Available online at www.sciencedirect.com
ScienceDirect
Journal of Phonetics 37 (2009) 46–65
Journal of Phonetics
www.elsevier.com/locate/phonetics

Perceptual normalization for speaking rate III: Effects of the rate of one voice on perception of another ☆

Rochelle S. Newman[a,*], James R. Sawusch[b]

[a]Department of Hearing and Speech Sciences and Program in Neuroscience & Cognitive Science, University of Maryland, 0100 Lefrak Hall, College Park, MD 20742, USA
[b]Department of Psychology and Center for Cognitive Science, State University of New York at Buffalo, Buffalo, NY, USA

Received 15 August 2007; received in revised form 13 August 2008; accepted 12 September 2008

Abstract

Individuals vary their speaking rate, and listeners use the speaking rate of precursor sentences to adjust for these changes [Kidd, G. R. (1989). Articulatory-rate context effects in phoneme identification. *Journal of Experimental Psychology: Human Perception and Performance*, *15*, 736–748]. Most of the research on this adjustment process has focused on situations in which there was only a single stream of speech over which such perceptual adjustment could occur. Yet listeners are often faced with environments in which multiple people are speaking simultaneously. Each of these voices provides speaking rate information. The challenge for the listener is to determine which sources of information should apply in a speech perception situation. Three studies examined when listeners would use rate information from one voice to adjust their perception of another voice. Results suggested that if only one source of duration information was available, listeners used that information, regardless of the speaker or the speaker's spatial location. When multiple sources were available, listeners primarily used information from the same source as the target item. However, even information from a source that differed in both location and talker still influenced perception to a slight degree.

1. Introduction

Speakers vary in the rate at which they typically speak (Crystal & House, 1988), and the same talker may speak at different rates at different points in time (Miller, Grosjean, & Lomanto, 1984). These changes in speaking rate result in changes in the duration of speech segments (Crystal & House, 1982). This poses a problem for perception, because some contrasts are cued, in whole or in part, by duration. For example, the /b/–/w/ manner contrast can be cued by differences in rate of change (a correlate of duration) alone (Miller & Liberman, 1979), and duration can be a sufficient cue to the /i/–/ɪ/ vowel contrast in English (Ainsworth, 1972). Yet when a speaker talks quickly, listeners do not perceive the intended /w/s as being /b/s, despite their shortened duration. Instead, listeners appear to adjust, or normalize, their perception for the rate at which a person speaks.

A number of studies have investigated how this process of rate normalization occurs (Hirata & Lambacher, 2004; Miller, 1981, 1987; Miller & Liberman, 1979; Newman & Sawusch, 1996). Normalization occurs on the basis of both prior (Kidd, 1989; Summerfield, 1981) and subsequent rate information (Miller & Liberman, 1979). Prior information appears to consist of two components, one based on the rate or rhythm of stressed syllables in the preceding phrase, and one based on the duration of the segment or segments immediately preceding the target item (see Kidd, 1989). Subsequent rate normalization is based on the segments immediately following the target item and appears to be limited to a temporal window of approximately 250 ms. Only segments that fall within this short temporal window can influence the perception of a preceding phoneme distinction (Newman & Sawusch, 1996; but see Hirata & Lambacher, 2004 for different results in Japanese). Within

☆The "III" in the title is because this is the third in a series of studies on speaking rate; the first (Newman & Sawusch, 1996) looked at how far from the target a source of rate information could be located and still have an effect; the second (Sawusch & Newman, 2000) addressed the effect of discontinuities within a signal.

*Corresponding author. Tel.: +1 301 405 4226; fax: +1 301 314 2023.
E-mail address: rnewman@hesp.umd.edu (R.S. Newman).

0095-4470/$ - see front matter © 2008 Elsevier Ltd. All rights reserved.
doi:10.1016/j.wocn.2008.09.001

this window, use of speaking rate information appears to be obligatory (Miller & Dexter, 1988), and is not influenced by factors such as the acoustic or perceptual similarity between the segments (Newman & Sawusch, 1996). That is, perception of a fricative is influenced by the duration of vocalic sounds to the same extent as by the duration of other fricatives.

Nearly all of the work to date has focused on situations in which there was only a single source of speech (or nonspeech) over which such perceptual adjustment could occur. In these studies, listeners hear a single talker speaking on each trial, with the speech occurring at either a fast (short duration), medium, or slow (long duration) speaking rate. Listeners appear to always make use of this speaking rate information, but there is no other source of speaking rate information present.

In daily life, listeners are often faced with environments in which multiple people are speaking simultaneously. Each of these voices provides speaking rate information, and the listener must select which sources of information to apply to any given speech perception situation. Thus, in deciding whether Jane said "ba" or "wa", the listener must not only adjust his or her perception for Jane's speaking rate, but must also select which information is most relevant to make that adjustment. Presumably, information about the duration of Jane's speech segments, but not duration information from other talkers, is relevant to perceiving Jane's phonetic segments. Thus the process of rate normalization and speech recognition depends on the ability of the listener to separate different sources of sound (or different voices) and assign them to different streams. Yet there has been little work investigating how these different processes might interact. (Although for a model that attempts to incorporate both processes, see Grossberg, 2003). In the next sections, we first review a portion of the literature on auditory stream segregation, and then discuss how this might interact with speaking rate normalization.

1.1. Auditory stream segregation and its effect on phonetic perception

Much of the work on auditory stream segregation has focused on either simplified auditory stimuli or on music (Bregman, 1990; see Snyder & Alain, 2007 for a recent review). However, there is also a small literature examining the perceptual organization of speech. Most has focused on selective attention, examining the cues adults can use to separate different sources of speech occurring simultaneously, although a few studies have focused more on cues used to group sequences across time (see Alain & Arnott, 2000; Ciocca, 2008 reviews). Adults perform better either at separating different sound streams or at attending selectively to a single stream when the streams differ on any of a variety of acoustic cues, including location in space (Broadbent, 1954; Cherry, 1953; Hirsh, 1950; Pollack & Pickett, 1958; Poulton, 1953; Spieth, Curtis, & Webster, 1954; see also Kidd, Arbogast, Mason, & Gallun, 2005), frequency range (Bregman & Pinker, 1978; Dannenbring & Bregman, 1978), sex of the talkers and their voice pitch for speech (Broadbent, 1952; Brokx & Nooteboom, 1982; Brungart, 2001; Brungart & Simpson, 2007), onset and offset times (Bregman & Pinker, 1978; Dannenbring & Bregman, 1978), and differences in amplitude modulation (Bregman, Abramson, Doehring, & Darwin, 1985; Grimault, Bacon, & Micheyl, 2002).

Interestingly, a number of studies suggest that the information that goes into a phonetic percept may not be limited to a single perceptual stream (at least as defined by low-level perceptual grouping cues; Remez, Rubin, Berns, Pardo, & Lang, 1994). For example, Broadbent and Ladefoged (1957) presented listeners with 2-formant synthetic speech signals in which the two formants were excited by separate, slightly different fundamental frequencies. Listeners reported hearing the same sentence spoken by two different talkers. In other words, their phonetic perception was based on information from both formants, even though listeners simultaneously reported hearing two different perceptual streams. Similarly, studies in duplex perception (Whalen & Liberman, 1987) have reported situations in which listeners report hearing a speech stream separate from a tone, despite the fact that information from the tone influenced phonetic perception within the speech stream. These studies suggest that perception of one or more streams at the level of conscious awareness is not necessarily directly tied to how signals are grouped for phonetic perception. Thus, in a multi-talker situation, even if a listener reports hearing separate talkers (and thus separate streams), this need not mean that speech perception was based on information from a single stream alone. Indeed, according to Remez et al. (1994), "phonetic organization diverges from auditory scene analysis early in perception and proceeds independently" (p. 151). If so, there is no reason to believe that speaking rate normalization would necessarily be limited to the speech from a single talker.

Other studies suggest that phonetic perception is altered when a change in talker occurs, indicating that the two are not as separate as the above argument would suggest. Dorman and colleagues reported that a change in talker (Dorman, Raphael, & Liberman, 1979) was sufficient not only to cause the segregation of a speech sequence into two separate perceptual streams, but that it simultaneously changed the interpretation of acoustic qualities as phonetic cues. When listeners were presented with the sequence "Please say" followed by the word "shop", a 50-ms silent gap before the final word was sufficient to change listeners' perception of the final word from "shop" to "chop". The silent gap was taken as an indication of vocal tract closure and thus as a cue to the presence of a following affricate. However, when the final word was spoken by a different talker, listeners consistently reported that the last word was "shop", regardless of the length of the silent gap. The silent gap was not interpreted as a phonetic cue (as closure duration) when it accompanied a change in talker. This

suggests that a change in source attribution altered phonetic perception, not just auditory scene analysis.

Despite this finding, it would appear that, in at least some cases, phonetic perception may be influenced by information coming from an alternate sound source. This suggests that it is worth exploring how the presence of multiple talkers could influence perceptual processes such as normalization for speaking rate.

1.2. The interaction between auditory stream segregation and rate normalization

How might rate normalization work in a multi-talker environment? One possibility is that rate normalization would occur only within a particular talker (or stream). If two individuals spoke simultaneously, or if one individual stopped speaking and a second began, listeners would not use speaking rate information from one voice to help them interpret a second voice. This notion implies that the acoustic information from the different talkers has been separated to some extent prior to the point at which rate normalization begins. In other words, rate normalization would have to occur after some stage or stages of perceptual organization, in which the incoming auditory signal is divided into different perceptual streams on the basis of talker identity.

Yet as noted above, phonetic perception does not necessarily follow the principles of auditory stream segregation (Remez et al., 1994) and it is possible that rate normalization likewise may not be tied to a single perceptual stream. Rate normalization could be based on whatever duration information is available, regardless of its sound source. In fact, there is some evidence to suggest that rate normalization is not always limited to information originating from a single voice. Sawusch and Newman (2000) presented listeners with the syllables /bʌlz/ and /pʌlz/ (“buhlz” and “puhlz”). One set was spoken entirely by a female talker. In a second set, the /lz/ of the female talker was digitally removed and replaced with the same segments spoken by a male talker. Perceptually, the edited items sounded as if one talker stopped speaking and another began speaking in a single syllable. Sawusch and Newman found that the duration information from the second talker influenced listeners’ perception of the initial consonant of the first talker, despite the fact that the two sound sources clearly represented different voices. This suggests that rate normalization will occur across two different sound sources, at least in some situations. Similarly, Green and colleagues (Green, Stevens, & Kuhl, 1994; but see Lotto, Kluender, & Green, 1996) presented listeners with syllables containing a change in formant structure partway through a vowel. Even though listeners heard this as a change in talker identity, their phonetic perception combined across the two portions of the vowel.

In a recent study, Wade and Holt (2005) used fast and slow sequences of tones followed by a /bɑ/–/wɑ/ test series. The rate of the tone sequence influenced listeners’ perception of the test series in much the same way that variation in speaking rate does. The fast precursor resulted in fewer /b/ responses and an earlier category boundary in the test series relative to the slow precursor series even though the precursor was clearly not speech. Our purpose in discussing the results of Wade and Holt is not to raise the issue of how speech and nonspeech signals interact in perception (see Wade and Holt for a recent summary of this issue). Rather, their results, like those of Sawusch and Newman (2000), show that speech from a single talker is not a prerequisite for the “usual” influences of speaking rate on phonetic perception.

In contrast to the above, there is also one study that reported that speaking rate influences did not cross a change in voice. Diehl, Souther, and Convis (1980) used the synthetic precursor phrase “Teddy hears”, followed by a synthetic syllable from a /gɑ/–/kɑ/ test series. Across three experiments, they varied either the fundamental frequencies, formant frequencies, or both of the precursor in relation to the target, so as to mimic a change in talker. In most cases, a difference in the synthesis parameters between the precursor phrase and the target eliminated any influence of the speaking rate of the precursor phrase upon the target. (This would appear to be consistent with the Dorman et al. (1979) results that phonetic integration is also blocked by a change in talker.) However, these results must be interpreted with caution. In particular, Diehl et al. found a reversal of the typical rate pattern when both precursor and target were synthesized as a female voice. Since other studies have not reported unusual patterns with female voices, there is reason to think that the results of Diehl et al. might not generalize. It is possible that their synthesis parameters for the precursor phrase may not have adequately mimicked changes in speaking rate or that the change in synthesis parameters across conditions led to other, unintended changes in how listeners perceived the precursor phrase and its relation to the target syllable (for example, perhaps their synthesis parameters did not create a clear pattern of stressed syllables, which is critical for long-range speaking rate effects; see Kidd, 1989). Certainly in light of the results of Sawusch and Newman (2000) and Wade and Holt (2005) that show talker continuity is not always necessary for rate normalization, the influence of a change in talker warrants further investigation.

It is possible that listeners in the Sawusch and Newman and Wade and Holt studies simply misperceived the two portions of the acoustic signals as originating from the same source, at least at an early stage of perceptual processing. While there were clearly some cues to suggest the two parts of the acoustic signal were from different sound sources (such as differences in fundamental frequency and spectral properties), there were also cues to suggest they were not. In particular, the two voices (or nonspeech and speech) were timed such that the second voice began speaking at the exact instant that the first ended. Such continuity is unlikely to occur in the real world unless the two segments actually come from the same

source. Moreover, both voices came from the same location in space, which is also an uncommon occurrence outside of a laboratory setting (except for multi-party telephone calls). Finally, the speech information across the two voices in Sawusch and Newman exhibited phonetic coherence (Remez et al., 1994). Put another way, while there were spectral changes that indicated a change in talker, the information specifying the phonetic segments changed in an orderly and well-delineated way such that it formed a single, coherent stream or group. These properties of the signal may have been sufficient to counteract the spectral discontinuity caused by the source change, and to have led the listener to (mistakenly) treat the two sequences as coming from the same source (for more on auditory stream segregation in general, see Bregman, 1990).

In most situations, there are likely to be multiple acoustic cues that guide the segregation of a signal into information from different sources, and that influence whether a signal is perceived as a single or multiple streams. In addition, the process of stream segregation may occur at different (or multiple) points in perceptual processing depending upon the information available to the listener and the listener's allocation of attention to the signal and task (for effects of attention, see for example, Cusack, Deeks, Aikman, & Carlyon, 2004). If this view is correct, rate normalization may occur across streams of speech in some situations, but not in others, depending on the number and quality of segregation and grouping cues (for work comparing the effectiveness of different cues for selective attention see Darwin & Hukin, 2000). In the real world, talkers often have different vocal qualities, their speech comes from distinct points in space, and the prosodic, syntactic, and semantic qualities may cohere to a greater or lesser extent. Any and all of these qualities may contribute to separating the sound into different streams and to keeping speaking rate information in one voice from influencing perception of another voice. It is also possible that there are some very-low-level acoustic principles (such as location in space; Broadbent, 1954) that may be more likely to induce segregation than are changes in talker identity. If so, rate normalization may occur across two different talkers, but would be less likely to occur across different locations in space.

The role of various grouping and segregation cues could also depend upon the number of voices or sound sources that the listener encounters. If prior context suggests that only one sound source is present in the environment, rate normalization may occur regardless of changes in source quality or location. In this situation, the presence of one voice would have already established that a single stream is present, and a syllable in a new voice would initially be processed as part of this continuing stream. Put another way, after hearing a female talker speak in isolation, a change in talker may not be sufficient to disrupt rate normalization. However, if prior context suggests that two or more sound sources are present simultaneously, rate normalization might occur only within one of these sources. Listeners would be less likely to treat speech sequences with different spectral properties as belonging to the same stream when an alternative perceptual organization is present and perceptual grouping processes are already being actively used to separate the voices. This approach implies that the same change in talker voice (or other sound source changes) might have different effects depending on the context in which the listener hears the target.

The present studies investigate how rate normalization interacts with effects of perceptual grouping. We do not purport to examine perceptual grouping, per se. Our approach is to present listeners with potential sources of duration information that differ from the target word on the basis of talker identity and/or spatial location, and examine when listeners normalize on the basis of this duration information. Thus, we explore what types of putative grouping cues disrupt the process of rate normalization. These studies focus on the situation in which there is a change in talker and/or location from the precursor phrase, which establishes the speaking rate, to the syllable with the target phonetic contrast.

Experiment 1 examines the effect of voice changes on the combined long-range and short-range effects (Kidd, 1989) of a precursor phrase in rate normalization. Four conditions were tested. In two of these conditions, the precursor phrase and target item were spoken in the same voice. In the other two conditions, there was a change in talker between the precursor phrase and the target word. The two conditions for each talker voice differed in terms of the location from which the precursor voice and the target word seemed to originate: either the precursor phrase appeared to come from the same spatial location as the target word, or it appeared to differ in its origination point. This study thus serves to examine the extent to which rate normalization occurs across different talkers, and across different locations in space. It also tests whether normalization effects are disrupted to a greater extent when there are multiple cues supporting segregation (a change in talker combined with a change in spatial location) as compared to when there is only a single cue supporting segregation (a change in talker or spatial location alone).

Experiments 2 and 3 examine the situation in which two voices are speaking simultaneously during the precursor phrase. In this situation, the listener has two potential sources of duration information, and has to select which of the two sources to use. This is in contrast to Experiment 1, in which there was only one potential source of information, and listeners could either make use of that information or not. Experiments 2 and 3 investigate whether rate information from only one voice, or from multiple voices, will influence processing of a duration-based contrast. Studies of divided attention (Gallun, Mason, & Kidd, 2007; Mullennix, Sawusch, & Garrison-Shaffer, 1992; Wood & Cowan, 1995) show that listening to two channels at once incurs a cost in processing, reducing performance on a listening task. This suggests that listeners are likely to

choose to attend to a single channel. At the same time, it is also clear that some information from an unattended channel in listening does receive processing (Nusbaum & Schwab, 1986; Treisman & Gelade, 1980; Wood & Cowan, 1995) and thus could influence perception of the attended signal/channel. That is, even assuming that listeners attend to a single channel, it is still possible that information from the alternative channel could influence processing. Thus, depending upon when speaking rate normalization occurs with respect to other basic perceptual processes, listeners faced with multiple sources of information may use only one source (talker) or more than one source (talker) in speaking rate normalization.

Together, the attention literature and prior research on auditory stream segregation suggest that perceptual grouping can occur at multiple levels of processing and is influenced by both stimulus information and the strategy used by the listener. For this reason, it seems reasonable to consider the possibility that perceptual grouping may not have been completed at the point at which rate normalization occurs, and thus that information from alternative streams of speech could influence perceptual processing. Ultimately, our understanding of how attention and perceptual grouping influence speaking rate normalization will necessitate studies that seek to control the individual's allocation of attention during listening. At present, however, this would be premature since little is known about the situations in which speaking rate normalization does, or does not, use information from multiple talkers. The experiments described here serve as a start to answering this question. If listeners consistently exploit information from a voice other than the target voice then future studies can explore the influence of listener strategies, task and attention on the extent of that information use.

2. Experiment 1

As noted above, previous work has demonstrated that rate normalization effects can occur across changes in talker identity (Sawusch & Newman, 2000) or even across speech and nonspeech stimuli (Wade & Holt, 2005). To the extent that normalization with precursor and subsequent information involve the same perceptual mechanism(s), we might expect that the spectral discontinuities associated with a change in talker from precursor to target would be likewise insufficient to disrupt normalization. As a test of this hypothesis, we presented listeners with a sentence in which the final "word" (a nonword syllable) contained a duration-based contrast, the VOT distinction between /g/ and /k/. The rest of the sentence (that is, the part up to the final word) was spoken at one of three different speaking rates. Four groups of listeners participated and all four were asked to identify the same target (final) syllables. All four heard a precursor phrase that varied in speaking rate. What differed among the four groups was whether that precursor phrase matched or mismatched the target word in talker, and matched or mismatched the target word in spatial location. For ease of communication, these four variants are henceforth referred to as the "same-voice/same-location", "different-voice/same-location", "same-voice/different-location", and "different-voice/different-location" conditions. We test whether the precursor phrase's duration influenced listeners' perception in each case.

Based on the research literature reviewed above, three different patterns of data appear plausible. First, we could find the same effect of speaking rate for the same-voice and different-voice conditions, supporting the proposal by Sawusch and Newman (2000) that all of the speaking rate information within a coherent signal is used in normalization. The effect size might, however, be reduced for the different-location conditions (regardless of the talker voice), as the low-level cue of spatial location differences could be enough to disrupt rate normalization.

Second, we might find a smaller (but still present) effect of speaking rate for the different-voice condition. Kidd (1989) argued that there are two different sources of information for rate normalization: a long-range system based on stressed syllables in the precursor and a short-range system that exploits the segment durations immediately preceding the target. The short-range system appeared to show cross-voice effects in Sawusch and Newman (2000), but it is possible that the long-range system requires talker continuity. If so, the influence of the single-talker precursor on the target would involve two sources of information (from the short- and long-range systems) while the influence of the different voice would involve only one. This would result in reliably smaller effects for both the different-voice and different-location conditions compared to the same-voice/same-location condition.

Finally, we might find no reliable effects of speaking rate in the different-voice condition, even when the two voices come from the same location in space. Though we do not consider this outcome likely, it is possible given the results of Diehl et al. (1980) and Dorman et al. (1979). If this were found, it would cast serious doubt on the interpretation offered by Sawusch and Newman (2000) for their results.

In this study, the last word of our precursor phrase ended in the stop consonant /d/ and our target (final) word began with a stop consonant (/g/ or /k/). Consequently, there is a closure interval between the end of the precursor phrase and the start of the target word. This closure can be thought of either as part of the precursor phrase (closure for the syllable final stop), as part of the target word itself (closure for the following, initial stop consonant), or as neither or both. How it is processed by the perceptual system has implications for how speaking rate variation should be implemented in this study (and the following ones). If closure duration is treated as part of the precursor phrase, then the closure is the most immediately adjacent segment to the target phoneme. Its duration should then have a contrastive effect on the target word, such that a longer closure should make the following stop consonant

seem short in contrast. If this is the case, it would be appropriate methodologically to treat the closure as part of the precursor phrase, and vary its duration along with that of the precursor (so that the entire precursor phrase varied in a similar way).

Alternatively, if this closure is perceived as being an integral part of the target word, then varying its duration is a variation in the acoustic correlates to the stop voicing distinction of the target word. In this case, the exact opposite pattern of results should occur: a *shorter* closure interval should result in perception of more voiced initial consonants since voiced stops are generally preceded by shorter closure intervals than voiceless consonants (see Lisker, 1986). In this situation closure duration is being treated as part of the target, rather than being a basis of comparison for it. If the closure duration is treated as part of both the precursor and the target, then both an increase and a decrease in voiced (/ɡ/) responses could occur because of different uses of the closure information, resulting in either no or a very small net influence of the variation in closure duration on listener responses. Finally, the speech processing system could simply ignore the closure period altogether (as was apparently the case in Dorman et al., 1979), in which case its duration should have no effect on listeners' perception.

Understanding how closure duration affects perception is thus critical to all other studies manipulating precursor duration. We therefore decided to explore the effect of the closure duration directly in Experiment 1A, prior to the primary focus of this experiment (Experiment 1B). Experiment 1A uses a constant duration precursor phrase, either in a male or female voice, but varies the closure duration, to see what effect this closure duration has on listeners' responses to the voicing contrast in the final, target syllable. Experiment 1B then explores the effect of precursor phrase and spatial location, as described above.

2.1. Method

2.1.1. Listeners

For Experiment 1A, the listeners were 24 undergraduates at the State University of New York at Buffalo who participated in exchange for either a cash payment or for partial course credit. For Experiment 1B, the listeners were 95 undergraduates at the University of Maryland and the State University of New York at Buffalo. All were native speakers of English with no reported history of either a speech or a hearing disorder. Data from 1 participant in Experiment 1A and 12 participants in Experiment 1B were excluded for the following reasons: computer errors (2), nonnative speaker (2), age (>70 years, raising concerns of hearing loss), failure to complete the study (4), or failure to respond on a sufficient number (85%) of trials (4).

2.1.2. Stimuli

Two speakers of Midwestern American English (the authors) recorded the sentence, "I heard him say the word gipe" at three different speaking rates: a normal rate, a fast rate, and a slow rate. In order to ensure that the three different rates were as similar as possible across talkers, the male talker produced multiple examples at each of the three rates. These phrases were examined with Praat (Boersma & Weenink, 2005) to find ones that best matched the rates in the female speaker's phrases in terms of syllable durations and prosodic contours. This "best match" was selected as the appropriate precursor.

The final nonword was excised from each sentence, and the remainder of the sequence served as the precursor phrase. This resulted in six versions of the precursor phrase: fast, medium, and slow versions in both a male and a female voice. The durations of each of these precursor phrases are shown in Table 1. All recordings were made at a 20 kHz sampling rate with 16 bits precision.

Each of the precursors contained voicing into the closure period. This was edited, digitally, to be 50 ms in duration (this involved reduplicating a pitch pulse for the fast rate, but removing pulses for the slow rates); a constant duration was used so as not to overly influence the target.

The 50 ms of final /d/ voicing was followed by silence. In Experiment 1A, the silent intervals were either 50 or 100 ms while for Experiment 1B, the silent interval was set at 50 ms. This value was chosen because it resulted in closure intervals intermediate between those typical of initial /ɡ/ and /k/.

In addition, the male talker recorded the syllables "gipe" (/ɡaɪp/) and "kipe" (/kaɪp/, rhymes with ripe) in isolation. The use of a nonword item is relatively standard in rate normalization research, and avoids effects of lexical frequency differences across endpoints in real word stimuli. The /ɡaɪp/ to /kaɪp/ series was created by replacing successively longer portions of the initial /ɡ/ in /ɡaɪp/ with the same duration release plus aspiration from /kaɪp/ (see Ganong, 1980). All editing was done digitally, at zero crossings in the waveform, to avoid the introduction of

Table 1
Precursor phrase durations, in ms, for both male and female voices in Experiments 1–3

Sentence	Fast	Medium	Slow
"I heard him say the word", male voice (single-voice condition, Exp. 1)	753	971	1220
"I heard him say the word", female voice (different-voice conditions, Exp. 1–3)	767	952	1254
"You wrote to me and said", male voice (correct/male voice, Exp. 2 and 3)	750	920	1222

Note: closure durations were 50 ms in all cases (except Exp. 1A), and were not included in the above measures.

pops or clicks in the waveform. The original /gɑɪp/ was used as the first stimulus in the series; the /g/ had a voice onset time (VOT) of 14 ms. The second stimulus was created by removing the initial 14 ms release burst of the /g/ and replacing it with the initial 14 ms of release from the /k/. Thus, the first two stimuli contained the same VOT, but different release bursts. The third stimulus was created by removing the release burst and the first vocal pulse from the onset of the /g/ and replacing it with the equivalent duration burst plus aspiration from the onset of the /k/. The fourth, fifth, sixth, seventh, and eighth stimuli involved removing the /g/ release and successively more vocal pulses (2, 3, 4, 5, and 6) and replacing them with equivalent duration release and aspiration from the /k/. This resulted in a series with VOTs of 14, 14, 25, 36, 45, 54, 63, and 72 ms. This method of stimulus creation has been widely used in the literature as a means of series construction, and results in clear, natural-sounding stimuli (see Ganong, 1980; Newman & Sawusch, 1996).

Each of these eight target syllables was then appended onto each of the six precursor phrases, resulting in a total of 48 items. Listeners in the same-voice condition of Experiment 1B heard the three series with the male precursor phrase and male target syllables. The male talker's voice pitch varied within each sentence, with averages across the sentences ranging from 93.4 to 94.9 Hz for the different speaking rates. The final /gɑɪp/, with a pitch of 98.3 Hz, sounded as if it came from the same speaker. Listeners in the different-voice condition heard the three series with the female precursor phrases (with pitch minimums of 218 Hz, very distinct from the male voice) and male target syllables. Listeners in Experiment 1A heard the medium-rate precursor phrases in both voices (but not the fast or slow precursors). Their items contained either the 50 ms silence or a slightly longer, 100-ms silence.

Finally, to create the stimuli for the different-location conditions in Experiment 1B, the same items were used, except that the precursor phrase was presented in only one earphone (monaural) while the target syllable was presented in both earphones (binaural). This is, in essence, a change in lateralization of stimuli, rather than localization, *per se*. Thus, the target items at the ends of the sentences were acoustically identical to those in the other two conditions. An apparent location change occurred between the precursor phrase, which appeared to come from one side of auditory space (or from one ear), and the target word, which appeared at the midline (or coming from the center of the listener's head, just as it did in the same-location conditions).[1] Half of the listeners in these different-location conditions heard the precursor phrase from the right ear only, and half from the left ear only, to counterbalance any effects of ear.

[1]This results in a location change both from the side to midline and from outside the listener's head (external) to an internal location. It is not clear which of these changes is most critical, or whether an external/internal change is responded to differently than would a change from one external location to another. As a beginning investigation of this issue, we felt it most important to maintain the acoustic identity of the target stimuli across conditions, resulting in that stimulus being played binaurally.

2.1.3. Procedure

All audio presentation, timing of intervals, and recording of responses was controlled by a Macintosh 7100AV computer. The listeners were divided into four groups, as described above. All participants were asked to identify the initial consonant in the final syllable of each sentence as either a /g/ or a /k/ using a six-point rating scale. The scale ranged from a 1 for a good, clear /g/ though 3 for an ambiguous /g/ and 4 for an ambiguous /k/ to 6 for a good, clear /k/. Listeners heard a practice block of 24 trials, followed by 6 test blocks consisting of three repetitions of each of the 24 syllables in their condition. The order of trials was randomized within each block.

2.2. Results and discussion, Experiment 1A

First, for each listener, an average rating was computed for each stimulus in each series. This average incorporates not just identification but also goodness ratings. These rating functions are shown in Fig. 1. As can be seen, participants' ratings were monotonic across the continua, with extremely high endpoint ratings and steep functions. This suggests that only the central items in the series were ambiguous, and items near either endpoint were perceived as excellent examples of their category. From these rating functions, the category boundary between /g/ and /k/ was determined by linear interpolation using the rating responses for the two stimuli on either side of the boundary. If the boundary was crossed more than once, an average was taken of the crossings going in the appropriate direction. The location of the category boundary reveals the influence of speaking rate on the putatively most ambiguous stimuli, and is the means of assessing effects of speaking rate on perception that has been used in most prior studies. Second, the total percentage of /g/ responses for each series was determined by summing the rating responses of 1, 2, and 3 across the stimuli in each series. This measure includes effects on stimuli away from the category boundary and may, as a result, be slightly more sensitive (see Samuel, 1986).

The effects of closure duration variation were assessed with two 2-way ANOVAs (one for each dependent measure), with the variables of precursor talker voice (male or female) and closure duration (short or long); partial η^2 (η_p^2) was calculated as a measure of effect size. Fig. 2 shows the results based on percentage "g" responding for the same-voice condition on the left, and for the different-voice condition on the right. Figures based on category boundaries look similar.

There was an overall effect of voice, such that listeners had an earlier category boundary when the precursor voice was male than female (by category boundaries, $F(1,23) = 10.84$, $p < .005$, $\eta_p^2 = .32$; by percentages, $F(1,23) = 6.72$, $p < .05$, $\eta_p^2 = .22$). More importantly, there was an overall

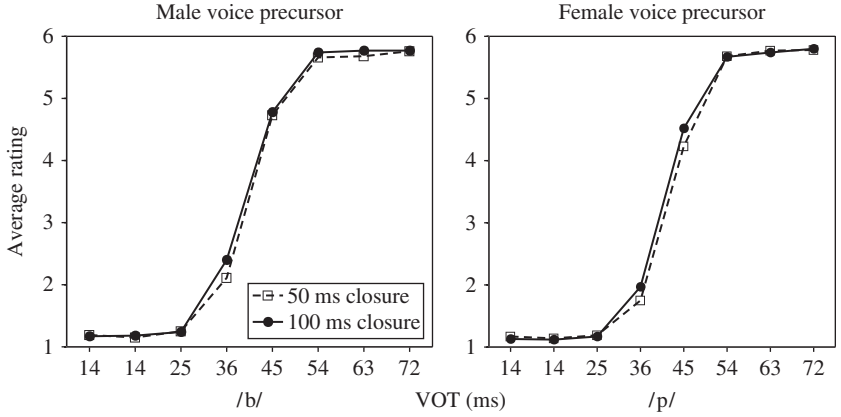

Fig. 1. Group identification functions for the gipe–kipe series, with a male precursor voice (left) and a female precursor voice (right).

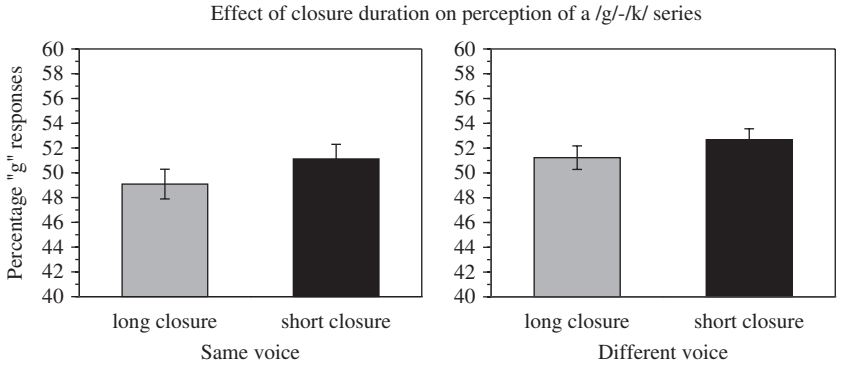

Fig. 2. Percentage of "g" responses with male (left) and female (right) precursor phrases, with different closure durations. Error bars reflect standard error.

effect of the closure duration (by category boundaries, $F(1,23) = 13.39$, $p < .005$, $\eta_p^2 = .37$; by percentages, $F(1,23) = 13.56$, $p < .005$, $\eta_p^2 = .37$), and no interaction between these effects (both $F < 1$). The effect of closure duration was such that a shorter closure duration led to perception of more items as being the shorter (voiced /g/) consonant, as can be seen in Fig. 1. When the closure duration was short, 52% of the items were perceived as being a /g/; when the closure was long, only 50% were perceived as being a /g/. That is, the direction of the effect suggested that the closure duration was integrated with the following consonant (cf. Lisker, 1986).

Clearly, these effects are small; this is not surprising, given the small change in closure duration that was presented. But the directionality and consistency of the results strongly suggest that the dominant influence of closure duration is as being part of the following stop consonant, rather than being part of the precursor phrase (or being ignored entirely). We therefore kept the closure duration constant for all other experiments in this paper (rather than having it vary as part of the precursor phrase).

Despite the clarity of the results, however, these findings do conflict with the prior work by Dorman et al. (1979). In their cross-voice study, Dorman et al. found that the duration of the closure did not influence perception of the following phoneme as being an affricate or fricative. That is, the closure period was not integrated with the following consonant in their study; to quote the authors, "it is as if their perceptual machinery 'knew' that, with two speakers, intersyllabic silence conveys no useful phonetic information" (p. 1529). Here, though, the duration of the closure *was* interpreted as conveying information relevant to phonetic perception. What might have caused the difference between these sets of results?

There are (at least) three differences between the studies. First, Dorman et al. used a range of closure durations (from 0 to 100 ms), as compared to our use of 2 durations; this variability in closure may have contributed to the formation of two separate streams of sound. Second, the studies differ with regard to the phonetic distinctions being made, and the way in which closure duration influences that distinction. In the present study, the closure duration is a potential acoustic correlate of the stop voicing distinction between /g/ and /k/. It is possible that the abrupt release of the /g/ and /k/ along with the presence of a clear silent interval in all cases led our listeners to treat the silent intervals as closures and integrate the closure with the stop voicing contrast. In contrast, in one of two experiments in Dorman et al., the fricative/affricate contrast (a distinction in manner of articulation) was used.

In their other experiment, the task was to recognize the place of articulation of a syllable final stop. If listeners had integrated the phonetic information across talkers, then with short closures the identity of the syllable final stop would be obscured (as was found for a single talker condition). Thus, this task may have encouraged perceptual segregation.

A third (and we think more likely) possibility is related to the duration of the "precursor" in these different studies. Ours was a natural speech phrase with normal intonation. Moreover, the target word completed the phrase to make a coherent, syntactically well-formed sentence. In addition, the precursor was substantially longer in our study (six syllables) than in either Dorman et al. (1979) or in Diehl et al. (1980) (one and three syllables, respectively). The long precursor in a single voice may be a strong acoustic cue to the presence of a single, coherent stream of sound and promote the integration of the target (in a different voice) with the precursor. In contrast, the single syllable precursor in Dorman et al. and the synthetic and potentially unnatural-sounding 3-syllable phrase in Diehl et al. may have provided very weak evidence of a single stream of sound. In turn, this could have led to the formation of separate streams of sound and a lack of integration across the change in talker.

Regardless of the actual reason for the difference between these previous results and Experiment 1A, the results of this study are clear. For the /g/ versus /k/ distinction used in the present paper, the duration of the closure appears to be integrated with the VOT information in listeners' perception. This suggests that it is most appropriate to leave the consonant closure duration constant so that the phonetic integration does not obscure the influences of precursor speaking rate on the target.

2.3. Results and discussion, Experiment 1B

Data were analyzed in the same manner as in Experiment 1A. Fig. 3 shows the results based on percentage "g" responding for the same-voice conditions on the left and the different-voice conditions on the right. The conditions with the same spatial location are shown on the upper half; those with a change in spatial location are shown on the bottom half.

The data were analyzed with a 2 (voice) × 2 (location) × 3 (speaking rate) ANOVA, with partial η^2 (η_p^2) as a measure of effect size. There was a main effect of voice, such that listeners had identified more items as being "g" with the female precursor ($F(1,91) = 8.29$, $p = .005$, $\eta_p^2 = .08$ by category boundaries, $F(1,91) = 11.73$, $p < .001$, $\eta_p^2 = .10$ by percentages). There was also a main effect of spatial location, such that listeners labeled more items as being "g" after hearing a binaural precursor (one in the same spatial location; $F(1,91) = 13.13$, $p < .001$, $\eta_p^2 = .12$ by

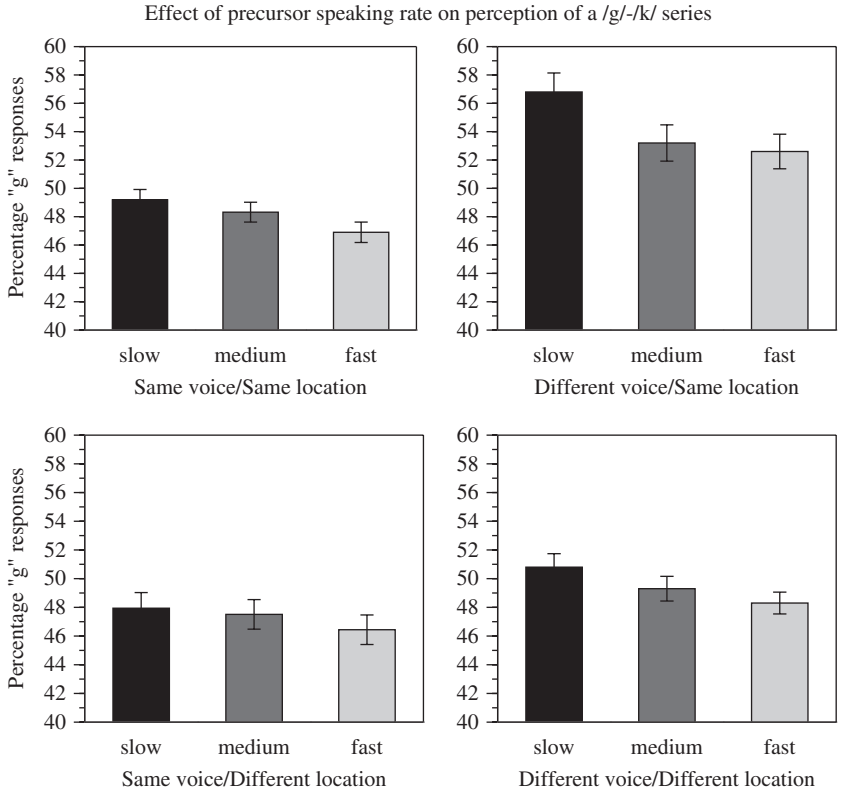

Fig. 3. Percentage of "g" responses with precursor phrases at three different speaking rates for the same-voice/same-location condition (upper left), the different-voice/same-location condition (upper right), the same-voice/different-location condition (lower left) and the different-voice/different-location condition (lower right). Error bars reflect standard error.

category boundaries; $F(1,91) = 10.15$, $p < .005$, $\eta_p^2 = .11$ by percentages). Both of these main effects, however, may be the result of general category boundary differences among listeners (for example, the participants hearing the female voice precursor may simply have had earlier category boundaries in general than those hearing the male precursor).

More importantly, however, there was a main effect of speaking rate ($F(2,182) = 71.09$, $p < .0001$, $\eta_p^2 = .44$ by category boundaries; $F(2,182) = 56.84$, $p < .0001$, $\eta_p^2 = .38$ by percentages). Overall, listeners had the earliest category boundaries in the fast precursor condition, and the latest boundaries in the slow precursor condition, as expected. Speaking rate also interacted with both talker ($F(2,182) = 2.87$, $p < .06$ (marginal), $\eta_p^2 = .05$ by category boundaries; $F(2,182) = 4.53$, $p < .02$, $\eta_p^2 = .06$ by percentages), and location ($F(2,182) = 4.42$, $p < .05$, $\eta_p^2 = .03$ by category boundaries; $F(2,182) = 5.77$, $p < .005$, $\eta_p^2 = .05$ by percentages), although both interactions were of a small effect size. The three-way interaction was not significant ($F(2,182) = 1.88$, $p > .10$ by category boundaries; $F < 1$ by percentages).

Given these overall effects and interactions, we then went on to explore the different conditions separately. For the same-voice/same-location condition, there was a significant effect of speaking rate in both the category boundary and percentage data (by category boundaries, $F(2,54) = 15.06$, $p < .0001$, $\eta_p^2 = .36$; by percentages, $F(2,54) = 21.67$, $p < .0001$, $\eta_p^2 = .44$). Follow-up t-tests showed that all three speaking rates differed from one another. The effect of the male voice was monotonic, with faster speaking rates in the precursor phrase producing fewer /g/ and more /k/ responses, as was expected from the prior literature.

For the different-voice/same-location condition, there was again a significant effect of speaking rate in both the category boundary and percentage /g/ data (by category boundaries, $F(2,30) = 25.45$, $p < .0001$, $\eta_p^2 = .63$; by percentages, $F(2,30) = 24.41$, $p < .0001$, $\eta_p^2 = .62$, a larger effect size than that for the same voice). Five of the 6 follow-up t-tests differed from one another significantly (all except the fast versus medium rate in the percent "g" response). Rate information from the female voice influenced listeners' judgment as to whether the consonant in the male voice was a /g/ or a /k/. As with the male voice, the effect of the female voice was monotonic, with faster speaking rates in the precursor phrase producing fewer /g/ and more /k/ responses.

Before moving on to the different-location conditions, we decided to compare the size of the effect for the same-voice and different-voice conditions. One way to do so is to simply compare effect sizes; looking at the partial η^2 values, it is clear that the rate effect in the female (different) voice is just as large as (indeed larger than) that in the male (same) voice (by category boundaries, η_p^2 of .63 for the female, .36 for the male; by percentage data, η_p^2 of .62 for the female versus .44 for the male). Another approach is to compare the two conditions statistically. Since the effects in both cases were monotonic, we ignored the medium rate, and simply looked at the size of the boundary shift between the two extreme precursor sentences (that is, the difference between the boundary location with the fast precursor phrase and that with the slow precursor phrase). This difference score provides us with a single value indicating the size of the normalization effect. For the same-voice condition, there was a shift of .26 in category boundary location, and 3% in "g" responses. For the different-voice condition, there was a shift of .38 in category boundary location, and 4% in "g" responses. These did not differ significantly (by category boundaries, $t(42) = 1.44$, $p > .10$; by percentages, $t(42) = 1.04$, $p > .10$), although the trend was again in the direction of larger effects in the different-voice condition (arguing that the absence of a greater effect size in the male voice was not for lack of power). Listeners showed the same size of effect of speaking rate when the rate information and target phoneme distinction came from the same talker as when they came from different talkers. Although both difference scores appear relatively small, the large effect sizes from the ANOVAs described above suggest that the effects are real and reliable, despite their small size.

For the same-voice/different-location condition, one-way ANOVAs showed a significant effect of precursor-phrase speaking rate in both the category boundaries, $F(2,62) = 20.03$, $p < .0001$, $\eta_p^2 = .39$, and in the percentage "g" responses, $F(2,62) = 7.11$, $p < .005$, $\eta_p^2 = .19$, both large effects (although the latter was smaller than in the same-location condition). There was no significant effect of which ear the precursor phrase occurred in, nor any interactions between ear and speaking rate, so all analyses were collapsed across ear. Follow-up t-tests showed that 5 of the 6 possible comparisons were significantly different from one another with only the medium versus the slow rate comparison in the percent "g" responses not being significant.

For the different-voice/different-location condition, one-way ANOVAs showed a significant effect of precursor-phrase speaking rate in both the category boundaries, $F(2,17) = 14.26$, $p < .0001$, $\eta_p^2 = .47$, and in the percentage "g" responses, $F(2,17) = 7.27$, $p < .005$, $\eta_p^2 = .32$, both large effect sizes (if smaller than that found for the same-location condition). There was again no effect of which ear the precursor phrase occurred in, nor any interactions between ear and speaking rate, so all analyses were collapsed across ear. Follow-up t-tests again showed that 5 of the 6 possible comparisons were significantly different from one another (the fast versus medium rate in the percent "g" response was not). Thus, rate normalization continues to occur even across changes in both talker and spatial location.

Did the change of location have any effect at all? In fact, it did. The change in spatial location appeared to reduce the overall size of the rate normalization effect for both same and different voices. To examine this, we first compared the overall size of the rate effect (the difference

between the fast and slow series) for the items varying only in spatial location. Comparing the same-voice/same-location condition to the same-voice/different-location condition, we found a significant difference in the percentage data ($t(58) = 2.53$, $p < .05$; for category boundary data, $t(58) = 1.01$, $p > .10$). Listeners' percentage responding changed by 3.2% in the same spatial location condition, but only by 1.5% in the different spatial location condition. Likewise, comparing the different-voice/same-location condition to the different-voice/different-location condition, we found a significant difference, this time only in the category boundaries (by category boundaries, $t(33) = 2.18$, $p < .05$; by percentage "g" responses, $t(33) = 1.46$, $p > .05$). If we instead compare effect sizes, we find that in the male voice, the category boundary data shows similar effect sizes for the same location and different location ($\eta_p^2 = .36$ versus .39), but the percentage data shows a smaller effect for the location change ($\eta_p^2 = .44$ versus .19). For the female voice, both types of data show smaller effects with a location change (by category boundaries, $\eta_p^2 = .63$ versus .47; by percentages, $\eta_p^2 = .62$ versus .32). It appears as if there is a trend towards less of a rate effect when there is a spatial location difference between the precursor phrase and the target word. Thus, while a change in spatial location did not entirely disrupt rate normalization effects, there was some evidence to suggest that it reduced the size of these effects.

It is not clear whether this is an overall reduction of the size of the effect on each trial, or whether the effect of speaking rate occurred on some trials but not others, leading to an overall reduction of the effect size. Furthermore, the reduction that was found could reflect changes in the use of the long-range information or the short-range information or both.

Despite this reduction in size, the fact that there were still significant overall effects in the different-location conditions suggests that the speech processing system requires a great deal of information in favor of segregation before different sound sources are treated as being completely distinct for the purposes of rate normalization. Moreover, the lack of any effect of talker voice suggests that talker identity is not taken into account for purposes of rate normalization. Listeners will use the speaking rate of one talker to help them interpret a segment produced by a different talker, even when the two talkers come from different apparent locations in space. Neither a change in sound source (talker), nor a change in location, was sufficient to eliminate rate normalization, and a change in talker did not even substantially reduce rate normalization.

As a final analysis, we compared the condition where two factors, voice plus spatial location, provided evidence for segregation, to the condition where only spatial location differed (that is, the different-voice/different-location condition versus the same-voice/different-location condition). Here we saw no difference: by category boundaries, $t(45) = .63$, $p > .50$; and by percentages, $t(45) = .68$, $p > .48$. Thus, there is no indication that adding voice distinctiveness on top of a spatial location cue decreases the effect of a precursor speaking rate. A change in the identity of a talker is apparently not a sufficient cue to disrupt rate normalization, even when added onto other cues for perceptual segregation.

This pattern of results appears to be quite comparable to that found by Sawusch and Newman (2000) for rate normalization on the basis of following speech information. (Moreover, that study used a female target voice, so there is little reason to think the gender of the target voice here would have influenced the results.) The fact that the different-voice condition was not reliably different from that of the same-voice condition suggests that both the long-range and short-range influences of speaking rate occurred even when the voice changed from the precursor to the target. These results are also consistent with Wade and Holt (2005) who found that the rate of a nonspeech tone sequence could alter phonetic perception. With phrasal duration precursors, the speech processing system appears to use whatever speaking rate information it has available to it as a source for rate normalization, regardless of its actual relevance.

However, in Experiment 1 and these previous studies, there was no competing organization to keep the target separate from the precursor. That is, there was still only one stream of sound present at any point in time, and only one source of rate information available. Moreover, the listeners were well aware of the locations from which the stimuli would be arriving (since this was constant across the experiment), and could have therefore chosen to focus their attention on the particular spatial location that was present in these stimuli, in a manner that might be different from typical real-world listening. Perhaps if both the male and the female voices had been present during the precursor period then listeners would have been less likely to use rate information from the mismatching voice. Experiments 2 and 3 investigate this possibility.

3. Experiment 2

In Experiment 1, listeners used rate information from a female precursor phrase (or from a precursor phrase coming from a different location) to help them interpret a duration-based contrast in a different (male) following voice. Yet there was no alternative source of rate information for listeners to use in that experiment. A number of studies have suggested that rate normalization is an obligatory process, making use of whatever information is available at the time that the listener processes and responds to the target (Miller & Dexter, 1988; Sawusch & Newman, 2000). Had there been information from both voices present during the precursor period, however, listeners may have been less likely to use the information that came from an incorrect voice. It is possible that rate normalization will follow whatever perceptual grouping has been already established. If the precursor phrases provide evidence that there are two different streams of

speech present, rate normalization may be limited to information within a single stream.

As a starting point to investigate these issues, we decided to pit an instantiation of two different segregation cues against one another. Listeners heard two different voices speaking simultaneously during a precursor phrase: a male voice and a female voice. One voice was presented in the right ear, while the other voice was presented in the left ear. The target syllable was presented in the male voice, as in the previous studies. However, this target syllable was presented in the same ear as the female voice precursor phrase. Thus there were two different cues to segregation present in this experiment: listeners could segregate the two talkers by either voice cues or by spatial location cues.

If listeners rely primarily on voice cues to guide their rate normalization, we might expect that the speaking rate of the male voice would influence the perception of the final syllable in that same voice. The speaking rate of the female voice would not be expected to have an effect in this case on the male voice target. In contrast, if listeners rely primarily on spatial location cues to guide their rate normalization, we would expect that the speaking rate of the female voice would influence the perception of the final syllable in the male voice, as both appear to originate at the same location in space. The speaking rate of the male voice in the other ear would not be expected to have an effect. Finally, listeners could be influenced by the speaking rate of both of the two precursor voices. This could suggest that neither spatial location nor talker identity provide sufficient cues for perceptual segregation with respect to speaking rate normalization. It might also suggest the possibility that normalization occurred prior to the point in time at which the target syllable was grouped with only one of the two precursor voices. Put another way, since both precursor phrases exhibit phonetic coherence with the target syllable (cf. Remez et al., 1994), both precursor phrases may influence listener responses to the target.

3.1. Method

3.1.1. Listeners

The listeners were 31 members of the University of Maryland community who participated in exchange for a cash payment. All were native speakers of English with no reported history of either a speech or a hearing disorder. Data from two additional participants were excluded for being a nonnative speaker ($n = 1$) or failure to identify the endpoint items at 80% accuracy ($n = 1$). Eight of the 31 participants had been in previous versions or pilot versions of this study.

3.1.2. Stimuli

The /gaɪp/–/kaɪp/ test series and the female precursor phrase from Experiment 1 were used here. In addition, the male talker recorded a new precursor phrase, ''You wrote to me and said gipe.'' This was recorded at three different speaking rates, as before, and the final syllable was excised, resulting in a new precursor phrase containing the same number of syllables as the one used in Experiment 1 (''I heard him say the word___''). The new precursor phrase was recorded so that the words and phonemes in the two voices, when presented dichotically, would be different. This, in turn, should reduce the likelihood that the information in the two voices would blend as opposed to being heard as two separate streams of speech. The amplitudes of the two precursor phrases were set so as to sound similar and the peak amplitudes were within .5 dB of one another.

Both precursor phrases end with the same syllable–final stop consonant. This allowed the two sequences to be presented dichotically such that they ended with the same consonant at the same point in time. The transitions into the /d/ were slightly different (because the words ''said'' and ''word'' have different vowels); transition durations ranged from 16 to 32 ms in the male voice, but 29–45 ms in the female voice. Each of the three speaking rate versions of the male precursor phrase was paired with each of the three versions (speaking rates) of the female precursor phrase, resulting in nine different combinations. As in Experiment 1B, the precursor was followed by a constant, 50 ms closure interval and then the target syllable.

It is worth noting that the precursor phrases were similar, but not identical, in length, as shown in Table 1. Since the two sequences ended at the same point in time, but were not identical in duration, they did not begin at the same point in time, even in the combinations in which both were spoken at the same rate.

The eight members of the /gaɪp/–/kaɪp/ series were appended to each of the nine dichotic precursors, resulting in a total of 72 precursor-target items. Unlike in the previous experiments, however, the final syllable in the male voice was presented in only one earphone, the same that presented the female voice precursor. This may be represented as

MALE VOICE, LEFT EAR:	You wrote to her and said
FEMALE VOICE, RIGHT EAR:	I heard him say the word
MALE VOICE, RIGHT EAR:	gipe

3.1.3. Procedure

Listeners heard a practice block of 24 selected items, followed by 6 blocks of all 72 items presented in random order. The presentation of precursor voice to ear was counterbalanced across participants (a 15/16 split, with 16 hearing the male precursor on the right, and female precursor/male target on the left and 15 hearing the reverse). All other aspects of the procedure were identical to that in Experiment 1.

Following the experiment, listeners were asked to identify what strategy they used during the experiment. They were asked whether they found themselves listening to a particular ear, a particular voice, to both sentences, or whether they used another strategy altogether.

3.2. Results and discussion

In terms of strategies used, the majority of participants ($N = 16$) reported that they attempted to ignore both precursor voices and attend only to the critical word. One reported listening to both precursor sentences, 8 reported attending to a specific ear, and 4 reported attending to a particular voice. Finally, 2 neglected to provide strategy information. A slightly higher ratio of those who had previously been in another (earlier) study reported ignoring the precursor voice (6 of 8, 75% versus 10 of 23, 44%), but this difference was not significant by Fisher's exact test.

The basic data reduction for the rating data was identical to that in Experiment 1. There was no effect of ear, nor any interaction between ear and any other factor. There was also no effect or interactions on the basis of reported strategy, nor any overall effect of having been in previous studies. There was a single interaction between previous participation and the effect of the different voice, which will be discussed later. Given the general lack of effects of these factors, we collapsed across them in the overall analysis. We performed two 3×3 ANOVAs (one based on category boundary data, the other on percentage "g" responding), with the within-subject factors of same-voice speaking rate and different-voice speaking rate.

There was a significant effect of the same (male) voice, even though it occurred in the ear opposite the target word (by category boundaries, $F(2,60) = 27.98$, $p < .0001$, $\eta_p^2 = .48$; by percentage responding, $F(2,60) = 15.59$, $p < .0001$, $\eta_p^2 = .94$), as seen in the left panel of Fig. 4. Follow-up tests showed that the fast speaking rate differed significantly from both the medium and slow speaking rates and the latter two did not differ from one another.

There was also an effect of the female precursor speaking rate (by category boundaries, $F(2,60) = 40.97$, $p < .0001$, $\eta_p^2 = .58$; by percentage responding, $F(2,60) = 33.47$, $p < .0001$, $\eta_p^2 = .78$), as seen in the right panel of Fig. 4. Follow-up t-tests showed that 5 of the 6 comparisons differed significantly, with only the fast versus medium rate in the percent "g" responses as an exception. Thus, even though listeners were using the speaking rate of the male voice to help them interpret the /g/–/k/ contrast, this did not prevent them (as a group) from also using the speaking rate of the female voice.

Finally, there was also an interaction between the two precursor speaking rates (by category boundaries, $F(4,120) = 7.28$, $p < .0001$, $\eta_p^2 = .19$; by percentage responding, $F(4,120) = 5.84$, $p < .0005$, $\eta_p^2 = .85$). Fig. 5 shows the influence of the female voice's speaking rate (separate curves for fast, medium and slow) as a function of male voice speaking rate (on the x-axis). The female voice had its greatest effect when the male voice was at a moderate (intermediate) speaking rate. When the male voice was at a slow speaking rate (points on the left side of Fig. 5), the female voice had a much smaller influence on listeners' responses. This might indicate that an extreme speaking rate in the male voice provided less opportunity for the speaking rate of the female voice to alter perception, although it could also be an accident of the particular tokens used here.

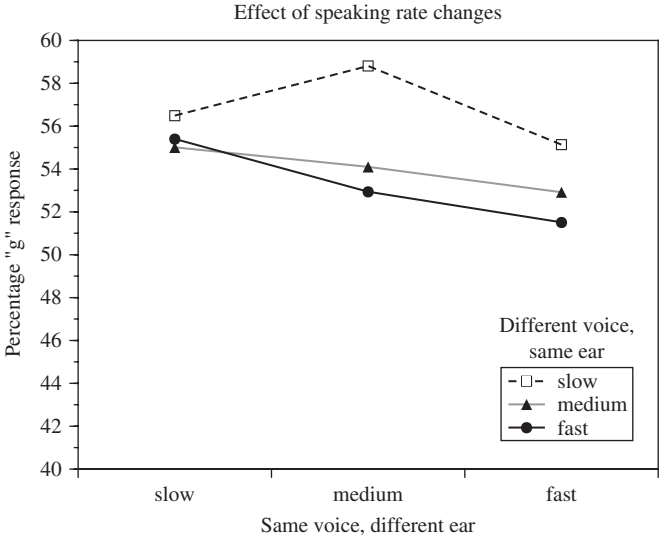

Fig. 5. Interaction between the speaking rate of the male and female voices in Experiment 2.

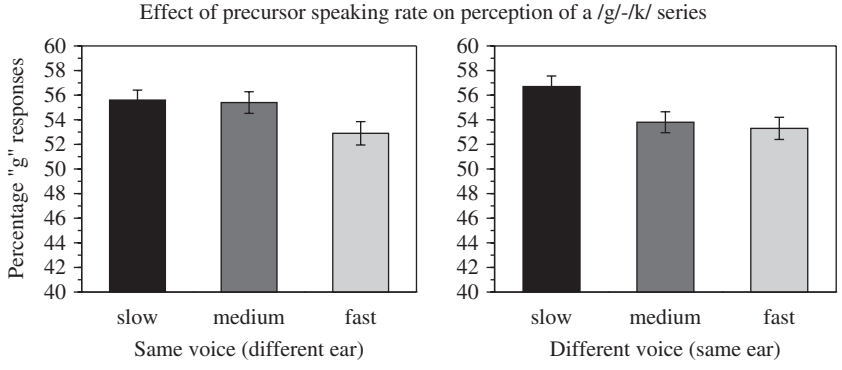

Fig. 4. Percentage of "g" responses with precursor sentences at three different speaking rates for the same (male) voice in a different ear (left) and for the different (female) voice in the same ear (right). These graphs show main effects only (averaging across the alternate voice). Error bars reflect standard error.

Overall, the size of the effect in the female voice did not differ from that in the male voice, although there was a trend for a slightly larger effect in the female voice (by category boundaries, $t(30) = 1.72$, $p < .10$; by percentages, $t(30) = 1.76$, $p < .10$). The size of the effect with the female voice precursor also did not differ from the effect with the female voice in Experiment 1, suggesting that the presence of an additional (second) voice did not reduce the size of the cross-voice effect (the interaction between study and speaking rate was $F < 1$ for both category boundaries and percentage data).

As mentioned before, the initial analyses suggested that there was an interaction between previous participation and the effect of the different voice. This was examined with two $3 \times 3 \times 2$ ANOVAs with the within-subject factors of speaking rate for the female voice and speaking rate for the male voice, and the between-subject factor of whether the individual participant had been in any prior rate normalization studies in the lab. This analysis replicated the overall effects in both voices and the interaction between voice that was described above. There was also an interaction between previous participation and the size of the effect in the different (female) voice (by category boundaries, $F(2,56) = 4.40$, $p < .05$, $\eta_p^2 = .14$; by percentages, $F(2,56) = 5.28$, $p < .01$, $\eta_p^2 = .16$). Follow-up tests comparing the size of the effect in the female voice revealed that those individuals who had been in one of the previous studies showed a smaller effect of the speaking rate of the female voice (by category boundaries $t(27) = 2.20$, $p < .05$; by percentages, $t(27) = 2.20$, $p < .05$). Individuals who had not been in any prior studies showed a 4.3% shift in responding on the basis of the female talker's speaking rate (.36 of a unit by category boundaries). Those who had been in prior studies showed a smaller shift of only 1.8% (.17 by category boundaries). Knowledge about the purposes of the study may have altered listeners' strategies, causing them to place slightly less emphasis on the use of the female voice. This implies that strategic knowledge may play a role in listener's ability to exclude information from other voices, but further studies will be needed to test this explicitly, especially given the lack of any effect of reported strategy itself.

Before proceeding with Experiment 3, one aspect of the results of Experiment 2 needs some further exploration. Aside from the influence of participating in a previous experiment, does any other strategic or presentation factor dominate or determine the pattern of results? One possible factor has to do with the durations of the two phrases on each trial. The rates of speaking of the two voices were independent and only the offsets of the two phrases were synchronized. Consequently, one of the two phrases always started before the other. Could the voice that started first have captured the listeners' attention and determined the pattern of the results? Since the female voice had a slightly longer overall duration at each speaking rate, it would have started first on two-thirds of the trials overall (in all cases except when the male voice was presented slowly and the female was presented at a moderate or fast rate, or when the male voice was presented at a moderate rate and the female voice at a fast rate). This is consistent with a slightly larger overall influence of the female voice speaking rate on listeners' responses to the target. Moreover, when the male precursor was long in duration (slow rate), the influence of the female voice was small, and only occurred when the female voice was also at a slow rate (and thus would have started first). The listener data for the fast and medium female precursors with the slow male precursor showed no difference, as would be expected if there were an influence of whichever voice that started first. Conversely, by this account the greatest influence of the male voice should occur when the female precursor was shortest (fast speaking rate) since this would allow the male precursor to start first at its slow and intermediate rates. The solid line with the circles in Fig. 5 seems to show the largest influence of the male speaking rate (steepest slope left to right) in just this situation.

However, the pattern of data for the female voice with the medium and fast rate male precursors does not follow the predicted pattern. Based on the idea that the voice that starts first has the greatest effect, the largest influence of the female voice should be with the fast male precursor (since in this situation, the female voice would always start first). However, the greatest influence of the female voice was actually found for the medium-rate male precursor, as shown in Fig. 5. Furthermore, the influence of the slow female precursor was substantially reduced with the fast male precursor. Consequently, the data cannot be described as simply showing an effect of the voice that started first, although this may be one factor that influences the data.

Were listeners actually using both voices at the same time? We have no way of accurately measuring whether listeners were truly being influenced by both duration sources simultaneously, or were being influenced by one precursor duration on some trials and the alternative precursor on other trials. Thus, we cannot say that the rate normalization mechanism actually takes in information from two potential sound sources at the same time. Rather, we can conclude that when there are multiple sources of information available, it is not limited to using information only from a source that matches in voice (or in location).

4. Experiment 3

In Experiment 2, listeners as a group used the speaking rate of both precursor voices to adjust their perception of the final duration-based contrast. Both a voice that mismatched in spatial location and a voice that mismatched in talker identity were used for rate normalization. Yet in both of the experiments thus far, there were some cues that supported grouping each precursor phrase with the target voice. In Experiment 1, the precursor phrase was the only precursor present and there was no competing voice or perceptual group to encourage segregation. In

Experiment 2 there were competing sound sources, but each stream matched the target in one feature, either voice or spatial location. Thus each precursor phrase had some cues that supported grouping with the target syllable at the same time that other cue(s) supported segregation.

The present experiment examines the case where all of these cues support segregation. As in Experiment 2, two voices were presented during the precursor phrase. One phrase matched in both talker identity and spatial location; the other phrase differed in both features. We expected that listeners would use the duration information from the "correct" voice in their rate normalization. The real question is whether the speaking rate of a voice that is clearly separate from the continuous, target-carrying voice could still influence rate normalization of the target. In other words, would listeners use rate information from a clearly "wrong" voice?

4.1. Method

4.1.1. Listeners

Thirty members of the University of Maryland community participated in exchange for extra credit or a cash payment. All were native speakers of English with no reported history of either a speech or a hearing disorder. Data from an additional 6 participants were excluded for being a nonnative speaker (3), for having a history of language or speech disorders or attention deficit disorder (2), or for demonstrating poor accuracy at identifying the endpoints of the series (1). None of the participants had been in previous versions of this study.

4.1.2. Stimuli

The gipe–kipe series and precursor phrases from Experiment 2 were used for this experiment, except that the ears for the two precursors were reversed. In this manner, the male precursor phrase now matched the final word in both talker voice and spatial location, and the female precursor phrase now mismatched on both features. This may be represented as

MALE VOICE, LEFT EAR: You wrote to her and said gipe
FEMALE VOICE, RIGHT EAR: I heard him say the word

4.1.3. Procedure

The procedure was identical to that in Experiment 2. One half of the listeners heard the male voice in the left ear and female voice in the right. For the other half of the listeners the voice-to-ear assignment was reversed.

4.2. Results and discussion

The basic data reduction was identical to that in the prior experiments. Surprisingly, preliminary analyses showed that there was an overall effect of ear (that is, which ears the voices occurred on), ($F(1,28) = 6.83$, $p < .05$, by category boundaries; $F(1,28) = 5.53$, $p < .05$, by percentages). It is not clear what this overall effect of ear implies, particularly as they were different participants; apparently, some participants had relatively later category boundaries (more /g/ responses) than others. However, this did not interact with any other factor, so we collapsed across this factor in the final analysis.

Interestingly, there was a significant interaction between listeners' reported strategy and the effect of speaking rate. Somewhat surprisingly, participants did not all report attending only to the correct voice (or ear). Although that was the predominant strategy (21 of 30 participants reported listening to the male voice), 5 participants reported listening to both voices, 3 participants reported ignoring both voices, and 1 reported listening to both voices when the sentences were fast but the female voice when it was slow. Given the interaction, this factor of strategy was retained in the final analysis. The data were therefore examined with two $3 \times 3 \times 3$ ANOVAs with the within-listener factors of speaking rate of the male voice (fast, medium, and slow) and speaking rate of the female voice (fast, medium, and slow), and the between-listener factor of reported strategy, which was grouped into 3 levels (attention to the male voice, to both voices, or to neither voice).

As expected, there was a significant effect of the speaking rate of the male voice (the voice which, according to all grouping cues, should be grouped with the final syllable), as shown in the left panel of Fig. 6. This was significant in both the category boundary data ($F(2,54) = 136.70$ $p < .0001$, $\eta_p^2 = .84$) and percentage "g" responding ($F(2,54) = 145.94$, $p < .0001$, $\eta_p^2 = .47$). Follow-up t-tests showed that all speaking rates differed significantly from one another. As the male voice spoke at a slower rate, there was a later category boundary, and more items were labeled as "g." This replicates the basic precursor normalization effect reported by Kidd (1989) and Summerfield (1981), and shown in the same-voice/same-location condition in Experiment 1.

This effect of the male voice was moderated by an interaction with strategy that was significant in the percentage data only ($F(4,54) = 1.72$, $p > .10$, $\eta_p^2 = .11$ by category boundaries; $F(4,54) = 2.55$, $p < .05$, $\eta_p^2 = .16$ by percentages), as seen in Fig. 7. Generally, those who attempted ignoring both voices (that is, neither listening to nor attending to either voice) actually showed a larger effect of the male voice than did those who reported attending to the male voice (a 16.2% difference between fast and slow speaking rates for those who attended to neither voice, a 12.3% difference for those attending to both, and a 10.9% difference for those attending to the male voice). Those listening to the male voice showed a significantly smaller effect of the male speaking rate than those attending to neither voice, $t(22) = 2.90$, $p < .01$. The effect size shown by those listening to both voices did not differ from either of the other two groups (both versus male, $t(25) = .80$; both versus neither, $t(7) = 1.11$, both

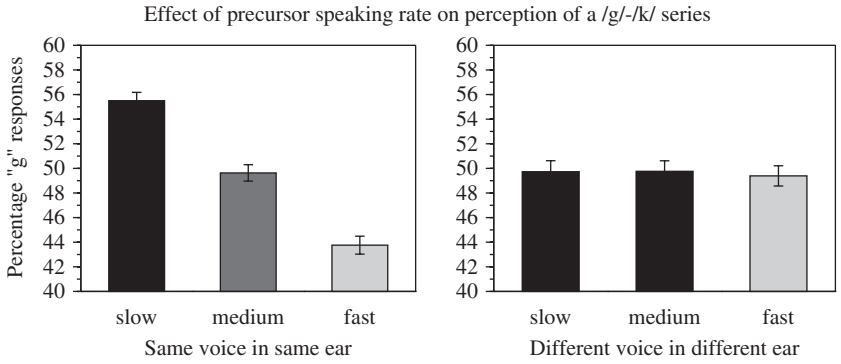

Fig. 6. Percentage of "g" responses with precursor sentences at three different speaking rates for the same voice/ear (left) and the different voice/ear voice (right). These graphs show main effects only (averaging across the alternate voice). Error bars reflect standard error.

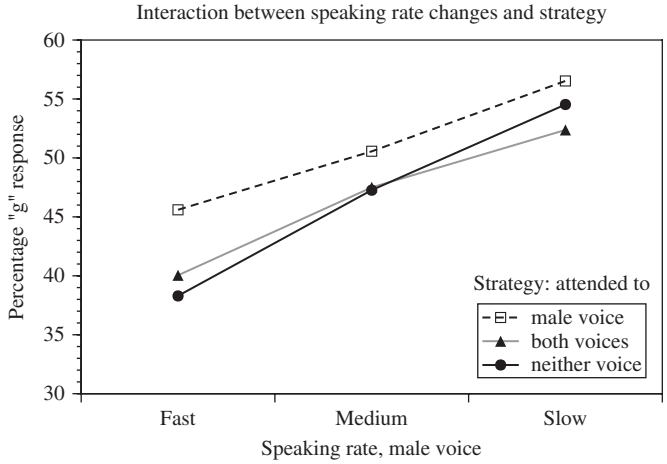

Fig. 7. Interaction between the speaking rate of the male voice and the listener's reported strategy.

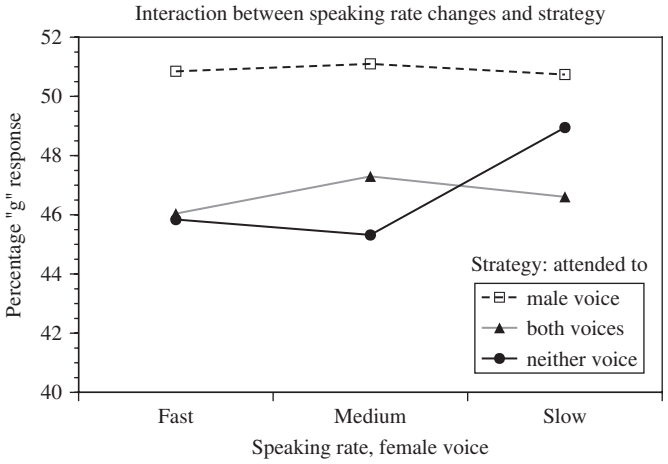

Fig. 8. Interaction between the speaking rate of the female voice and the listener's reported strategy.

$p > .05$). It is not clear what might be the reason for this unusual pattern. Since participants were not randomly assigned to a particular strategy, this may reflect other aspects of those participants or the use of different strategies on different trials.

It is worth noting that the effect in the male voice appears to be much larger than that found for any other experiment in this set, including the classic rate normalization case of the same-voice/same-location condition in Experiment 1 (roughly an 11% change versus a 3% change). It is not clear why this might be the case. Although it is tempting to argue that the smaller effects of the wrong-voice and wrong-location conditions in Experiments 1 and 2 are the result of the fact that these precursors are being segregated from the target word to some degree, this cannot explain the difference between the size of the effect in this experiment and that of the male voice in Experiment 1, since in both cases the precursor phrase was in the correct voice and the correct location. Since the two precursor phrases are different sentences, and the participants in the two studies differed, it is not clear how much to make of this difference. However, one possibility is that the presence of another voice (that also varied in speaking rate across trials) highlighted the rate variation occurring in this study.

Surprisingly, there was an effect of the "incorrect" female voice, although this was significant only in the category boundary data (by category boundaries, $F(2,54) = 4.49$, $p < .02$, $\eta_p^2 = .14$; by percentages, $F(2,54) = 1.86$, $p > .10$, $\eta_p^2 = .06$). Category boundaries were located at 4.45 stimulus units for the fast speaking rate, 4.46 for the medium speaking rate, and 4.50 for the slow speaking rate. The small overall effect of the female voice's speaking rate is shown on the right side of Fig. 6.

There was also a marginal interaction between the female voice precursor speaking rate and strategy ($F(4,54) = 2.48$, $p < .06$, $\eta_p^2 = .16$ by categories, $F(4,54) = 2.01$, $p < .11$, $\eta_p^2 = .13$ by percentages). Those individuals who reported listening to neither voice showed a larger effect of the female talker's voice than those individuals who reported listening only to the male voice or to both voices, as shown in Fig. 8. The size of the influence of the female speaking rate for those who reported listening to the male voice was significantly smaller than the effect for listeners who reported attending to neither voice, $t(22) = 2.11$, $p < .05$. The effect of the female speaking rate

for participants who reported listening to both voices did not differ from either of the other two groups (both versus male, $t(25) = .30$; both versus neither, $t(7) = 1.55$; both $p > .05$). Put another way, those individuals who reported trying to ignore both voices actually showed larger effects of both the male voice's speaking rate and the female voice's speaking rate. Listeners' shifts in category boundaries from the slow to fast female speaking rates were only .02 units for those who attended to the male voice, .04 units for those who reported attending to both voices, but .31 units for those who reported attending to neither voice (based on percentage responding, these values were −.11%, .57%, and 3.11%). There was no interaction between the male and female speaking rates, $F(4,108) = 1.69$, $p > .05$ by category boundaries, $F < 1$ by percentages, and no three-way interaction, $F(8,108) = 1.76$, $p < .10$ by category boundaries, $F < 1$ by percentages.

Despite the significant overall effect of the female voice in the category boundary data, the effect of the male voice was considerably larger. Participants in this experiment were primarily influenced by the "correct" voice. This suggests that, in multi-talker environments outside the laboratory, listeners are unlikely to show substantial effects of the speaking rate of voices other than the one they are actually attending to. Nonetheless, finding *any* significant effect of the female voice in this experiment suggests that even clearly incorrect voices may occasionally influence perception.

5. General discussion

The present studies explored how listeners adjust for speaking rate in multi-talker environments. Or, to put it another way, the studies explored the interactions between the processes of speaking rate normalization and perceptual grouping. Experiment 1 demonstrated that speaking rate information from a precursor phrase in one voice influences phonetic perception in a subsequent voice. Likewise, speaking rate information from one location in space influences phonetic perception in a subsequent location in space. The difference in spatial location appears to have reduced, but not eliminated, this speaking rate effect. These data are consistent with the proposal that the use of duration information in speaking rate normalization is a relatively early part of perception that makes obligatory use of whatever duration information is available (Miller & Dexter, 1988; Sawusch & Newman, 2000). Even when spectral and/or spatial information suggests that the information is from a different sound source in the environment, listeners nonetheless continued to use that information for speaking rate normalization to some degree. That said, low-level perceptual grouping information (such as spatial location) appeared to play a role in the degree to which that duration information had an effect. There was a smaller effect of rate normalization when the location changed. In contrast, the effects of both same-voice and different-voice precursors were virtually identical and a change in voice did not diminish the influence of the speaking rate of the precursor upon the target.

Interpreting these findings from Experiment 1 to some degree depends on whether the change in voice was sufficient to result in the perception of two streams at some level of processing. Perhaps a change in location results in the perception of two source streams in a way in which a change in talker does not. Prior research has certainly suggested that changes in talker (as well as changes in spatial location) are audible and can serve as cues for perceptual grouping, although we did not assess this directly in the present study. Even though we do not have an independent measure of stream segregation (other than its effect on rate normalization), prior research (Dorman et al., 1979) has demonstrated that a change in voice can result in stream separation at the level of phonetic perception. Clearly, then, such a manipulation could influence rate normalization as well; that it seems not to do so is unlikely to be an indication that the manipulation is inaudible but rather an indication of the role of auditory stream formation in perception. In the present studies, the subjective impression of two talkers (and, by implication, two streams of speech) seems to have occurred after the process of speaking rate normalization had already influenced perception of the target.

Experiments 2 and 3 demonstrate that rate normalization is not completely limited to information from a single sound source. When two sources of speech were presented simultaneously, listeners were influenced by both of them. As noted earlier, it could be the case that both sources of rate information were used within a trial or that different sources were used on different trials or that both possibilities can occur and the influence of the precursor in a multi- (two or more) talker situation is heavily modulated by attention.

However, Experiment 3 also suggests that there are clear limits to this process. When all of the available acoustic cues (voice and location in space) suggest that one source of sound is relevant for rate normalization and another is not, listeners do not treat the sources equivalently. In Experiment 3, perception was dominated by effects of the appropriate (male) voice. Surprisingly, though, there was still a very small influence of the female voice in this study. Even in the case where one precursor voice matched the target word in both talker identity and spatial location, a voice that mismatched on both dimensions still had a small effect on listeners' perception.

Despite this effect of the duration of a "wrong voice", rate normalization is clearly not occurring without reference to grouping cues. Spatial location cues, in particular, seem to play a large role in listeners' rate normalization. In Experiment 1, effects of the precursor duration were reduced when a spatial location change occurred (regardless of talker voice). Moreover, in Experiment 3, a precursor that mismatched in both voice and spatial location had only a very small effect on rate

normalization, much smaller than that of the appropriate voice in the same spatial location as the target.

These studies provide information about the role of various grouping processes in speech as indexed by speaking rate normalization. First, the number and type of cues supporting separation versus grouping appears to be an important factor in the degree to which two voices are kept separate during phonetic processing. In Experiment 1, information from a voice that differed in spatial location from the final item was used to a lesser extent than speaking rate information that came from the same location in space. In Experiment 3, a precursor that immediately preceded the target word, but differed in voice and location, had a much smaller effect than a simultaneous precursor that matched on both factors. These findings suggest that grouping of information for phonetic perception is influenced by fairly low-level properties of the signal. Yet information from a wrong source continued to have an influence, despite the fact that it was clearly treated differently than information from a correct voice. That is, even though the processing system had determined that the female voice in Experiment 3 was not the appropriate source of information (since this information was used to a lesser extent than the male voice), this information was not entirely ignored. One way of reconciling these apparent discrepancies is to assume that grouping is not a single process, but may instead be influenced by information at multiple stages of speech processing. Even though the female precursor had been at least partially separated from the male talker at the point when normalization occurred, it had not been fully separated. If this explanation is in fact true, future research will need to explore what cues to perceptual grouping are used at different points in perceptual processing.

Although the experiments described here were not explicitly designed to separate the influences of the rate of stressed syllables in the precursor from the duration of the speech segments right before the target (the long-range and short-range systems described by Kidd, 1989), it is tempting to speculate on the role of these two systems in the results. We propose that the long-range influence of the stressed syllable rate in the precursor is more likely to be disrupted by the allocation of attention and strategy used by the listener and by processes that contribute to auditory stream formation and perceptual grouping late in processing. Thus, even when the precursor and the target appear to originate at different points in perceptual space, the precursor can influence the long-range normalization process when the listener groups the precursor and the target together based on continuity in voice or the selective allocation of attention. Prior research suggests that the influences of immediately adjacent segment durations in the short-range system, in contrast, are relatively uninfluenced by strategy or the allocation of attention and are primarily influenced by stimulus factors (cf. Miller & Dexter, 1988; Sawusch & Newman, 2000). Continuity in spatial location may be one such stimulus factor. Thus some of the reduction in influence of precursor speaking rate with a change in spatial location (Experiments 1 and 3) may have come from the short-range system.

Another stimulus-driven factor may be the duration of the precursor. In these studies the phrase contained six syllables that, together, varied between 750 and 1250 ms in duration (see Table 1). In Wade and Holt (2005), the precursors were all about 1200 ms duration, and made up of either 10 long or 30 short tones. Thus, both of the studies that have found an influence of a precursor when the source changed from precursor to target used long precursors with multiple "segments". In contrast, Dorman et al. used a single-syllable precursor (durations were not provided), and Diehl et al. (1980) used a three-syllable precursor (with durations of 445 and 730 ms for fast and slow rates). One way to resolve the discrepancy between the rate effects found here (and in Wade & Holt, 2005) but not by Diehl et al. (1980), and between the phonetic integration across talkers found in Experiment 1 but not by Dorman et al. (1979), is to propose that coherent longer precursors capture processing and promote the formation of a single stream that will bridge other stimulus manipulations such as a change in location or talker. Further empirical investigation is needed, however, before accepting this proposal.

From an applied perspective, these results also provide a novel reason for why listening in an environment with multiple conversations can be a difficult task. Most explanations have focused on masking effects between voices (in cases where the voices spoke simultaneously), or on the requirement to adjust perception for the presence of a new and different voice (for tasks in which one voice followed another, see Sommers, 1997). The present study suggests that background voices, even when clearly distinct from the target voice, may also influence normalization processes, and thus may alter perception of an attended voice. All that is required is that the listener momentarily alter his or her attentional allocation so that information from another voice is processed or that there are stimulus conditions that promote momentary grouping of information from an irrelevant talker with the speech of the target talker. Either or both of these circumstances could alter a listener's perception of fluent speech segments.

Future research should also explore the role of attention on rate normalization. The significant interactions with strategy in Experiment 3 suggest that attention may influence perception in some situations. However, since individuals were not assigned randomly to any particular strategy, this may simply be an artifact. Moreover, although most listeners reported attending to the male voice, they likely had no strong incentive to focus on that voice alone. When trying to listen to a friend at a party, an individual may invoke more effort to focus on that particular voice than in our repetitive lab setting, and this may moderate the extent to which information from an incorrect voice can influence perception. Future research

should investigate this aspect directly, by determining whether instructional manipulations or selective payoffs for performance could eliminate the effect of the inappropriate precursor voice.

Another issue for future research has to do with the length of time over which these voice interactions occur. In the present studies, the change in voice occurred immediately prior to the closure for the target distinction. Presumably, rate information from a different voice would only be used for some limited time span. Information from an incorrect voice that is temporally removed from the target would not be used. Future research will be needed to map the time course over which this cross-voice information is used, and whether this time course actually differs for a different voice as compared to the same voice.

Interactions across voices might also potentially change across the lifespan. Elderly adults have been shown to have particular difficulty with changes in talker (Sommers, 1997), suggesting they are not able to adjust for a new talker's voice as quickly as a young adult can. This might imply that they would continue to show the cross-voice rate effect even in situations in which younger adults would not do so, or that this cross-voice effect would continue across longer intervening sequences for elderly listeners than for young adult listeners. Since using speaking rate information from an incorrect voice is only likely to lead to false interpretations when the voices actually differ in speaking rate, this also implies that older listeners' difficulties listening in multi-talker environments would be exacerbated (relative to young adults) when the voices were more different from one another in features such as dialect, speaking rate, or spectral information.

In conclusion, the present studies examined when (and how) speaking rate information from one talker influenced the perceptual processing of an alternate stream of speech. Results are generally consistent with the notion that rate normalization is an early occurring, obligatory process that makes use of whatever information is available to it at the time that phonetic processing of a target is completed. When multiple, consistent sources of information are present, rate normalization appears to be based primarily on information from the appropriate stream of speech. Thus, while rate normalization may make use of incorrect information when that is the only information available, or when different cues to stream segregation favor different voices, it is less likely to do so when there are multiple, convergent cues to a single voice as the appropriate source of speaking rate information. Despite this fact, even information from an alternate stream of speech can influence perception, at least to a small degree.

Acknowledgments

The research reported here was partially supported by NIH (NIDCD) Grant R01-DC00219 to the University at Buffalo. Parts of this research were previously reported at the 143rd meeting of the Acoustical Society of America, June 2002. The authors wish to thank Nina Azhdam, Christine Beagle, Emilie Clingerman, Nicole Craver, Eva Derecskei, Hayley Derris, Annie Ferruggiaro, Sarah Haszko, Maria Hernandez, Lacey Kahner, Becky Karman, Micaela Knebel, Hannah Kim, Lisa Loder, Keren Malaky, Jamie Mowbray, Robin Nicoletti, Jessica Pecora, Jamie Ratner, Antonia Rodriguez, Kate Shapiro, Lauren Simpson, Emily Singer, Cheryl Tarbous, Stephanie Weinberg, Donnia Zack-Williams, and Jenni Zobler for assistance in subject running.

References

Ainsworth, W. A. (1972). Duration as a cue in the recognition of synthetic vowels. *Journal of the Acoustical Society of America*, *51*, 648–651.

Alain, C., & Arnott, S. R. (2000). Selectively attending to auditory objects. *Frontiers in Bioscience*, *5*, d202–d212.

Boersma, P., & Weenink, D., (2005). *Praat: Doing phonetics by computer (Version 4.3.28)* [*computer program*]. Retrieved from 〈http://www.praat.org/〉.

Bregman, A. S. (1990). *Auditory scene analysis: The perceptual organization of sound*. Cambridge, MA: MIT Press.

Bregman, A. S., Abramson, J., Doehring, P., & Darwin, C. J. (1985). Spectral integration based on common amplitude modulation. *Perception & Psychophysics*, *37*, 483–493.

Bregman, A. S., & Pinker, S. (1978). Auditory streaming and the building of timbre. *Canadian Journal of Psychology*, *32*, 19–31.

Broadbent, D. E. (1952). Listening to one of two synchronous messages. *Journal of Experimental Psychology*, *44*, 51–55.

Broadbent, D. E. (1954). The role of auditory localization in attention and memory span. *Journal of Experimental Psychology*, *47*, 191–196.

Broadbent, D. E., & Ladefoged, P. (1957). On the fusion of sounds reaching different sense organs. *Journal of the Acoustical Society of America*, *29*(5), 708–710.

Brokx, J. P. L., & Nooteboom, S. G. (1982). Intonation and the perceptual separation of simultaneous voices. *Journal of Phonetics*, *10*, 23–36.

Brungart, D. S. (2001). Informational and energetic masking effects in the perception of two simultaneous talkers. *Journal of the Acoustical Society of America*, *109*(3), 1101–1109.

Brungart, D. S., & Simpson, B. D. (2007). Effect of target-masker similarity on across-ear interference in a dichotic cocktail-party listening task. *Journal of the Acoustical Society of America*, *122*(3), 1724–1734.

Cherry, E. C. (1953). Some experiments on the recognition of speech, with one and with two ears. *Journal of the Acoustical Society of America*, *25*, 975–979.

Ciocca, V. (2008). The auditory organization of complex sounds. *Frontiers in Bioscience*, *13*, 148–169.

Crystal, T. H., & House, A. S. (1982). Segmental duration in connected-speech signals: Preliminary results. *Journal of the Acoustical Society of America*, *72*, 705–716.

Crystal, T. H., & House, A. S. (1988). Segmental durations in connected-speech signals: Current results. *Journal of the Acoustical Society of America*, *83*, 1553–1573.

Cusack, R., Deeks, J., Aikman, G., & Carlyon, R. P. (2004). Effects of location, frequency region, and time course of selective attention on auditory scene analysis. *Journal of Experimental Psychology: Human Perception & Performance*, *30*(4), 643–656.

Dannenbring, G. L., & Bregman, A. S. (1978). Streaming vs. fusion of sinusoidal components of complex tones. *Perception & Psychophysics*, *24*, 369–376.

Darwin, C. J., & Hukin, R. W. (2000). Effectiveness of spatial cues, prosody, and talker characteristics in selective attention. *Journal of the Acoustical Society of America*, *107*, 970–977.

Diehl, R. L., Souther, A. F., & Convis, C. L. (1980). Conditions on rate normalization in speech perception. *Perception & Psychophysics*, *27*, 435–443.

Dorman, M. F., Raphael, L. J., & Liberman, A. M. (1979). Some experiments on the sound of silence in phonetic perception. *Journal of the Acoustical Society of America*, *65*(6), 1518–1532.

Gallun, F. J., Mason, C. R., & Kidd, G. (2007). Task-dependent costs in processing two simultaneous auditory stimuli. *Perception & Psychophysics*, *69*(5), 757–771.

Ganong, W. F. (1980). Phonetic categorization in auditory word perception. *Journal of Experimental Psychology: Human Perception & Performance*, *6*, 110–125.

Green, K. P., Stevens, E. B., & Kuhl, P. K. (1994). Talker continuity and the use of rate information during phonetic perception. *Perception & Psychophysics*, *55*, 249–260.

Grimault, N., Bacon, S. D., & Micheyl, C. (2002). Auditory stream segregation on the basis of amplitude-modulation rate. *Journal of the Acoustical Society of America*, *111*(3), 1340–1348.

Grossberg, S. (2003). Resonant neural dynamics of speech perception. *Journal of Phonetics*, *31*(3/4), 423–455.

Hirata, Y., & Lambacher, S. G. (2004). Role of word-external contexts in native speakers' identification of vowel length in Japanese. *Phonetica*, *61*, 177–200.

Hirsh, I. J. (1950). The relation between localization and intelligibility. *Journal of the Acoustical Society of America*, *22*(2), 196–200.

Kidd, G. R. (1989). Articulatory-rate context effects in phoneme identification. *Journal of Experimental Psychology: Human Perception and Performance*, *15*, 736–748.

Kidd, G., Arbogast, T. L., Mason, C. R., & Gallun, F. J. (2005). The advantage of knowing where to listen. *Journal of the Acoustical Society of America*, *118*(6), 3804–3815.

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

Lotto, A. J., Kluender, K. R., & Green, K. P. (1996). Spectral discontinuities and the vowel length effect. *Perception & Psychophysics*, *58*, 1005–1014.

Miller, J. L. (1981). Effects of speaking rate on segmental distinctions. In P. D. Eimas, & J. L. Miller (Eds.), *Perspectives on the study of speech* (pp. 39–74). Hillsdale, NJ: Lawrence Erlbaum.

Miller, J. L. (1987). Rate dependent processing in speech perception. In A. W. Ellis (Ed.), *Progress in the psychology of language*, Vol. 3. London: Lawrence Erlbaum.

Miller, J. L., & Dexter, E. R. (1988). Effects of speaking rate and lexical status on phonetic perception. *Journal of Experimental Psychology: Human Perception and Performance*, *14*, 369–378.

Miller, J. L., Grosjean, F., & Lomanto, C. (1984). Articulation rate and its variability in spontaneous speech: A reanalysis and some implications. *Phonetica*, *41*, 215–225.

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.

Mullennix, J. W., Sawusch, J. R., & Garrison-Shaffer, L. (1992). Automaticity and the detection of speech. *Memory and Cognition*, *20*, 40–50.

Newman, R. S., & Sawusch, J. R. (1996). Perceptual normalization for speaking rate: Effects of temporal distance. *Perception & Psychophysics*, *58*(4), 540–560.

Nusbaum, H. C., & Schwab, E. C. (1986). 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 (pp. 113–158). New York: Academic Press.

Pollack, I., & Pickett, J. M. (1958). Stereophonic listening and speech intelligibility against voice babble. *Journal of the Acoustical Society of America*, *30*(2), 131–133.

Poulton, E. C. (1953). Two-channel listening. *Journal of Experimental Psychology*, *46*(2), 91–96.

Remez, R. E., Rubin, P. E., Berns, S. M., Pardo, J. S., & Lang, J. M. (1994). On the perceptual organization of speech. *Psychological Review*, *101*, 129–156.

Samuel, A. (1986). Red herring detectors and speech perception: In defense of selective adaptation. *Cognitive Psychology*, *18*, 452–499.

Sawusch, J. R., & Newman, R. S. (2000). Perceptual normalization for speaking rate II: Effects of signal discontinuities. *Perception & Psychophysics*, *62*(2), 285–300.

Snyder, J. S., & Alain, C. (2007). Toward a neuropsychological theory of auditory stream segregation. *Psychological Bulletin*, *133*(5), 780–799.

Sommers, M. S. (1997). Stimulus variability and spoken word recognition, II: The effects of age and hearing impairment. *Journal of the Acoustical Society of America*, *101*(4), 2278–2288.

Spieth, W., Curtis, J. F., & Webster, J. C. (1954). Responding to one of two simultaneous messages. *Journal of the Acoustical Society of America*, *26*(3), 391–396.

Summerfield, Q. (1981). Articulatory rate and perceptual constancy in phonetic perception. *Journal of Experimental Psychology: Human Perception and Performance*, *7*, 1074–1095.

Treisman, A. M., & Gelade, G. (1980). A feature integration theory of attention. *Cognitive Psychology*, *12*, 97–136.

Wade, T., & Holt, L. L. (2005). Perceptual effects of preceding nonspeech rate on temporal properties of speech categories. *Perception & Psychophysics*, *67*, 939–950.

Whalen, D. H., & Liberman, A. (1987). Speech perception takes precedence over nonspeech perception. *Science*, *237*(4811), 169–171.

Wood, N. L., & Cowan, N. (1995). The cocktail party phenomenon revisited: Attention and memory in the classic selective listening procedure of Cherry (1953). *Journal of Experimental Psychology: General*, *124*(3), 243–262.