# PAYING ATTENTION TO SPEAKING RATE

*Alexander L. Francis*
*Howard C. Nusbaum*

Center for Computational Psychology
Department of Psychology
The University of Chicago

## ABSTRACT

Variability in speaking rate results in a many-to-many mapping between acoustic properties in speech and the linguistic interpretation of an utterance. In order to recognize the phonetic structure of an utterance, listeners must calibrate their phonetic decisions against the rate at which the speech was produced. This process of rate normalization is fast and effective allowing listeners to maintain phonetic constancy in spite of changes in speaking rate. Most of the research on rate normalization has investigated the sources of information used by listeners to determine the speaking rate. There is an assumption in much of this research that the normalization process is a passive, automatized filtering process that strips the effects of rate variation away from the signal prior to recognition.

The present study starts from a different perspective by assuming that speech perception is carried out by an active perceptual process that is specifically needed to address the lack of invariance problem (Nusbaum & Henly, in press). This perspective predicts that increased variability from any source, including rate variability, should increase the cognitive load during speech perception. Our results support this prediction.

## 1. SPEAKING RATE VARIABILITY

Listeners are constantly exposed to a wide range of speaking rates. Talkers differ in their characteristic speaking rates. Individual talkers will also vary their speaking rate, sometimes even within a single utterance. This variation affects the acoustic patterns of speech by restructuring the relationship between acoustic cues and phonetic categories (see Miller, 1981). The durations of cues change nonlinearly with speaking rate and there may also be changes in spectral patterns. This results in a lack of invariance in mapping acoustic cues onto phonetic categories (cf. Liberman, Cooper, Shankweiler, & Studdert-Kennedy, 1967).

At one speaking rate, the distribution of acoustic properties that denote different phonetic categories may overlap substantially; across different speaking rates this overlap becomes much more extreme (e.g., Miller & Baer, 1983). Transition durations that correspond to /w/ at a fast speaking rate might correspond to a /b/ at a slow speaking rate. This means that in order to interpret a particular acoustic property as the intended phonetic segment, listeners must know something about the speaking rate at which the utterance was produced. Listeners must calibrate their phonetic decisions against the context of the speaking rate, or in other words, they must normalize rate differences.

There is evidence that listeners derive information about speaking rate from the context of the utterance in which a particular cue is judged. The source of this may be from the overall structure of the sentence (e.g., Gordon, 1988; Miller & Grosjean, 1981) or from the structure of the syllable carrying the target property (e.g., Miller & Liberman, 1979; Newman & Sawusch, 1996; Port & Dalby, 1982). However, although research has investigated some of the sources of information listeners use in rate normalization, there are few explicit models of rate normalization. Miller and Liberman (1979) emphasized that the perceiver was affected by the *articulatory* rate rather than the physical rate represented by the acoustic durations. If rate normalization is based on the underlying articulatory characteristics of speech rather than on the acoustic properties of the signal, this suggests that listeners might normalize speaking rate differences by using an approach based on motor theory (cf. Liberman & Mattingly, 1985). This would emphasize the role of articulatory knowledge in resolving the lack of invariance problem resulting from variation in speaking rate.

By contrast, Pisoni, Carrell, and Gans (1983) questioned whether this rate normalization effect was based on articulatory knowledge. Using nonspeech analogs of Miller and Liberman's (1979) speech stimuli, they found the same pattern of results as those reported by Miller and Liberman. However since the nonspeech analogs were not heard as speech, these stimuli must have been immune from the operation of articulatory knowledge. In addition, Gordon (1988) demonstrated that a time-varying sinusoid matched to the F0 of a context sentence produced changes in classification of speech stimuli consistent with rate normalization, even though this sinusoid was not perceived as speech. These kinds of experiments suggest that rate normalization is a function of more general auditory mechanisms rather than specific to articulatory knowledge.

Regardless of whether the claim is that listeners use articulatory knowledge or general auditory mechanisms, it is interesting to note that the difference in theoretical perspective is based on the *type* of information that is being processed. This concern with the type of information or knowledge used to resolve the lack of invariance in mapping acoustic cues to phonetic categories has typically framed much of the theoretical debate in speech research. The contrast between articulatory theories and auditory theories of speech perception is well known. Furthermore, theories of talker normalization have also generally focused on this issue of determining the the sources

of information and the nature of knowledge needed to overcome the effects of talker variability.

## 2. TALKER VARIABILITY

There is much in common between the problems of rate variability and talker variability. Differences among talkers result in overlapping distributions of acoustic cues such that any particular acoustic pattern may correspond to more than one phonetic category and one phonetic category may be cued by several different acoustic patterns (e.g., Peterson & Barney, 1952). As a result, in order to interpret an acoustic pattern as the intended phonetic segment, it is necessary for a listener to know something about the vocal characteristics of the talker. Just as the listener appears to derive information about the speaking rate from the overall context of the utterance or the intrinsic structure of the syllable that is being recognized, listeners also appear to use extrinsic context and intrinsic structure to derive information about the vocal characteristics of the talker (Ainsworth, 1975; Neary, 1989).

Despite the apparent similarities in these problems, there has been little in common in the theoretical perspectives taken on rate normalization and talker normalization. Theories of talker normalization have focused specifically on the problem of vocal tract scaling (cf. Fant, 1973). For example, models of talker normalization have been proposed using extrinsic information about point vowels derived from prior context (Gerstman, 1968) or using the intrinsic structure of the vowel including F0 and F3 (Syrdal & Gopal, 1986) to scale F1 and F2 for talker-independent interpretation. Since these theories are concerned with scaling spectral patterns in the context of vocal tract differences, these mechanisms seem irrelevant to explaining the process of temporal scaling in rate normalization.

However, Nusbaum and Magnuson (in press) have argued that there is an important, equivalent computation-theoretic structure to all of the manifestations of the lack of invariance in acoustic-phonetic mapping. They distinguish, on computational theoretic grounds, the case in which one linguistic interpretation has multiple alternative acoustic instantiations and the case in which one acoustic cue maps onto multiple linguistic interpretations. The first of these can be processed by any simple deterministic finite state automaton and therefore nearly any kind of simple computational device and poses no real theoretic problem. The second case is different since it represents a basic nondeterministic relationship between patterns and the classification of those patterns which requires a different kind of computational solution. Nusbaum and Magnuson (in press) argued that this kind of computational problem cannot be resolved by changing the nature of the knowledge used in processing but instead requires a different computational control structure--an actively controlled computational mechanism (see Nusbaum & Schwab, 1986).

One hallmark of an active computational mechanism is that the listener's cognitive load should be affected by the computational demands on an active mechanism (Nusbaum & Schwab, 1986). Several studies have demonstrated that talker variability increases recognition time (Mullennix & Pisoni, 1990; Nusbaum & Morin, 1992; Summerfield & Haggard, 1975) and that this increased response time is due to increased cognitive load when there is talker variability (Mullennix & Pisoni, 1990; Nusbaum & Morin, 1992). These results suggest that listeners may use an actively controlled process to recognize speech when there is talker variability (Nusbaum & Magnuson, in press). Nusbaum and Henly (in press) have suggested that this may reflect the operation of a more general set of cognitive principles that govern perception whenever there is lack of invariance in acoustic-phonetic relationships. If this more general claim is valid, and speaking rate variability represents the same kind of lack of invariance problem, in computational terms of a nondeterministic mapping between cues and categories, we should find that rate normalization also increases the cognitive load on the listener.

## 3. ATTENTION AND RATE NORMALIZATION

Although the attentional demands of rate normalization have not been investigated explicitly, there is some evidence suggesting that rate normalization may be a passive filtering process rather than an active attention-demanding process. Miller, Green, and Schermer (1984) argued that rate normalization is obligatory and not under active attentional control since listeners could not strategically avoid using rate information from a context sentence in phonetic classification, even though they could ignore semantic information. Similarly, Miller and Dexter (1988) found that listeners could not avoid using rate information in classifying segments, but they could ignore lexical status under certain task constraints. When listeners respond quickly, they still carry out rate normalization during phonetic classification even when they do not use other sources of higher-order linguistic knowledge. However, one finding in the Miller and Dexter (1988) is at odds with the idea of a passive, automatized rate normalization mechanism: Listeners used different sources of rate information under different task constraints.

This kind of flexibility in shifting attention to different sources of information is more consistent with an active perceptual mechanism (see Nusbaum & Schwab, 1986). Although these studies are suggestive, they were not designed test specifically the question of whether rate normalization is a passive, automatized process or an active, controlled process. A more direct test of this question would assess whether rate variability increases the listener's cognitive load in the same way as has been observed with talker variability (Nusbaum & Morin, 1992).

If rate normalization is carried out by an active process, we would expect rate variability to slow down phoneme

recognition. Furthermore, if this slowing is due to an increase in cognitive load when there is rate variability, we would expect manipulating cognitive load through a secondary task (see Nusbaum & Morin, 1992) should interact with rate variability. Under high cognitive load, when there are few resources for speech perception, rate variability should slow down recognition even more than under low cognitive load, whereas when there is no rate variability, cognitive load manipulations should have little effect on recognition time. On the other hand, if rate normalization is carried out by automatized processes, manipulating cognitive load with a secondary task should not differentially affect speech perception with there is rate variability and when there is no rate variability.

## 3.1. Method

We used the speeded target monitoring procedure that was used by Nusbaum and Morin (1992) to investigate the effects oftalker variability. Subjects would see a target phoneme displayed on a computer screen and then would listen for that target throughout a sequence of 16 syllables. Four of these utterances were targets randomly located in the sequence and subjects were told to respond as quickly and accurately as possible by pressing a response key when they recognized the target.

The stimuli consisted of syllables produced by a text-to-speech system (DECtalk) at two different rates, 250 words per minute (wpm) corresponding to a very fast but still intelligible rate and 150 wpm, a slow, but still natural-sounding rate. For any particular trial, the target was selected from the set [ba], [pa], [wa], [ga], [ka], [ya]. Distractors were selected from the set [ta], [da], [ma], [na], [fa], [va], and [za] and any of the members of the target set that were not targets on that trial.

Twenty-four subjects were assigned to two different groups. For one group of subjects, trials were blocked by speaking rate so on any particular trial all syllables were produced at the same rate. For a second group of subjects, each trial consisted of targets and distractors from both speaking rates mixed randomly. Both groups also received a secondary digit preload memory task designed to manipulate the capacity available for carrying out the speech task (see Nusbaum & Morin, 1992). In the low-load condition, subjects were given a single two-digit number to remember while performing the target recognition task. In the high-load condition, a list of three two-digit numbers was presented before each monitoring trial.

On each trial, subjects would see the numbers to be remembered displayed on a computer screen. Following this display, they would see the target phoneme. They would then monitor a sequence of 16 syllables, listeing for the four stimuli that matched the target. After each monitoring trial was complete, subjects were prompted to recall the numbers from the digit preload task.

## 3.2. Results

Overall, accuracy was affected by load, for both hit rates and false alarm rates, but accuracy was not affected by speaking rate variability. The mean hit rates were over .9 in all conditions and the false alarm rates were under .1. Recall of the visually presented numbers in the digit preload task was affected only by the length of the list of digits. Rate variability did not affect recall which varied from around 93% correct for one two-digit number to 77% correct for three two-digit numbers.

As in our previous research on talker variability, recognition times were longer in the mixed-rate condition (478 msec) than in the blocked-rate condition (450 msec) although this did not reach significance, due to a between-subjects comparison, $F(1,22) = 3.48$, *n.s.* Subjects were significantly slower to recognize phoneme targets in the high cognitive load condition (473 msec) than in the low cognitive load condition (455 msec), $F(1,22) = 4.86$, $p < .05$. Finally, there was a significant interaction between cognitive load and rate variability, $F(1,22) = 5.21$, $p < .05$. When there was no rate variability, there was no difference in recognition times between the low cognitive load condition (451 msec) and the high cognitive load condition (450 msec). However when there was a mix of speaking rates presented on each trial, recognition time was longer in the high cognitive load condition (497 msec) than in the low cognitive load condition (459 msec). This is the pattern of results predicted if rate normalization increases the cognitive load of the listener due to active processing.

# 4. CONCLUSIONS

The present results demonstrate that rate variability increases the cognitive load of the listener. This finding argues against passive, automatized explanations of rate normalization and demonstrates a similarity in the perceptual processing of talker variability and rate variability. Nusbaum and Henly (in press) have argued that a common set of cognitive principles could explain how phonetic constancy is achieved in spite of the lack of invariance problem. These principles assume that speech perception is carried out by an active mechanism rather than a collection of specialized passive filtering devices (see Nusbaum & Schwab, 1986). The claim is that whenever there is an increase in variability, whether it is talker variability, rate variability, or contextual variability, recognition time will increase specifically because of an increase in the number of alternative linguistic interpretations of acoustic patterns. This increase in recognition time is due to the increased cognitive load incurred as listeners attempt to test among these alternatives. This testing process is carried out by shifting attention to acoustic properties or other sources of information or knowledge that discriminate among the alternatives. The present results support this view and suggest that further studies should be carried out to investigate the operation of attention in rate normalization.

## 5. ACKNOWLEDGEMENTS

This research was supported, in part, by a grant from the Social Sciences Division Research Fund at the University of Chicago.

## 6. REFERENCES

1. Ainsworth, W. Intrinsic and extrinsic factors in vowel judgments. In G. Fant & M. Tatham (Eds.), *Auditory analysis and perception of speech*. London: Academic Press, 103-113, 1975.

2. Fant, G. *Speech sounds and features*. Cambridge: MIT Press, 1973.

3. Gerstman, L. J. Classification of self-normalized vowels, *IEEE Transactions on Audio Electroacoustics, AU-16*, 78-80, 1968.

4. Gordon, P. C. Induction of rate-dependent processing by coarse-grained aspects of speech. *Perception & Psychophysics, 43*, 137-146, 1988.

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

6. Liberman, A. M., & Mattingly, I. G. The motor theory of speech perception revised. *Cognition, 21*, 1-36, 1985.

7. Miller, J. L. Effects of speaking rate on segmental distinctions. In P. D. Eimas & J. L. Miller (Eds.), *Perspectives on the study of speech*, LEA, Hillsdale: NJ, 39-74.

8. Miller, J. L., & Baer, T. Some effects of speaking rate on the production of /b/ and /w/. *Journal of the Acoustical Society of America, 73*, 1751-1755.

9. Miller, J. L., & Dexter, E. R. Effects of speaking rate and lexical status on phonetic perception. *Journal of Experimental Psychology: Human Perception and Performance, 14*, 360-378.

10. Miller, J. L., Green, K., & Schermer, T. M. A distinction between the effects of sentential speaking rate and semantic congruity on word identification. *Perception & Psychophysics, 36*, 329-337, 1984.

11. Miller, J. L., & Grosjean, F. How the components of speaking rate influence perception of phonetic segments. *Journal of Experimental Psychology: Human Perception and Performance, 7*, 208-215.

12. Miller, J. L., & Liberman, A. M. Some effects of later-occurring information on the perception of stop consonant and semivowel. *Perception & Psychophysics, 25*, 457-465, 1979.

13. Mullennix, J. W., & Pisoni, D. B. Stimulus variability and processing dependencies in speech perception. *Perception & Psychophysics, 47*, 379-380, 1990.

14. Nearey, T.M. Static, dynamic, and relational properties in vowel perception. *Journal of the Acoustic Society of America, 85*, 2088-2113, 1989.

15. Newman, R. S., & Sawusch, J. R. Perceptual normalization for speaking rate: Effects of temporal distance. *Perception & Psychophysics, 58*, 540-560, 1996.

16. Nusbaum, H. C., & Henly, A. S. Understanding speech perception from the perspective of cognitive psychology. In J. Charles-Luce, P. A. Luce, & J. R. Sawusch (Eds.), *Theories in spoken language: Perception, production, and development.* Norwood, NJ: Ablex Publishing, in press.

17. Nusbaum, H. C., & Magnuson, J. S. Talker normalization: Phonetic constancy as a cognitive process. In K.A. Johnson & J.W. Mullennix (Eds.), *Talker variability and speech processing*. Academic Press, New York, in press.

18. Nusbaum, H. C., & Morin, T. M. Paying attention to differences among talkers. In Y. Tohkura, Y. Sagisaka, & E. Vatikiotis-Bateson (Eds.), *Speech perception, production, and linguistic structure*. Tokyo: OHM Publishing Company, 113-134, 1992.

19. Nusbaum, H.C., & Schwab, E.C. The role of attention and active processing in speech perception. In E.C. Schwab & H.C. Nusbaum (Eds.), *Pattern recognition by humans and machines: Vol. 1. Speech perception*. San Diego: Academic Press, 113-157, 1986.

20. Peterson, G., & Barney, H. Control methods used in a study of the vowels. *Journal of the Acoustical Society of America, 24*, 175-184, 1952.

21. Pisoni, D. B., Carrell, T. D., & Gans, S. J. Perception of the duration of rapid spectrum changes in speech and nonspeech signals. *Perception & Psychophysics, 34*, 314-322, 1983.

22. Port, R. F., & Dalby, J. Consonant/vowel ratio as a cue for voicing in English. *Perception & Psychophysics, 32*, 141-152, 1982.

23. Summerfield, Q., & Haggard, M. Vocal tract normalization as demonstrated by reaction times. In G. Fant & M. Tatham (Eds.), *Auditory analysis and perception of speech*. London: Academic Press, 115-141, 1975.

24. Syrdal, A.K. & Gopal, H.S. (1986). A perceptual model of vowel recognition based on the auditory representation of American English vowels. *Journal of the Acoustic Society of America, 79*, 1086-1100, 1986.