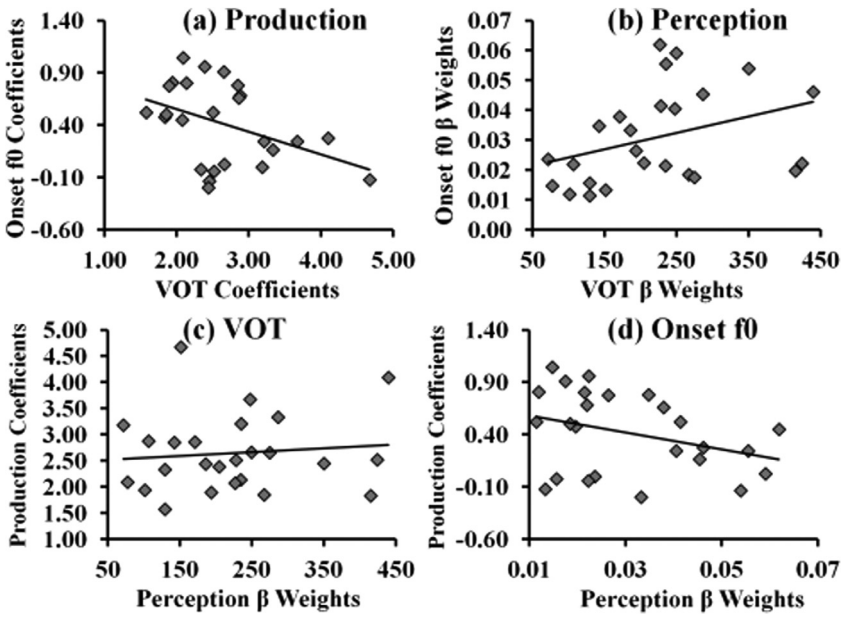

Fig. 1. Scatter plots illustrating comparisons discussed in text. Dark line indicates linear regression line. (a) Production: Total-sample canonical coefficients across VOT and onset *f*0 for each participant, $r\ (23) = -0.42$, $p = 0.037$. (b) Perception: β weights across VOT and onset *f*0 for each participant, $r\ (23) = 0.36$, $p = 0.076$. (c) VOT: Weightings across perception and production for each participant, $r\ (23) = 0.1$, $p = 0.623$. (d) Onset *f*0: Weightings across perception and production for each participant, $r\ (23) = -0.34$, $p = 0.101$.

comparison to results of the logistic regression analyses of perceptual data, because the comparatively small number of data points per subject (~40) in the production data led to unreliable estimates of β weights (p values > 0.90 for all but one subject).

Perceptual weights for VOT and onset *f*0 were calculated using logistic regression analysis of data from individual participants (cf. Kondaurova and Francis, 2008), to obtain standardized logit coefficients, or β weights, as a measure of the strength of the contribution of each variable (VOT and onset *f*0) to the category ([b] or [p]) as shown in Fig. 1(b). Unlike in production, the correlation between VOT and onset *f*0 in perception was positive, but not significantly so $r\ (23) = 0.36$, $p = 0.076$.

The canonical coefficients from the production analysis and the β weights from the perception analysis were compared for each of the two cues (VOT and onset *f*0), as shown in Figs. 1(c) (VOT) and 1(d) (Onset *f*0). For VOT there was a slightly positive but not significant correlation between production and perception, $r\ (23) = 0.1$, $p = 0.623$, while for onset *f*0 the trend was toward a negative correlation, but this was likewise not significant, $r\ (23) = -0.34$, $p = 0.101$.

4. Discussion

The observation of a significant negative correlation between VOT and onset *f*0 in production suggests that individuals who distinguish between voicing categories using large differences in VOT tend to make smaller differences between the two categories according to onset *f*0, and vice versa. The overall magnitude of the standardized canonical coefficients suggests that all of the participants are primarily users of VOT, but some emphasize VOT more than others, and, in turn, further de-emphasize onset *f*0. This is what would be expected if VOT and onset *f*0 are in a trading relationship (Repp, 1982), and is consistent with the idea that the various acoustic correlates of voicing combine into an integral "intermediate perceptual property" (Kingston *et al.*, 2008).

In perception, results were more equivocal. A significant positive correlation would suggest that individuals with sharper category boundaries along one dimension also make more categorical decisions along the other, extending theories of individual differences in acuity (e.g., Perkell *et al.*, 2004) into a multiple-feature contrast. Because the results were not statistically significant, this interpretation cannot be considered conclusive. However, the present results still showed a discrepancy between the treatment of these two cues in production, where they were significantly negatively correlated, and in perception, where they were not.

This suggests that the relative weighting of acoustic cues to the same phonetic contrast may reflect different goals in production and perception. If the goal of production is the efficient generation of an integral or combinatoric acoustic property (Kingston *et al.*, 2008), the observed correlation might indicate a type of "trading relation" in reverse: Talkers may compensate for idiosyncratic variation in one cue by scaling the other inversely, essentially never producing "too much" of a combined cue. Similarly, the observed lack of correlation in perception is consistent with an efficient coding model of speech perception in which cue weighting varies as a function of the distribution of cues in the input (Ming and Holt, 2009; Stilp and Kluender, 2011). That is, a listeners' pattern of cue weighting in the present experiment may simply reflect the fact that VOT and onset *f*0 were uncorrelated in the stimuli they received.

The lack of any significant correlation between weighting of VOT in production and perception poses a problem for theories that assume such a correspondence (Fowler, 1986; Liberman and Mattingly, 1985) or predict that it may arise through experience (Bradlow *et al.*, 1997; Shiller *et al.*, 2009). However, some caution is warranted. First, although non-significant, for VOT there was a trend toward a slightly positive correlation between production and perception, consistent with previous findings of such a correspondence (Newman, 2003) and suggesting that results more consistent with previous research might be forthcoming with more statistical power. On

the other hand, interpretation of the onset *f*0 results is less clear. Onset *f*0 is not a primary cue for English listeners (Newman, 2003), and both the canonical coefficients from the production analysis and the β weights from the perception analysis were quite small. If listeners do not give much weight to the cue in perception, they may also put little emphasis on it in production, and the lack of a correlation between the two here may simply reflect the practical irrelevance of this cue under these experimental conditions.

In conclusion, the goals of efficient perception may not correspond to those of efficient production, and this may be reflected in cue weighting differences in the two domains. In perception, listeners may weight cues partly as a function of context, showing no correlation between weighting of VOT and onset *f*0 when there is no correlation in the input. In contrast, in speech production speakers show a significant trade-off between the two cues, consistent with the idea that the goal of production is to produce a composite cue structured in terms of a correlation among more basic properties (Kingston *et al.*, 2008). The lack of significant correlations between production and perception in this study may reflect the fact that the functional goals of listening and speaking exert different pressures on cue weighting.

Acknowledgments

We are grateful to Olga Dmitrieva, Lori L. Holt, and two anonymous reviewers for comments and suggestions, and to the Linguistics Program, College of Liberal Arts, Purdue University, for research support.

References and links

[1]Further research is needed to elucidate the phenomenon of prevoicing in English (cf. Newman, 2003). For example, data collected for this project suggest that frequent prevoicing might be related to exposure to Spanish at a young age, or speaking a Southern or Western U.S. dialect. Similarly, both people with hearing loss (excluded from the present analyses) showed anomalous cue weights, suggesting that hearing loss may affect cue weighting as well (for a further discussion of individual variation in prevoicing this data set, see Shultz, 2011).

Boersma, P., and Weenik, D. (**2010**). *Praat: Doing Phonetics by Computer* (Versions 5.1.25 and 5.2) [Computer program and manual]. Retrieved January 20, 2010, from http://www.praat.org/.

Bradlow, A. R., Pisoni, D. B., Akahane-Yamada, R., and Tohkura, Y. (**1997**). "Training Japanese listeners to identify English /r/ and /l/: IV. Some effects of perceptual learning on speech productions," J. Acoust. Soc. Am. **101**, 2299–2310.

Fowler, C. A. (**1986**). "An event approach to the study of speech perception from a direct-realist approach," J. Phonetics **14**, 3–28.

Haggard, M., Ambler, S., and Callow, M. (**1970**). "Pitch as voicing cue," J. Acoust. Soc. Am. **47**, 613–617.

Kessinger, R. H., and Blumstein, S. E. (**1998**). "Effects of speaking rate on voice-onset time and vowel production: Some implications for perception studies," J. Phonetics **26**, 117–128.

Kingston, J., Diehl, R. L., Kirk, C. J., and Castleman, W. A. (**2008**). "On the internal perceptual structure of distinctive features: The [voice] contrast," J. Phonetics **36**, 28–54.

Klatt, D. (**1980**). "Software for a cascade/parallel formant synthesizer," J. Acoust. Soc. Am. **67**, 971–995.

Kondaurova, M. V., and Francis, A. L. (**2008**). "The relationship between native allophonic experience with vowel duration and perception of the English tense/lax vowel contrast by Spanish and Russian listeners," J. Acoust. Soc. Am. **124**, 3959–3971.

Liberman, A. M., and Mattingly, I. G. (**1985**). "The motor theory of speech revised," Cognition **21**, 1–36.

Lisker, L. (**1986**). "'Voicing' in English: A catalogue of acoustic features signaling [b] versus [p] in trochees," Lang Speech **29**, 3–11.

Lisker, L., and Abramson, A. S. (**1964**). "A cross-language study of voicing in initial stops," Word **20**, 384–422.

Löfqvist, A., Baer, T., McGarr, N. S., and Seider Story, R. (**1989**). "The cricothyroid muscle in voicing control," J. Acoust. Soc. Am. **85**, 1314–1321.

Löfqvist, A., Koenig, L. L., and McGowan, R. S. (**1995**). "Vocal tract aerodynamics in /aCa/ utterances: Measurements," Speech Commun. **16**, 49–66.

MathWorks. (**2010**). *MATLAB* (Version 7.10) [Computer Software]. The Mathworks, Inc., Natick, MA.
Ming, V., and Holt, L. L. (**2009**). "Efficient coding in human auditory perception," J. Acoust. Soc. Am. **126**, 1312–1320.
Newman, R. S. (**2003**). "Using links between speech perception and speech production to evaluate different acoustic metrics: A preliminary report," J. Acoust. Soc. Am. **113**, 2850–2860.
Perkell, J. S., Guenther, F. H., Lane, H., Matthies, M. L., Stockmann, E., Tiede, M., and Zandipour, M. (**2004**). "The distinctness of speakers' productions of vowel contrasts is related to their discrimination of the contrasts," J. Acoust. Soc. Am. **116**, 2338–2344.
Repp, B. H. (**1982**). "Phonetic trading relations and context effects: New experimental evidence for a speech mode of perception," Psychol. Bull. **92**, 81–110.
Schneider, W., Eschman, A., and Zuccolotto, A. (**2002**). *E-Prime User's Guide* (Psychology Software Tools Inc., Pittsburg, PA).
Shiller, D. M., Sato, M., Gracco, V. L., and Baum, S. R. (**2009**). "Perceptual recalibration of speech sounds following speech motor learning" J. Acoust. Soc. Am. **125**, 1103–1113.
Shultz, A. A., (**2011**). "Individual differences in cue weighting of stop consonant voicing in perception and production," Master's Thesis, Purdue University, West Lafayette, IN.
Stilp, C. E., and Kluender, K. R. (**2011**). "Non-isomorphism in efficient coding of complex sound properties," J. Acoust. Soc. Am. **130**(5), EL352–EL357.
Washington University in St. Louis, Speech & Hearing Lab Neighborhood Database. Available from http://128.252.27.56/Neighborhood/SearchHome.asp.
Whalen, D. H., Abramson, A. S., Lisker, L., and Mody, M. (**1993**). "F0 gives voicing information even with unambiguous voice onset times," J. Acoust. Soc. Am. **47**, 36–49.