Journal of Phonetics (1999) **27,** 107–123
Article No. jpho.1999.0087
Available online at http://www.idealibrary.com on IDEAL®

Auditory accounts of temporal factors in the perception of Norwegian disyllables and speech analogs

Wim A. van Dommelen

Department of Linguistics, Norwegian University of Science and Technology, N-7491 Trondheim, Norway

Received 23rd December 1997, revised 5th October 1998, and accepted 27 January 1999

The first part of this study investigates temporal factors in the perception of V:C vs. VC: rhymes in Norwegian. To that aim, a listening test was performed with stimuli of the form /mV:Cə/ vs. /mVC:ə/with varying vowel, consonant closure and schwa duration. For a group of native speakers, boundaries in the perception of /V:/-/V/were established. Both longer consonantal closure and schwa duration appeared to cause a perceptual shortening of the vowel. These results were interpreted by appealing to Kingston & Diehl's (1994) duration ratio hypothesis, which states that the durations of a vowel and a following consonant are mutually enhancing acoustic cues. In two listening tests on speech analogs, listeners judged the duration of the first tone in tone-gap-tone sequences. Mimicking the speech stimuli, these sequences featured varying gap and second tone durations. Longer durations of these two signal parts turned out to perceptually lengthen first tone duration. The divergent results from the speech as opposed to the non-speech stimuli were explained by assuming different perceptual strategies for the two types of signals. While in speech the listeners can rely on well-established temporal patterns, such a framework is absent in the case of non-speech.

1. Introduction

The temporal organization of human speech is characterized by a complex of rule-governed patterns. An issue from this domain that has been investigated for a number of different languages is the durational patterning involved in the phonological opposition of postvocalic voiced-voiceless obstruents. In many languages, voiced consonants have a shorter duration than their voiceless counterparts. Moreover, vowels preceding voiced consonants tend to be longer than vowels preceding voiceless consonants. During the last decade attempts have been made to give a perception-oriented account of temporal regularities like those found in the production of the voiced-voiceless distinction.

1.1. *The auditory enhancement hypothesis*

Kluender, Diehl, and Wright (1988) investigated this question by using /aba/-/apa/ stimuli and square-wave analogs temporally mimicking these speech stimuli. In both

0095-4470/99/010107 + 17 $30.00/0

speech and non-speech conditions, the initial segment (vowel or tone) of a stimulus was either long or short, while the medial silent interval varied in 10 ms steps from 20 ms to 110 ms. Following training runs with endpoint stimuli, listeners classified test stimuli "corresponding to the endpoint that each stimulus sounded more like" (p. 159). The results showed parallel patterns for the speech and the non-speech conditions. In both cases a long, as opposed to a short, initial segment caused a statistically significant increase of short-gap responses. Kluender et al. interpreted this outcome as supporting their *auditory enhancement hypothesis*, stating that vowel duration and consonantal closure duration are acoustic cues that have mutually reinforcing effects: a long vowel makes a phonologically voiced, brief consonant appear shorter and, consequently, perceptually more "voiced". Similarly, a shorter vowel will cause a following phonologically voiceless, long consonant to be perceived as longer and, hence, more "voiceless". Given the parallel between the speech and non-speech results, the effect was considered to be attributable to auditory factors not specific to speech.

Investigating Kluender *et al.*'s (1988) auditory enhancement hypothesis in a carefully designed study, Fowler (1992) presented counterevidence advocating the existence of assimilation, rather than contrast effects. In her first experiment, one group of listeners had to classify the silent interval in synthetic VCə disyllables with varying vowel and consonant durations as long or short. Another group of listeners gave similar judgments on vowel duration. For both groups of listeners assimilation effects were found; longer vowels increased the number of "long closure" judgments, while longer silent intervals increased "long vowel" judgments. However, for a third group of subjects, who classified the medial consonant of the same stimuli as "b" or "p", longer vowels enhanced the number of "b" judgments. Without the data on estimated closure duration, this latter result could presumably have been explained by postulating a contrast effect with perceptual shortening of the closure due to increasing vowel durations. In Fowler's second experiment, listeners were presented with square-wave analogs of the VCə stimuli from the first experiment. Here, for listeners judging the square-wave stimuli as /apa/-/aba/ analogs, first tone duration failed to systematically bias the number of /p/-/b/ responses. In contrast, for subjects who had been told that they would listen to nonsense sounds, a longer first tone caused an increase of "long gap" responses. The latter result is the reverse of the contrast effects found by Kluender *et al.*

It should be pointed out that Fowler (1990, 1991, 1992) argued that responses to speech and non-speech stimuli in fact are not comparable. According to the direct realist view adopted by Fowler, our perceptual system will always try to recover distal causes of incoming signals. In decoding speech, the source is a human vocal tract and the signals will be interpreted accordingly. However, as far as the perception of non-speech is concerned, it may be impossible to account for the specific mechanisms involved. The question is whether response patterns to non-speech reflect the specific way our auditory system processes acoustic signals or, alternatively, whether such patterns are attributable to *ad hoc* distal-source interpretations by the auditory system. It could be argued that applying this direct realist approach to the auditory processing of duration parameters is questionable. Duration can be considered to be a basic property of any acoustic signal, both speech and non-speech, its perceptual assessment representing one of several stages in the processing of acoustic stimuli. It is conceivable that duration together with other basic properties like fundamental frequency and amplitude are processed in a manner independent of the identity of the distal cause. For example, it seems plausible that a listener will perceive the duration of a signal regardless of whether it is identified as, for

example, a musical tone, a computer bleep, or a voiceless fricative. For an elaborate discussion of the interpretability of speech/non-speech comparisons the reader is referred to Diehl, Walsh, and Kluender (1991) and Fowler (1990, 1991, 1992).

1.2. *The duration ratio hypothesis*

In a more recent paper, Kingston and Diehl (1994) gave a new account of durational contrast phenomena in speech. Rather than focusing on duration properties of single vowels and consonants, the authors considered the C/V duration ratio to be one of the relevant contrastive perceptual properties of the phonological voiced-voiceless contrast. They stated that "a short consonant and a long preceding vowel are mutually enhancing in that they both contribute to the relatively small C/V duration characteristic of [+voice] consonants" (p. 442). Further, it was argued that another feature of [voice], viz. phonetic voicing during a consonantal closure, apart from contributing to the voicing distinction as a separate cue, makes the C/V duration ratio more distinctive: the voicing causes not only a perceptual lengthening of the vowel but also an apparent shortening of the closure.

Diehl and Castleman (1996) showed that for the affricate/fricative distinction, such as in /tʃ/-/ʃ/, the ratio of silence interval to fricative duration is an effective measure to separate the two categories. Like the C/V duration ratio for intervocalic consonants, the paper postulates mutually enhancing effects to be involved in the silence interval/fricative duration property.

However, neither Kingston and Diehl (1994) nor Diehl and Castleman (1996) formulated explicit hypotheses concerning the issue of auditorily based durational contrast effects: will a physically shorter consonant duration yield a more distinct [+voice] percept not only by virtue of a decreased C/V duration ratio but also due to a contrastive lengthening of the vowel? The latter effect would cause the C/V duration ratio to be shifted even more towards [+voice].

The lack of further investigation of durational contrast effects in the studies just mentioned can readily be explained by the fact that the authors abandoned the auditory contrast hypothesis to replace it by the durational ratio hypothesis referred to above. To the best of my knowledge, this shift has not been stated explicitly in the literature yet, but it has been expressed informally by reviewer Randy Diehl. For the sake of discussion, the present paper will take both the auditory enhancement hypothesis and the duration ratio hypothesis as a point of departure, with special emphasis on the latter.

1.3. *The present study*

The first part of the present study investigates timing patterns of intervocalic consonants in Norwegian CVCə disyllables. Through tests on non-speech stimuli the second part tries to account for the response patterns found for the speech stimuli in the light of the hypotheses presented above.

In Norwegian there is a phonological short-long vowel length distinction that is accompanied by phonetic durational differences in a following obstruent: a long vowel is followed by a phonetically short consonant, while consonantal duration is longer after a short vowel (Fintoft, 1961). Norwegian has only V:C and VC: patterns; minimal pairs of the form VC vs. VC: or V: C vs. V: C: do not occur (such combinations being found in only a relatively few languages, such as Finnish). The V: C vs. VC: opposition was chosen

as a paradigm for the present study because of its high frequency of occurrence in modern Norwegian. However, the phonological opposition of postvocalic voiced-voiceless obstruents also occurs.

For the first part of this study, a listening test on the perception of CVCə disyllables was run. The stimuli in the listening test differed in vowel duration and closure duration of the intervocalic consonant such that the interaction of both segmental durations in perception could be investigated. In view of the patterns found in production, a shorter closure was postulated to bias perception towards a "long vowel". In addition, the duration of the schwa in the CVCə words was varied. In contrast with the phonologically conditioned differences in temporal organization, varying schwa durations can occur in production as a function of different speech rates (e.g., Port, 1981; Crystal & House, 1990; Fourakis, 1991). Perceptual studies have shown that durational criteria in the estimation of speech segments depend on perceived speech rate (Repp, Liberman, Eccardt & Pesetsky, 1978; Miller & Liberman, 1979; Port, 1979; Miller, 1981; Wayland, Miller & Volaitis, 1994; Newman & Sawusch, 1996; Miller, O'Rourke & Volaitis, 1997). Therefore, it was speculated that a shortened schwa in the CVCə stimuli would be interpreted by a listener as an increased speech tempo and that, in turn, this higher speech rate would cause the preceding vowel to be perceived as longer. In this connection it should be noted that in the perception of closure duration, speech rate effects can be confounded with variation of duration conditioned by vowel quantity. Similar to the effect of a shortened schwa, a shorter closure duration might be interpreted as a higher speech rate and so give rise to perceptual lengthening of the preceding vowel.

2. Perception of speech stimuli

As speech material the Norwegian words "mate" ([ma:tə]; "to feed") and "matte" ([mat:ə]; "mat", adj. plural) were chosen. Starting from a "mate" token, manipulations of three speech segments were performed, all of them in the temporal domain: (1) the long vowel was shortened in small steps so as to create a vowel duration continuum, (2) occlusion duration of the medial consonant was varied in two steps ("short"/"long") and, finally, (3) the original schwa was shortened so as to create the conditions "long"/"short" schwa.

In the listening test, an interactive up-down method was used to obtain individual subjects' boundaries in /a:/-/a/ perception (see 2.1.3). A classification of stimuli as either of two words ("mate"/matte") fairly well parallels everyday speech perception and was preferred to metalinguistic tasks as the "short gap"/"long gap" judgments for the "aba"/"apa" stimuli used in Kluender *et al.* (1988). As has been shown by van Dommelen (1997), the results of the up-down method are comparable with those acquired with the more frequently used method of constants.

2.1. *Methods*

2.1.1. *Stimulus materials*

A trained male phonetician who is a native speaker of the eastern dialect of Norwegian was recorded producing the Norwegian words "mate" and "matte". Recordings were made in the studio at the Linguistics Department using a Milab LSR 1000 microphone

TABLE I. Segment durations in ms used in the "mate/matte" listening test. The different /t/ occlusion and schwa conditions were combined orthogonally with the vowel duration continuum

vowel	/t/ occlusion	schwa
214 to 80	170 ("short") 250 ("long")	72 ("short") 143 ("long")

and a Fostex D-10 digital recorder. The recordings were stored, via Sound Designer software, as Signalyze™ files with a sampling rate of 44.1 kHz and 16 bit resolution. All manipulations were performed with Signalyze™ (Keller, 1994) and were restricted to editing the speech waveform in the time domain. Specifically, vowel duration was manipulated with the *AutoEdit* facility which modifies segment durations by adding or excising pitch periods at zero-crossings while attempting to preserve the original fundamental frequency contour and formant structure.

The point of departure in stimulus generation was a token of "mate". The original /a:/vowel (242 ms) was shortened by excising single periods, resulting in steps of approximately 8 ms each. In the listening tests, only a vowel duration continuum varying in 17 steps from 214 ms to 80 ms was used (Table I), since in pilot tests this range appeared to be sufficient to allow unambiguous "mate" and "matte" responses. Based on the speaker's natural productions of "mate" and "matte", two different /t/ occlusions were created with durations of 170 ms and 250 ms, respectively. Furthermore, both the original schwa (143 ms) in "mate" and the same segment shortened by half (to 72 ms) was used. Using these segments four different stimulus continua were generated, all having the same range of vowel durations and having: (1) short occlusion/short schwa (SS), (2) long occlusion/long schwa (LL), (3) short occlusion/long schwa (SL), (4) long occlusion/short schwa (LS).

2.1.2. *Subjects*

A total of 45 subjects aged 19–28 participated in the listening tests. Eight of them were students of Linguistics taking part on a voluntary basis. The other listeners were university students from different disciplines and were paid for their participation.

2.1.3. *Procedure*

Stimuli were presented over two Genelec 1031A loudspeakers to individual listeners seated in the studio of the Linguistics Department. Using the CSRE (Computerized Speech Research Environment) system, the listeners' task was to identify whether the word they heard was "mate" or "matte" and click on the corresponding word displayed on a computer screen.

The test design was an up-down method with variable step size similar to procedures described by Levitt (1971). An example of this method is visually illustrated in Fig. 1. Choosing one of the alternatives "mate/matte" prompts the program to present a stimulus that lies a certain number of steps away from the previous one towards the other end of the continuum. Consecutively clicking the same alternative gives rise to a stimulus still further away in that direction. Step-size is user-definable and, in addition, there can be

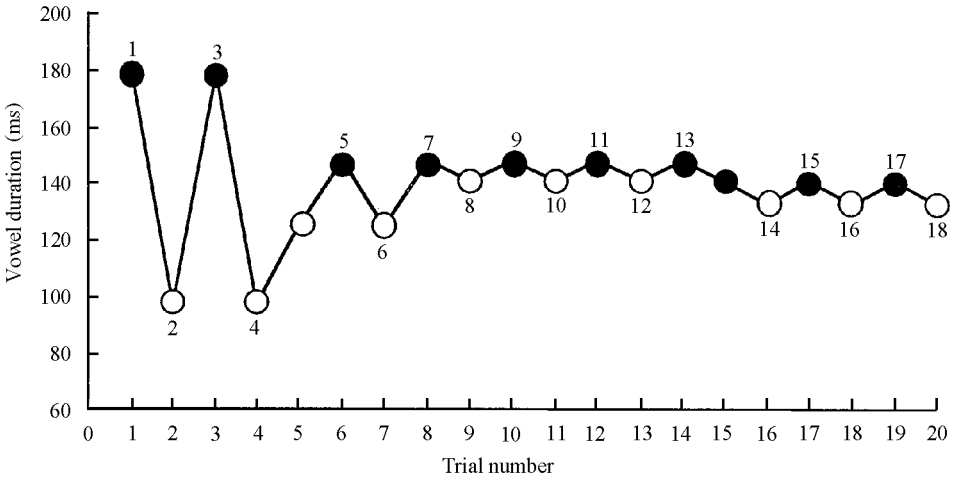

Figure 1. Reactions of a typical subject to "mate/matte" stimuli with short occlusion and short schwa. The 18 reversals are indicated (the first response being counted as a reversal). Filled circles indicate "mate" responses, open circles "matte" responses. For this subject, vowel durations for trial numbers 9–20 were used for calculation of the phoneme boundary (140.8 ms).

three different step-sizes within one experimental run. The program stops automatically after a certain, user-definable number of reversals (set at 18 in the present experiments). The present experiments had as a first step-size 10 stimuli (the current program's maximum) and started with the presentation of perceptually unambiguous tokens of "mate/matte" with vowel durations of 179 ms (e.g., trial 1 in Fig. 1) and 98 ms (e.g., trial 2 in Fig. 1), respectively, followed by a second step-size of three stimuli (e.g., between trials 4 and 5 in Fig. 1). The smallest step-size was set at one stimulus (corresponding to the duration of one glottal period; between each of the trials from 8 to 20). The program switched from the largest to the medium step size after four reversals (e.g., reversal 4 in Fig. 1) and from there to the smallest step size after three further reversals (e.g., reversal 7 in Fig. 1).

To investigate the influence of possible learning effects, the four experimental conditions (cf. 2.1.1) were presented in three different orders: SL-LL-SS-LS; (2) LL-SL-LS-SS; (3) SS-LS-SL-LL. The 45 listeners were evenly divided across the three orders of presentation on a randomized basis.

2.1.4. *Statistical evaluation*

Phoneme boundaries were calculated as the arithmetic mean of the vowel duration values chosen by the listener with the smallest step-size condition, starting from the second reversal within this region. In the example given in Fig. 1 the vowel duration values belonging to reversals 8–18 (trial numbers 9–20) formed the basis for this calculation. Due to the interactive procedure followed in determining the boundaries, the number of values in this calculation varies. The individual phoneme boundaries determined in this way were treated statistically with a repeated measures ANOVA with order of presentation, occlusion duration, and schwa duration as factors.

TABLE II. Analysis of variance for perception of "mate/matte": Effect of the factors order of presentation, occlusion duration, and schwa duration on /a:/-/a/ boundary position

Source of variation	*F*	*df*	*p*
Order of presentation	0.00	2, 42	0.9982
Occlusion duration	176.31	1, 42	< 0.0001
Schwa duration	85.35	1, 42	< 0.0001
Order × occlusion duration	1.24	2, 42	0.3007
Order × schwa duration	0.77	2, 42	0.4688
Occlusion × schwa duration	27.07	1, 42	< 0.0001

2.2. *Results*

As is apparent in the results of the ANOVA (Table II), the order of presentation of the four experimental conditions did not have a reliable impact on boundary position; consequently, the data of the three groups of subjects have been pooled for the discussion that follows.

Figure 2 presents the results for the perception of "mate/matte" pooled across 45 listeners. The top part of the figure displays the mean phoneme boundary locations measured for the four experimental conditions. For comparison with results to be presented in section 3 below, it should be noted that a boundary shift toward *longer* durations corresponds to an effective perceptual *shortening* of the vowel, and vice versa. Thus the change in effective duration between two conditions can be estimated by the magnitude of the boundary shift between those conditions, but with a reversed sign. The lower panel of Fig. 2 depicts such estimated changes in effective perceived duration, using the short occlusion, long schwa condition as a reference.

As can be seen from Fig. 2, the hypotheses concerning perception of speech stimuli formulated above (cf. 1.3) are confirmed by the data. For both the long and short schwa duration condition, a long occlusion duration appears to shorten the perceived duration of the preceding vowel. The shortening effect amounts to $144.4 - 136.2 = 8.2$ ms and 135.3–116.0 = 19.3 ms, respectively, and is statistically highly significant (see Table II). Similarly, the boundary location is affected by the schwa duration: A shorter schwa generally gives rise to a perceptual lengthening of the vowel (shifts of $136.2 - 116.0 = 20.2$ ms and $144.4 - 135.3 = 9.1$ ms, respectively). The stronger influence of occlusion duration for the short schwa than for the long schwa condition is reflected in the significant interaction between the two factors of occlusion duration and schwa duration. Pooled over the two schwa conditions, the shift of perceived vowel duration due to varying occlusion duration amounts to $(8.2 + 19.3)/2 = 13.8$ ms, which is about as strong as the shift of $(20.2 + 9.1)/2 = 14.7$ ms caused by the pooled influence of varying schwa duration. Given that the ranges of manipulated occlusion and schwa duration were similar (80 ms and 71 ms, respectively; cf. 2.1.1), it is concluded that the occlusion and schwa exert similar influence on the perceived vowel duration.

So far, the picture that has emerged for the perception of the speech material is coherent: lengthening of consonantal occlusion duration caused a perceived shortening of the preceding vowel, while a shorter schwa made vowel duration appear longer. The

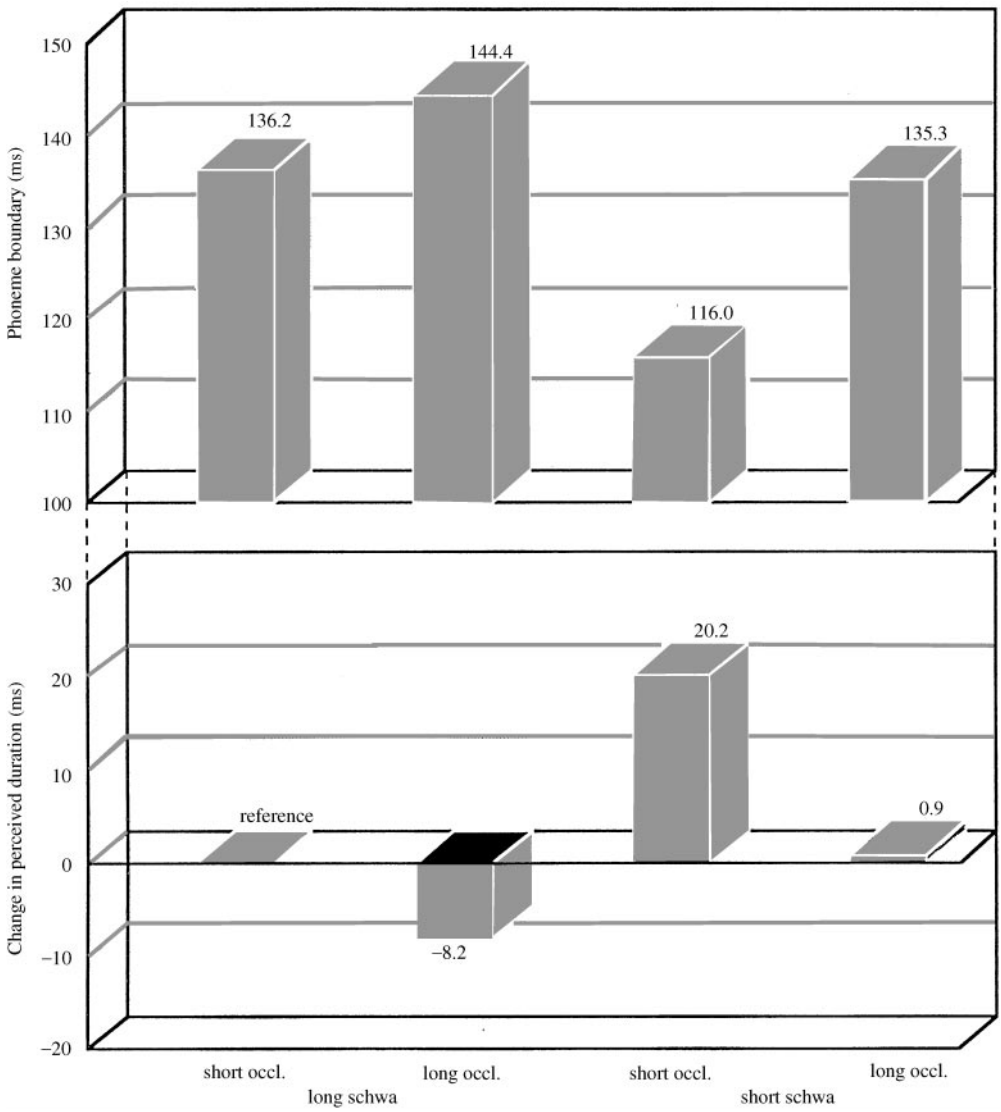

Figure 2. Top: Mean /a:/-/a/ boundaries in ms for perception of "mate/matte" with short vs. long /t/ occlusion and short vs. long schwa. $n = 45$ listeners. Bottom: Mean changes in perceived vowel duration. The values indicate shifts in boundary location, a boundary (top figure) towards larger values representing perceptual shortening of the vowel and vice versa. The stimulus condition short occlusion/long schwa was taken as a reference.

next step will be to investigate whether the results can be explained by either the auditory contrast or the duration ratio hypothesis. With that aim, the second listening test examined perceived duration of speech analogs with temporal patterns corresponding to those used for "mate/matte".

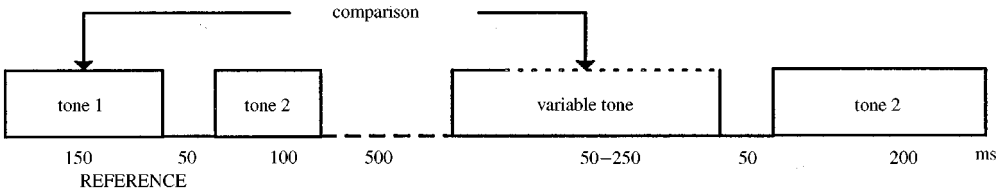

Figure 3. Example of stimulus configuration for the reference G50S (gap duration of 50 ms and short second tone duration of 100 ms). Listeners adjusted the variable tone after comparison with the reference's first tone, which always had a duration of 150 ms.

3. Perception of non-speech stimuli

From the viewpoint of the listeners, the tasks in experiments using non-speech stimuli often may seem to be more difficult and require training prior to the actual tests. Preceding their labeling task, subjects in the experiments described in Kluender *et al.* (1988) trained with endpoint stimuli (cf. 1.1). One group of listeners in Fowler's (1992) investigation was trained to identify speech analogs as having a short gap (A) or a long gap (B). Following training, they judged the test stimuli according to whether they sounded more like A or more like B. Another group labeled gap durations as short (S) or long (L). In contrast, the present study aimed to measure perceived duration more directly by means of a listening task using tone duration adjustments.

A typical signal used in this experiment consisted of two tones intermediated by a gap, thus roughly mimicking the alternation of voiced (nasal; vowels) and voiceless (consonantal closure) segments of the CVCə in the speech material. Six so-called references were created by orthogonally combining three different gap durations with two different durations of the second one. All six references had the same first tone duration. The first experiment with non-speech materials (3.1.2.1) used four references, and the second experiment (3.1.2.2) used six references.

In all cases, perceived duration of the reference's first tone was measured through the procedure represented in Fig. 3. For each trial a reference was followed by another tone-gap-tone sequence. While the gap and the second tone of this signal always had a fixed duration, the duration of the first tone varied from relatively short to relatively long. In the listening tests, the subjects' task was to adjust this duration to match perceived duration of the reference's first tone using an up-down method as described above (cf. 2.1.3).

Predictions about the perceptual influence of varying temporal structure on the perceived first tone duration are more complex than for the speech material. To make the formulation of hypotheses more transparent, the terms *assimilation* and *contrast* will be used in the following discussion. For the time being, these terms only denote surface phenomena without necessarily postulating perceptual reality. First, according to the assimilation hypothesis, a longer gap duration is expected to result in a perceptually longer first tone. The interaction of these two segments represent a straightforward case of *contact assimilation.* But what kind of assimilation can be postulated between the first and second tone? It seems reasonable to assume that listeners will primarily compare

signals of the same type, for example, comparing first and second tones of a reference and ignoring the gap. Such a hypothesis of *distant assimilation* predicts that a first tone would be perceptually lengthened when the second tone is relatively long. Note, however, that it is also possible to postulate a two-stage contact assimilation mechanism by which a long second tone would give rise to a perceptually longer gap, which in turn lengthens the perceived duration of the first tone. In this way the predictions made by the contact assimilation and the distant assimilation hypothesis coincide.

What can we expect if we adopt the contrast hypothesis? As far as the relation between a first tone and the succeeding gap is concerned, this hypothesis predicts that a longer gap duration in the reference will subjectively shorten the duration of the first tone in the reference. Predictions about the impact of the duration of the second tone, however, are less straightforward and depend on whether we assume contact or distant contrast. The latter type predicts that a longer second tone will directly result in the first tone being perceived as shorter. Contact contrast, on the other hand, would predict that the gap would be perceived as shorter due to a longer second tone, and in turn, this subjectively shorter gap would result in the first tone being perceived as longer. Thus the predictions for the effect of second tone duration diverge.

At this point it can be noted that confirmation of the contrast hypothesis for gap duration in combination with a distant contrast effect for second tone duration could provide an explanation for the perceived duration observed in the speech stimuli. In that case, changes in perceived vowel duration due to varying occlusion and schwa duration (section 2.2) might be explained by assuming auditory enhancement due to the two contrast effects.

3.1. *Methods*

3.1.1. *Stimulus materials*

In designing the non-speech stimuli, no special attempt was made to replicate the signals used in Kluender *et al.* (1998) and Fowler (1992). Specifically, the complex wave generated for the present purposes had fewer (four) partials with fixed frequencies, rather than having a fundamental frequency contour falling/rising in the vicinity of the gap as in the studies mentioned above. The reasoning behind the design chosen here was that focusing on durational variations without mimicking the speech stimuli in further detail would create sufficient conditions to test the present hypotheses.

In order to obtain the complex wave needed for the stimulus materials, sine waves were generated with a Kyoritsu K-128 audio generator. The tones had frequencies of 300, 500, 700, and 900 Hz. Adding these single components with relative amplitudes of 0 dB, 0 dB, 6 dB, and 12 dB, respectively, with the aid of Signalyze™ resulted in a complex wave form which was pleasant to listen to. The resulting tone's amplitude was held constant, apart from 10-ms rise and decay slopes.

Using this wave form, two types of signals were created, the first of which served as a reference. A typical reference consisted of a fixed 150-ms tone followed by a gap, which in turn was followed by a second tone. In total, six different reference signals were created having gap durations of 50 ms, 100 ms, or 150 ms, while the tone following the gap was either 100 ms or 200 ms long. Note that all six references had a first tone duration of 150 ms.

The second type of signal was designed to be compared with the references and consisted of a tone of varying duration followed by a fixed 50-ms gap and a fixed 200-ms tone. The duration of the first tone varied in 10-ms steps from 50 ms to 250 ms.

3.1.2. *Procedure*

Fig. 3 depicts the configuration of the signals which served as stimuli in the listening tests. In each test, the reference and the variable second signal were separated by a pause of 500 ms duration. The listeners' task was to compare the first tone of the second signal with the first tone of the reference and to decide whether it sounded shorter or longer than in the reference. In order to clarify their task, subjects were presented with a schematic drawing of the stimulus configuration similar to Fig. 3, but without the durations indicated. Listeners were especially encouraged to focus on the first tone of each of the pairs and to take no special notice of other signal parts. In addition to the oral explanation, the experimenter demonstrated the test procedure for a reference with a gap duration of 50 ms and a second tone of 200 ms, intentionally including examples from both ends of the tone duration continuum. No information about the different reference conditions was given.

In the listening tests, the same up-down method as applied for the speech stimuli was followed (cf. 2.1.3), using the response alternatives "too short" and "too long", respectively. In all cases, the program started switching between tone durations of 210 ms and 110 ms. The medium step size was set at three stimuli (30 ms), the finest at one stimulus (10 ms).

Two-step gap duration. In the first experiment with non-speech material, comparisons were performed using four different reference signals having gap (G) durations of 50 ms and 100 ms, and a second tone of either 200 ms (L = long) and 100 ms (S = short), respectively. Tests were run in the order G50L, G100L, G50S, and G100S.

Three-step gap duration. In order to get more information on the role of gap duration and possible effects of presentation order, a second experiment was run with a wider range of conditions. The test included reference signals with a longer gap duration (150 ms) and three different orders of presentation conditions. Using the coding from the preceding paragraph, these presentation orders were (1) G50L, G100L, G150L; G50S, G100S, G150S, (2) G150S, G100S, G50S; G150L, G100L, G50L, and (3) randomized.

3.1.3. *Subjects*

A group of ten subjects participated in the first experiment (3.1.2.1). Apart from one older (61 years) subject, their ages ranged between 21 and 30. In the second experiment (3.1.2.2), 45 listeners between 19 and 30 years old were evenly and randomly distributed among the three order of presentation conditions.

3.1.4. *Statistical evaluation*

Following the same evaluation procedures as for the speech stimuli (2.1.4), boundaries between durations for "too short" and "too long" responses were calculated for each

reference condition. These boundary values represent adjusted tone durations. Statistical calculations were similar to those described for the speech material.

3.2. *Results and discussion*

3.2.1. *Two-step gap duration*

Figure 4 displays mean adjusted tone durations for the two-step gap duration experiment. Note that, unlike the top panel of Fig. 2, which it superficially resembles, Fig. 4 represents perceived (matched) vowel duration directly. (Hence, changes in bar height across conditions in Fig. 4 should be compared to changes in effective perceived durations in the *bottom* panel of Fig. 2.) In evaluating the listeners' behavior, the reference condition represented by the left-most bar in Fig. 4 can be taken as an anchor point. For this condition, the durations of both gap (50 ms) and second tone (200 ms) in the reference are the same as the corresponding parts of the signal-to-be-adjusted. Mean adjusted duration for this condition amounts to 150.7 ms, which is very close to the reference's first tone duration of 150 ms. In comparison, a longer gap of 100 ms in the reference gave rise to longer adjusted durations. This is true for a long and, to a lesser degree, short second tone duration (lengthening effects of $159.7 - 150.7 = 9.0$ ms and $144.1 - 139.5 = 4.6$ ms, respectively). Based on an ANOVA with gap duration and second tone duration as factors, the effect of varying gap duration is statistically significant (cf. Table III). A higher level of significance is found for the influence of the second tone, suggesting that shorter second tone durations cause a perceptual shortening of the first tone. Pooled over the two gap durations, this shortening amounts to $((150.7 + 159.7) - (139.5 + 144.1))/2 = 13.4$ ms, which is about twice as long as the pooled shift due to

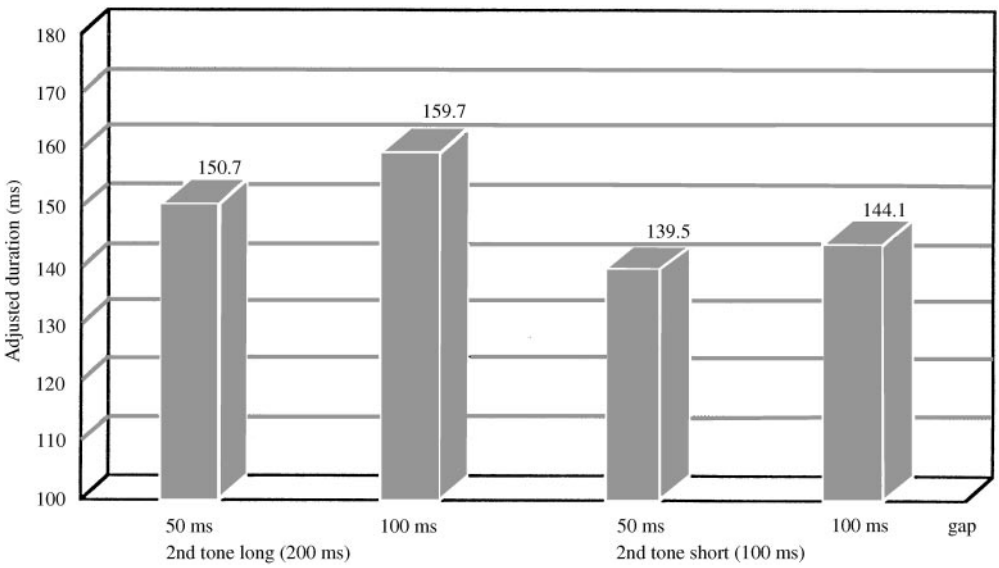

Figure 4. Mean adjusted first tone durations in ms for references with gap durations of 50 ms and 100 ms, and second tone durations of 200 ms and 100 ms, respectively. $n = 10$ listeners.

TABLE III. Analysis of variance for non-speech stimuli: Effect of the factors gap duration and duration of second tone in the reference on adjusted tone duration

Source of variation	*F*	*df*	*p*
Gap duration	5.29	1, 9	0.0470
Second tone duration	9.38	1, 9	0.0135
Gap × second tone duration	2.09	1, 9	0.1823

TABLE IV. Analysis of variance for non-speech stimuli: Effect of the factors order of presentation, gap duration, and duration of second tone in the reference on adjusted tone duration

Source of variation	*F*	*df*	*p*
Order of presentation	1.38	2, 42	0.2632
Gap duration	6.63	2, 84	0.0021
Second tone duration	66.42	1, 42	< 0.0001
Order × gap duration	0.36	4, 84	0.8399
Order × second tone duration	2.98	2, 42	0.0612
Gap × second tone duration	5.97	2, 84	0.0038

varying gap duration: $(9.0 + 4.6)/2 = 6.8$ ms. In this connection, however, it should be kept in mind that second tone duration (200 ms vs. 100 ms) varied twice as much as gap duration (100 ms vs. 50 ms). In the next experiment gap duration differences of $150 - 50 = 100$ ms were used, thus allowing a more direct comparison of the influence of gap duration vs. second tone duration. The interaction between the two factors, gap duration and second tone duration, turned out to be statistically non-significant.

3.2.2. *Three-step gap duration*

As indicated by the statistical results presented in Table IV, order of presentation of the six different reference conditions did not have a systematic influence on the listeners' responses. Therefore, in the discussion that follows, data from the three orders of presentation are pooled as is displayed in Fig. 5. From the figure it can be seen that the picture emerging from the three-step gap duration test is consistent with the results for the two-step condition, with an apparent positive relationship between adjusted tone duration and both gap and second tone duration. As in the previous experiment, the range of manipulated second tone duration (200 ms vs. 100 ms) amounted to 100 ms. Across the three gap durations, the influence of this factor averages $[(154.5 - 143.4) + (159.9 - 147.5) + (167.0 - 144.1)]/3 = 15.5$ ms. To get an impression of the relative influence of gap duration, one can compare this amount with shifts in adjusted tone duration for 50 ms vs. 150 ms gaps. For a second tone duration of 200 ms the effect amounts to $(167.0 - 154.5) = 12.5$ ms, thus somewhat less than the overall influence of second tone duration.

For a short second tone the situation is different; 50 ms vs. 150 ms gap durations result in approximately the same adjusted tone durations (143.4 ms and 144.1 ms, respectively). The latter value disturbs the patterns found elsewhere and accounts for the significant interaction between the factors gap duration and second tone duration (Table IV). The

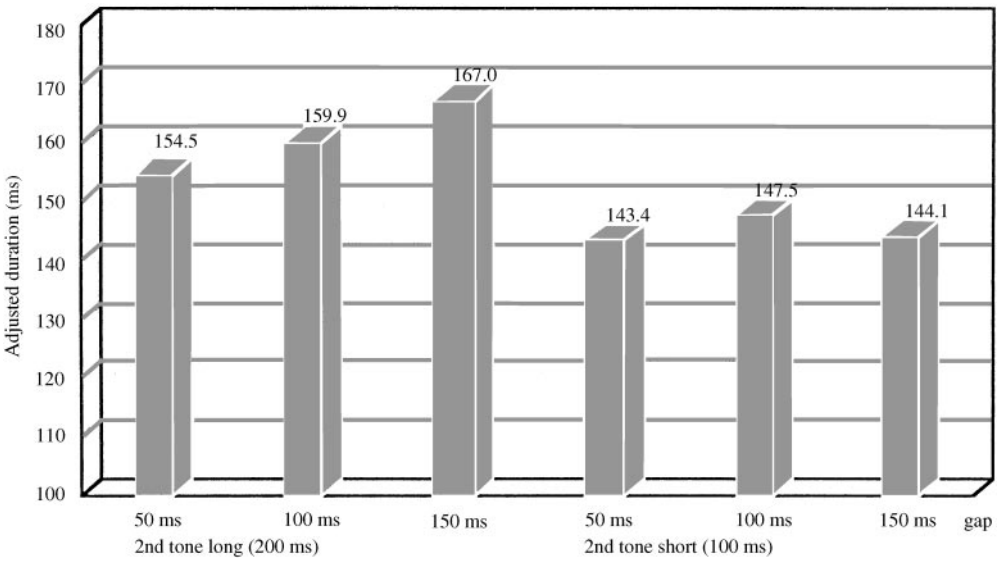

Figure 5. Mean adjusted first tone durations in ms for references with gap durations of 50 ms, 100 ms, and 150 ms, and second tone durations of 200 ms and 100 ms, respectively. $n = 45$ listeners.

reason for the non-monotonic effect of gap duration under the short second tone condition is not obvious. The only potentially relevant difference between the G150S reference and the other ones lies in the relation between the durations of the gap and the second tone. Only in this case is the former with 150 ms longer than the latter (100 ms). To a certain extent, this configuration might have caused a separation of the two tones in the reference. In turn, this might have reduced the perceptual influence of the second tone on perceived duration of the first tone. Of course, this explanation is highly speculative and further investigation would be needed to clarify this point. The general discussion will come back to this issue.

4. General discussion

4.1. *Perception of speech stimuli*

The results of the first experiment on Norwegian CVCə disyllables confirmed the hypothesis formulated at the outset: a physical variation of consonantal closure duration appeared to cause a systematic shift of the perceptual /a:/-/a/ boundary. This reflects phonological patterns in Norwegian, where a phonologically long vs. short vowel is followed by a phonetically short/long consonant, respectively. It seems reasonable to explain the phoneme boundary shifts as shifts of the vowel duration criterion, shorter consonants making the preceding vowel subjectively longer and vice versa. It is, however, a non-trivial problem how the segment durations are evaluated in detail by the listener. On the one hand, perceived durations of the single segments serve as a basis for establishing phonological categories like long/short vowel. At any time, these durations must

be evaluated relative to the temporal extensions of the surrounding context, such as the duration of the postvocalic consonant. On the other hand, all segment durations contribute to the perceptual estimation of speech rate. However, perceived speech rate is, in its turn, one of the factors against which single segment durations are measured. In the present case of a short/long occlusion duration, the levels of segment identification and prosodic rate information are intimately intertwined. A shorter closure may give rise to a perceptually longer vowel due to phonologically determined duration patterns as well as to a perceived increase of local speech rate.

It seems very well possible to explain the influence of closure duration on perceived vowel duration by appealing to Kingston and Diehl's (1994) duration ratio hypothesis. To discuss this, let us consider a certain C/V duration ratio which corresponds to a listeners' phoneme boundary in "mate/matte" perception. Lengthening the closure would cause the C/V ratio to increase, which implies that the relative vowel duration would be shorter. In order to restore the original "boundary" value of the ratio, the vowel would have to be lengthened. This is exactly what happened in the present experiments, though it should be noted that the compensation was only partial. The amounts of physical lengthening and consequent perceptual shortening differed considerably.

Unlike from the interdependency of perceived vowel duration and closure duration, the impact of schwa duration on /a:/-/a/-decisions must be ascribed to changes in perceived speech rate only. The results suggest that listeners use schwa duration in a direct comparison with the preceding vowel to adjust their long/short vowel criterion. Extending the duration ratio hypothesis to comprise larger frames than only VC units, a mechanism similar to that for vowel and closure duration can be postulated: when the schwa/V ratio is decreased due to a shorter schwa, this effect can be counterbalanced perceptually by a shorter vowel.

4.2. *Perception of non-speech stimuli*

The results of the non-speech material seemed to fully confirm the apparent assimilation effects reported by Fowler (1992). Increasing the gap in a tone-gap-tone sequence appeared to lengthen, rather than shorten the perceived duration of the first tone. The same positive relationship was found between the first and second tone durations. This finding supports the conclusion from the speech results that the time window can comprise at least three segments, or approximately 400–500 ms.

The perceptual effects measured for the non-speech stimuli can be explained by postulating the following mechanism (This was also suggested by reviewer Randy Diehl). In estimating the duration of the reference's first tone, the listener presumably judges the duration of a larger window, which possibly comprises the first tone plus the following gap. Let us consider the perceptual interaction of the first tone and gap duration. Recall that the reference had a fixed first tone duration, while the stimulus-to-be-adjusted had a fixed gap duration. Due to a longer gap in the reference, the total duration of first tone plus gap will increase correspondingly. It seems conceivable that the listener uses this total duration as a reference in judging the duration of the first tone of the stimulus-to-be-adjusted. Since the gap is shorter here, a (partial) perceptual compensation will take place by a lengthening of the first tone.

Instead of the total duration of the first tone plus the gap, it is also possible that the listener judges overall stimulus duration, i.e., the total duration of the first tone, the gap plus the second tone. This explanation offers itself for the case of apparent assimilation

between the first and second tones. Here, a shorter second tone in the reference shortens overall stimulus duration, which can be compensated for by a longer first tone.

The results of the present three-step gap duration experiment showed a conspicuous interaction between gap duration and second tone duration, such that for a gap of 150 ms and a second tone of 100 ms no clear assimilation effect was found. The initial interpretation proposed above tried to explain this deviation from the general pattern by pointing to the special temporal relations for this reference condition, where the duration of the gap exceeds that of the second tone. As is shown by the work of ten Hoopen, Hilkhuysen, Vis, Nakajima, Yamauchi and Sasaki (1993), the question of how listeners judge durations of seemingly simple events like silent intervals or tones is a very intricate one. According to their data, when judging two adjacent silent intervals delimited by very short (5–10 ms) markers, listeners tend to underestimate the duration of the second interval. The effect is almost exclusively found in those cases where the preceding duration is shorter. Informal tests indicated that this "time-shrinking" (as they called it) also occurs with temporal patterns comprising quasi-musical tones. Furthermore, the shrinking effect is asymmetric, preceding intervals being more influential than succeeding intervals. Also, the time illusion is confined to interval durations of approximately 200–240 ms. It is conceivable that effects like those reported by ten Hoopen *et al.* might have had an impact on the present listeners' duration judgements.

4.3. *Conclusion*

On the surface, the present results for the perception of speech stimuli appear to be the exact opposite from what was found for the non-speech stimuli. While in the latter case apparent assimilation occurred, the former was characterized by apparent contrast effects. The results seemed to suggest that it is the character of the acoustic signal the listener has to judge that is decisive. Speech stimuli, as opposed to similarly structured non-speech stimuli, will elicite divergent response patterns. Speech signals will be interpreted by the listener within the framework of temporal patterns that are determined by the rules for a specific language. In contrast, when confronted with non-speech stimuli, the listener cannot rely on existing temporal patterns. As a consequence, for the perceptual evaluation of such stimuli different strategies must be applied.

In this connection it can be mentioned that even within the domain of speech perception, apparently diverging results due to different perceptual tasks can be found. The interaction of perceived segment duration and f_0 contour in a segment might serve as an example. Investigating German VCən disyllables, van Dommelen (1993) found vowels with a falling contour to be perceived as shorter than vowels having a flat f_0 contour. Using the same contours in vowels preceding voiced vs. voiceless stops in disyllables resulted in an increase of "voiced" judgments with a dynamic contour. However, in view of the data on perceived vowel duration, this bias towards the voiced category could not be explained by assuming a perceptually longer vowel mediated by the dynamic f_0 contour (cf., e.g., Lehiste & Shockey, 1980). In other words, it was not possible to explain the interaction of f_0 contour and perception of voiced/voiceless stops by referring to f_0-influenced perception of vowel duration in a closely related paradigm. From this it may be concluded that, in general, we should be cautious in drawing parallels between phenomena that are seemingly akin.

This paper profited much from the very constructive criticisms given by Randy Diehl, Richard Wright, and an anonymous reviewer. Furthermore, I am indebted to Dawn Behne for her numerous useful comments and untiring help.

References

Crystal, T. H. & House, A. S. (1990) Articulation rate and the duration of syllables and stress groups in connected speech. *Journal of the Acoustical Society of America*, **88,** 101–112.

Diehl, R. L. & Castleman, W. A. (1996) Integrated perceptual properties: The affricate/fricative distinction. In *Arbeitsberichte des Instituts für Phonetik und digitale Sprachverarbeitung* (AIPUK), **31,** 191–200.

Diehl, R. L., Walsh, M. A. & Kluender, K. R. (1991) On the interpretability of speech/nonspeech comparisons: A reply to Fowler. *Journal of the Acoustical Society of America*, **89,** 2905–2909.

Dommelen, W. A. van (1993) Does dynamic F0 increase perceived duration? New light on an old issue. *Journal of Phonetics*, **21,** 367–386.

Dommelen, W.A. van (1997) A comparison of two methods for measuring perceptual boundaries. In *Reports from the Department of Phonetics, Umeå University* (PHONUM) **4,** 93–96.

Fintoft, K. (1961) The duration of some Norwegian speech sounds. *Phonetica*, **7,** 19–39.

Fourakis, M. (1991) Tempo, stress, and vowel reduction in American English. *Journal of the Acoustical Society of America*, **90,** 1816–1827.

Fowler, C. (1990) Sound-producing sources as objects of perception: Rate normalization and nonspeech perception. *Journal of the Acoustical Society of America*, **88,** 1236–1249.

Fowler, C. (1991) Auditory perception is not special: We see the world, we feel the world, we hear the world. *Journal of the Acoustical Society of America*, **89,** 2910–2915.

Fowler, C. (1992) Vowel duration and closure duration in voiced and unvoiced stops: there are no contrast effects here. *Journal of Phonetics*, **20,** 143–165.

Hoopen, G. ten, Hilkhuysen, G., Vis, G., Nakajima, Y., Yamauchi, F. & Sasaki, T. (1993) A new illusion of time perception – II. *Music Perception*, **11,** 15–38.

Keller, E. (1994) *Signalyze*™, version 3.12, Signal analysis for speech and sound.

Kingston, J. & Diehl, R. L. (1994) Phonetic knowledge. *Language*, **70,** 419–454.

Kluender, K. R., Diehl, R. L. & Wright, B. A. (1988) Vowel-length differences before voiced and voiceless consonants: an auditory explanation. *Journal of Phonetics*, **16,** 153–169.

Lehiste, I. & Shockey, L. (1980) Labeling, discrimination and repetition of stimuli with level and changing fundamental frequency. *Journal of Phonetics*, **8,** 469–474.

Levitt, H. (1971) Transformed up-down methods in psychoacoustics. *Journal of the Acoustical Society of America*, **49,** 467–477.

Miller, J. L. (1981) Some effects of speaking rate on phonetic perception. *Phonetica*, **38,** 159–180.

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

Miller, J. L., O'Rourke, T. B. & Volaitis, L. E. (1997) Internal structure of phonetic categories: effects of speaking rate. *Phonetica*, **54,** 121–137.

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

Port, R. F. (1979) The influence of tempo on stop closure duration as a cue for voicing and place. *Journal of Phonetics*, **7,** 45–56.

Port, R. F. (1981) Linguistic timing factors in combination. *Journal of the Acoustical Society of America*, **69,** 262–274.

Repp, B. H., Liberman, A. M., Eccardt, Th. & Pesetsky, D. (1978) Perceptual integration of acoustic cues for stop, fricative, and affricate manner. *Journal of Experimental Psychology*: *Human Perception and Performance*, **4,** 621–637.

Wayland, S. C., Miller, J. & Volaitis, L. E. (1994) The influence of sentential speaking rate on the internal structure of phonetic categories. *Journal of the Acoustical Society of America*, **95,** 2694–2701.