The 146th Meeting of the ASA at Austin TX, November 11, 2003

2aSC23 Perceptual rate normalization in naturally produced bilabial stops

Kyoko Nagao Kenneth de Jong

(knagao@indiana.edu kdejong@indiana.edu)

Department of Linguistics, Indiana University www.indiana.edu/~lsl

INTRODUCTION

Influence of speech rate

1. Speaking rate affects acoustic properties of voicing production
(Miller, Green, & Reeves, 1986; Volaitis & Miller, 1992)

- VOT systematically increases as syllable duration increases.
- VOT for voiceless consonants changes more with rate than voiced consonants.
- The range of VOT distribution increases as syllable duration increases.

2. Speaking rate affects voicing contrast perception *(Summerfield 1981; Miller & Volaitis, 1989; Volaitis & Miller, 1992)*

- Changing rate shifts perceptual boundary. The boundary between perceived /b/ and /p/ is at a longer VOT when syllable duration is longer.
- Changing rate shifts best example. VOT values of the highest rated /p/ are longer when syllable duration is longer.

Characteristics of previous studies

- Production experiments examined rates which were self-controlled by speakers. *(Summerfield, 1981; Miller, Green, & Reeves, 1986; Volaitis & Miller, 1992)*
- Perception experiments used synthesized syllables. *(Summerfield, 1981; Miller & Volaitis, 1989; Volaitis & Miller, 1992)*

Studies by Miller and colleagues (*Miller & Volaitis, 1989; Volaitis & Miller, 1992*)

- VOT was modified in proportion to the total syllable duration.
- Stimuli included three categories, /b/, /p/, and a third unnatural category called the exaggerated voiceless consonant (*/p/). The third category was introduced to demarcate the category for /p/. The stimulus space is shown below.

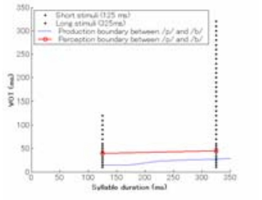

FIG 1. Stimuli space and perceptual /b/-/p/ boundary (Volaitis & Miller 1992) and /b/-/p/ boundary in production study (Miller et al 1989).

- The perceptual boundary did not match the boundary estimated from natural speech *(Miller, Green, & Reeves, 1986*) nor other studies (*Lisker & Abramson, 1970, Summerfield, 1981*). Perceptual VOT boundaries were at much higher values than estimates from production studies.

Table 1. Estimated VOT values of the category boundary and of the best /p/. VOT boundaries in other studies (Lisker & Abramson, 1970; Summerfield, 1980) are 20-30ms.

Syllable duration (ms)	125	325
VOT boundary based on production (Miller et al, 1989)	14	37
Perceptual /b/-/p/ boundary (Volaitis & Miller, 1992)	39.18	45.22
Best /p/ VOT (Volaitis & Miller, 1992)	49.17	69.17

RESEARCH QUESTIONS

- Is rate-controlled speech similar to rate-self-controlled speech ?
- Since production and perception studies find different boundaries, how does perception of naturally rate-varied speech compare with studies of previous synthetic speech?
- How do produced rate variation in voicing and perceived rate-normalized voicing judgments correlate with each other?

EXPERIMENT 1 (Acoustic analysis)

METHOD

Speakers: 4 native speakers of American English (2 female, 2 male)

Speech materials:

- Speech corpus originally collected for studying speech production (de Jong, 2001a, 2001b).
- Speakers repeated the same syllable approximately 25 times.
- Syllables used in current study were /pi/ and /bi/.
- Each syllable was repeated in time with a metronome.
- The metronome increased the rate of production throughout each utterance.

Acoustic measurements:

- The fastest 21 syllables of each /bi/ and /pi/ utterance were measured.
- Measures:
  - **VOT:** Duration from the consonant release to the beginning of voicing
  - **Syllable duration:** Duration from the consonant release to the ending of vowel
- The /b/-/p/ boundary was estimated by logistic regression analysis.

RESULTS

- Rate-induced method successfully elicits fast rate of speech.
- As syllable duration increases, VOT values for /p/ increase.
- This rate effect on VOT is large for /p/, whereas there is little rate effect on VOT for /b/.
- Range of VOT distribution of /p/ is wider than /b/.
- VOT values for /b/ and /p/ are overlapped at fast rates.
- Some tokens were produced with VOT values not typical of their category; long VOT for voiceless stops, and short VOT for voiced stops.
- Rate-dependent optimal VOT values for /b/-/p/ boundary in Miller et al (1989) does not differentiate categories successfully at fast rates.
- Stimulus VOT ranges in the previous perceptual studies greatly exceeded the ranges in natural speech.

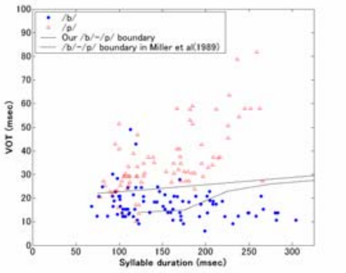

FIG 2. VOT values as function of syllable duration for /b/ and /p/. The dotted line is the optimal VOT values to differentiate /b/ and /p/ given in Miller et al (1989), while the solid line is the estimated boundary from the current study.

EXPERIMENT 2 (Voicing ID)

METHOD

Speech materials: The speech corpus from EXPERIMENT 1 was used.

Stimuli:

- 21 stimuli were spliced from each repetitive utterance as shown in FIG 3.
- Each stimulus consists of three repeated syllables.
- VOT and Syllable duration for each stimulus were based on the middle syllable in the stimulus.

Listeners: 18 native listeners of American English
Task: Four-alternative forced choice test (*'pea', 'bee', 'eep', and 'eeb'*)
Analysis: We discuss the results for voicing judgments only, not for syllable structure.

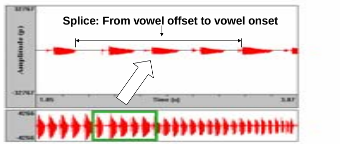

FIG 3. Example of stimulus spliced from the repetitive syllables.

RESULTS

- Identification of /p/ is accurate even at fast rates.
- Accuracy of /b/ identification decreases at fast rates.
- The perceptual boundary matches the VOT boundaries found in EXP 1.
- Boundary VOT values were estimated by logistic regression analysis with VOT and SYLLABLE DURATION as predicted variables. Positive slope indicates that as syllable duration increases, boundary VOT values increase.
- The perceptual boundary from responses to natural speech matches speakers' produced voicing distinction more accurately than the perceptual boundary from responses to synthesized speech.

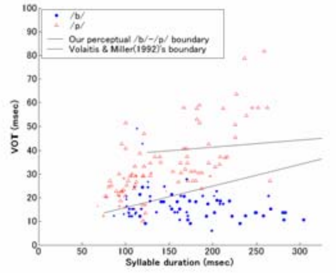

FIG 4. Relationship between perception and production results. The size of markers indicates the degree of performance on voicing identifications (The larger, the better). The dotted line is the estimated perceptual /b/-/p/ boundary in Volaitis & Miller (1992) and the straight line is the estimated perceptual boundary from the current study.

EXPERIMENT 3 (Goodness Judgments)

METHOD

Stimuli: The same stimuli from EXPERIMENT 2 were used.
Listeners: 17 native listeners of American English (different from EXP 2).
Tasks:
1. Two two-alternative forced choice tests, one for consonant identification (*'p' or 'b'*) and one for syllable structure identification (*'Consonant before Vowel'* or *'Vowel before Consonant'*).
2. Goodness rating of identified consonant from 1 (=Terrible) to 10 (=Excellent).

Analysis: We focus on goodness judgments for voicing. Mean goodness ratings were separately calculated for correct and incorrect identification. Stimuli were grouped in terms of values of VOT and syllable duration.

RESULTS

- The consonant identification results from EXP 2 were replicated. Two-tailed paired t-test of (ArcSine transformed) mean percent of /p/ responses revealed no significant difference between EXP 2 and EXP3; p=0.458.
- The highest rated /b/ and /p/ both had greater VOT values as syllable duration increased. Syllable duration affects goodness judgments of consonants.
- The listeners tended to rate /b/ with relatively short VOT and /p/ with relatively long VOT as better than their more moderate counterparts. This appears to be a 'Hyperspace effect'. Presumably it would be even more evident with the synthesized stimuli in previous studies.
- When VOT values were typical of their category, and the listeners categorized such tokens in terms of their VOT, goodness ratings were high.
- When VOT were typical of their category, and the listeners misidentified a consonant, goodness ratings were lower than when they correctly identified it. Hence, listeners apparently were aware of the stimuli's properties which mismatched their identifications, and were basing their identifications on some other property of the stimuli.

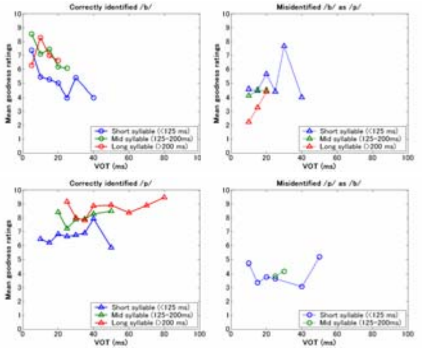

FIG 5. Mean goodness ratings for correctly identified /b/- and /p/-inputs (left top panel and left bottom panel) and for misidentified /b/- and /p/-inputs (right top panel and right bottom panel) as function of VOT values.

SUMMARY

- VOT values in production are rate sensitive. /p/ and /b/ values overlap at fast speech rates.
- In general, identification of voicing categories is accurate.
- Identification of /b/ is more affected by speech rate than identification of /p/.
- The perceptual boundary between voicing categories is rate sensitive.
- The perceptual VOT boundaries from natural speech match the values found in production. Perceptual boundaries with synthesized speech are much higher, possibly due to task-related effects in previous work, or perhaps due to a hyperspace effect in response to synthetic speech.
- Goodness judgments are also affected by speech rate.

CONCLUSIONS

- Rate normalization effects from synthesized speech also occur in the perception of natural speech.
- Speech rate affects both production and perception in a similar manner. The perceptual identification system is neatly tuned to the distributions found in production.
- The results of goodness ratings indicate that listeners store fine-grained information to distinguish voicing contrasts.
- Accurate identification of segments with aberrant VOT values suggests listeners use signal attributes in addition to VOT to differentiate the contrast.
- Even though rate-varied repetitive speech is uncommonly encountered, listeners effectively deal with rate-induced variation in categorization tasks. This suggests an active component in listeners' perceptual systems which generalizes to novel circumstances.

Acknowledgements

This work is supported by the NIDCD (grant# R03 DC04095) and the NSF (grant# BCS-9910701). We also appreciate the Indiana University Linguistics Club for their support.

Appendix

Summary of Logistic regression analysis: /p/-inputs and /p/-responses were coded 1, and /b/-inputs and /b/-responses were coded 0. VOT values at 50% cutoff points (logit = 0.5) were computed as the /p/-/b/ boundaries.

logit = - 4.564 + .232*VOT - 0.007*SYLLABLE DURATION (EXP1)
logit = 1.810 + .036*VOT - 0.021*SYLLABLE DURATION (EXP2)
logit = - 1.297 + .213*VOT - 0.021*SYLLABLE DURATION (EXP3)

Syllable duration (ms)	Estimated VOT at boundary (EXP 1)	Estimated VOT at boundary (EXP 2)	Estimated VOT at boundary (EXP 3)
75	21.9	13.5	13.5
125	23.4	18.1	18.4
175	25.0	22.7	23.3
225	26.5	27.2	28.3
275	28.0	31.8	33.2
325	29.5	36.4	38.1

References

de Jong, K. (2001a). Effects of syllable affiliation and consonant voicing on temporal adjustment in a repetitive speech-production task. *Journal of Speech, Language, and Hearing Research, 44*, 826-840.
de Jong, K. (2001b). Rate-induced resyllabification revisited. *Language and Speech, 44*, 197-216.
Lisker, L., & Abramson, A. S. (1970). The voicing dimension: Some experiments in comparative phonetics. In *Proceedings of the 6th International Congress of Phonetic Sciences* (pp. 563-567). Prague: Academia.
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*(6), 505-512.
Summerfield, Q. (1981). Articulatory rate and perceptual constancy in phonetic perception. *Journal of Experimental Psychology: Human Perception and Performance, 7*(5), 1074-1095.
Volaitis, L. E., & Miller, J. L. (1992). Phonetic prototypes: Influence of place of articulation and speaking rate on the internal structure of voicing categories. *Journal of the Acoustical Society of America, 92*(2), 723-735.