Vowel normalization and the perception of speaker changes: An exploration of the contextual tuning hypothesis

Santiago Barreda[a)]
Department of Linguistics, University of Alberta, Edmonton T6G 2E7, Canada

(Received 17 April 2012; revised 23 July 2012; accepted 30 July 2012)

Many experiments have reported a perceptual advantage for vowels presented in blocked-versus mixed-voice conditions. Nusbaum and colleagues [Nusbaum and Morin (1992). in *Speech Perception, Speech Production, and Linguistic Structure*, edited by Y. Tohkura, Y. Sagisaka, and E. Vatikiotis-Bateson (OHM, Tokyo), pp. 113–134; Magnuson and Nusbaum (2007). J. Exp. Psychol. Hum. Percept. Perform. **33**(2), 391–409] present results which suggest that the size of this advantage may be related to the facility with which listeners can detect speaker changes, so that combinations of less similar voices can result in better performance than combinations of more similar voices. To test this, a series of synthetic voices (differing in their source characteristics and/or formant-spaces) was used in a speeded-monitoring task. Vowels were presented in blocks made up of tokens from one or two synthetic voices. Results indicate that formant-space differences, in the absence of source differences between voices in a block, were unlikely to result in the perception of multiple voices, leading to lower accuracy and relatively faster reaction times. Source differences between voices in a block resulted in the perception of multiple voices, increased reaction times, and a decreased negative effect of formant-space differences between voices on identification accuracy. These results are consistent with a process in which the detection of speaker changes guides the appropriate or inappropriate use of extrinsic information in normalization.
 [http://dx.doi.org/10.1121/1.4747011]

PACS number(s): 43.71.An, 43.71.Bp [JMH] Pages: 3453–3464

I. INTRODUCTION

There is a many-to-many relationship between vowel categories and the acoustic characteristics listeners use to determine vowel quality (Peterson and Barney 1952). Productions of a single vowel category by different speakers might be very different acoustically, just as productions of different vowel categories by different speakers might be very similar acoustically. Differences in production between speakers may arise from differences in speaker gender, size, age, dialect, or any number of other factors. Despite potentially large differences in the acoustic characteristics of a vowel when produced by different people, listeners generally identify vowel tokens with good accuracy. Even for isolated vowels, free from any consonantal context, identification can be quite high (Assmann *et al*., 1982; Macchi, 1980; Rakerd *et al*., 1984). However, it is well known that for a given set of listening conditions, speech presented in a mixed-voice condition tends to be identified less accurately and more slowly than when similar stimuli are presented blocked by voice (Assmann *et al*., 1982; Creelman, 1957; Magnuson and Nusbaum, 2007; Mullennix, Pisoni, and Martin, 1989; Verbrugge *et al*., 1974; Nusbaum and Morin, 1992). The drop-off in identification performance in mixed-voice listening conditions relative to single-voice conditions for the same task will be referred to as the "mixed-voice" effect.

The mixed-voice effect is also associated with additional processing relative to single-voice conditions. Wong *et al*. (2004) report that listeners demonstrate increased activity in areas of the brain involved in speech perception in mixed-voice vs. single-voice listening conditions, indicating that mixed-voice listening conditions bear an added cognitive burden. Nusbaum and Morin (1992) asked participants to remember a series of numbers during a speech identification task and found that this increased reaction times only in mixed-voice conditions, indicating that the process of adapting to differences between speakers interacts with working-memory load.

Similarly, Martin *et al*. (1989) found that serial recall of word-lists is worse when the words are produced by multiple voices, relative to when they are produced by a single voice. Although the exact nature of the mixed-voice effect, and the cause of the additional processing observed in mixed-voice listening conditions, is not exactly known, it seems likely to arise from the mechanism by which listeners account for differences between speakers.

The process by which listeners account for speaker-to-speaker differences in the production of vowels is commonly referred to as "normalization." Many theories of normalization involve the estimation of a speaker-dependent formant-space (Joos, 1948; Ladefoged and Broadbent, 1957; Gerstman, 1968; Ainsworth, 1975; Nearey, 1978; Nearey, 1989; Nearey and Assmann, 2007). The estimate of the speaker's formant-space need only be detailed enough to let the listener know what formant frequencies should be expected for a given vowel category, when produced by the speaker. The listener then determines vowel quality in reference to the estimate of the speaker's formant-space, rather than by considering the acoustic information carried by vowel sounds in

a)Author to whom correspondence should be addressed. Electronic mail: sbarreda@ualberta.ca

absolute terms. Following this tradition, the term normalization will be used to refer to the process by which a listener arrives at an estimate of a speaker-dependent formant-space in order to interpret the vowels produced by a speaker.

If normalization were carried out for each vowel token in turn, without reference to what has been heard previously, one would expect that identification rates for vowels produced by a given speaker would not depend on the number of voices in the round. In addition, reaction times associated with the identification of a given set of speech sounds should not vary based on whether they were presented in a mixed- or single-voice listening condition. Instead, the existence of a mixed-voice effect strongly suggests the importance of extrinsic information in vowel perception, and for the process of normalization.[1] In single-voice blocks, the listener is presented with vowels produced by a single voice so that information from previously heard vowels may be used in order to more accurately identify upcoming vowels. In mixed-voice blocks, the formant-spaces of speakers may differ in such a way that considering vowels produced by one speaker relative to the formant-space of a second speaker may lead to errors. This fundamental difference between mixed-and single-voice listening conditions may help to explain some of the causes of the mixed-voice effect.

A. Contextual tuning theory

Nusbaum and Morin (1992) and Magnuson and Nusbaum (2007) suggest that normalization is controlled by a process they refer to as "contextual tuning." This approach to normalization is summarized in Nusbaum and Morin (1992): "attentional demands increase [in mixed-voice conditions] because the presence of this variability in relationships between speech and linguistic responses requires active processing to reduce the set of possible responses to a single response (Nusbaum and Schwab, 1986). This active processing uses information contained within a single token of speech to provide the context for recognizing the linguistic structure of the utterance, namely a representation of the talker's vocal characteristics. When the listener can develop a mental representation of the talker's vocal characteristics to constrain the representation of subsequent utterances, the demands on attention are reduced." (p. 125).

This formulation of contextual tuning suggests that a listener arrives at a formant-space estimate using information carried by the first speech sound produced by a new voice to interpret subsequent productions by that same voice, and is thus generally compatible with a (conditional) extrinsic-normalization framework. Magnuson and Nusbaum (2007) refine the theory, stating that: "a change in talker triggers normalization procedures that operate until a stable mapping between the talker and internal phonetic categories is achieved. The stable mapping is then maintained until a talker change is indicated acoustically (e.g., by large changes in F0) or more implicitly (e.g., via failures of lexical access)" (p. 393).

They later note that: "The problem of adjusting to changes in talker characteristics then might be thought of as the same kind of computational problem as recognizing phonetic structure (cf., Nusbaum and Magnuson, 1997). In other words, detecting talker differences that require perceptual accommodation is itself a perceptual problem that may not be handled automatically or passively" (p. 402).

1. An elaboration of the contextual tuning approach

Magnuson and Nusbaum (2007) make it clear that their intent is not to investigate the specific mechanisms involved in normalization or the detection of speaker changes. Their goal is only to investigate the cognitive mechanisms by which the normalization process is controlled, stating, "[t]he heavy lifting of identifying specific mechanisms remains" (p. 406). Although a full-fledged identification of such mechanism will not be attempted here, it is useful to explore some modest extrapolations of this general framework that relate in part to somewhat more specific proposals about normalization from the literature that can be subjected to empirical test.

According to contextual tuning, the important factor governing the use of extrinsic information in vowel perception is not whether there has been an actual speaker change, but whether the listener *thinks* that there has been a speaker change. Because of the many-to-many relationship between the acoustic characteristics of a speech sound and speaker changes, it is difficult to delineate the exact conditions under which a listener will detect a speaker change. For example, Magnuson and Nusbaum (2007) report an experiment (Experiment 4) in which listeners performed a speeded monitoring task for blocks made up of synthetic voices which differed only slightly in their f0 (150 Hz vs 160 Hz), but were identical in all other respects. One group of listeners was told that the blocks contained a single voice while the other group was told that the blocks contained multiple voices. The group which was instructed that blocks contained multiple voices exhibited a significant increase in reaction times relative to the group which was told that the blocks contained a single voice. Presumably, listeners who were instructed to expect multiple speakers treated the condition as a mixed-voice one, thereby leading to the longer reaction times typically observed in such tasks. The group which was instructed to expect one voice did not detect speaker changes and did not exhibit the increase in reaction times, despite being presented with identical stimuli.

Contextual tuning is composed of two processes which may result in additional cognitive demands and may help explain the increase in reaction times present in mixed-voice conditions. First, the estimation of the speaker-dependant formant-space may be a cognitively burdensome process, which results in increased reaction times. Although the refinement of the formant-space estimate may be an ongoing process in single-voice conditions, it seems reasonable to think that at some point a listener may become familiar enough with a speaker's voice so that normalization is no longer necessary (i.e., a "stable mapping" between acoustic input and internal representations has been achieved). In a block in which voices (and their related formant-spaces) change from trial to trial in an unpredictable manner, a listener may never arrive at this level of confidence. Another

possibility is that the initial estimation of the formant-space is the most cognitively burdensome, and that refinements to this estimate are less costly. If this were the case an increase in reaction times in mixed-voice conditions would also be observed even if listeners performed formant-space estimations for each vowel since mixed-voice listening conditions would result in relatively more initial estimations than refinements.

Secondly, the detection of speaker changes, or the diversion of some cognitive capacity in order to detect speaker changes, may slow the identification of speech sounds. Although it is reasonable to think that listeners may also monitor for speaker changes in single-voice conditions, this process may not be given a high priority in situations in which listeners do not expect a speaker change. Furthermore, in the event that a speaker change is detected, secondary processes that bear an additional cognitive load may become active. For example, when a likely speaker change is detected, the listener may attempt to estimate the characteristics of the new speaker (e.g., gender, height, age, socioeconomic status, dialect). The listener may also attempt to assess how necessary it is to re-initiate normalization completely, or whether any evidence from previous speech sounds might be used to inform the new re-estimation.

Contextual tuning may also help explain some of the decrease in identification rates for mixed-voice conditions. Because of the uncertainty involved in the detection of speaker changes in a mixed-voice block, listeners may fail to notice a speaker change, just as they might think that there has been a speaker change in cases where there has not. When there are large formant-space differences between speakers, failing to notice a speaker change, and combining extrinsic information from multiple voices, may lead to errors. This suggests that at least some of the decrease in performance associated with the mixed-voice effect is due to the inability of listeners to correctly detect speaker changes in situations where it would be beneficial to do so to maintain high identification accuracy. If this view of normalization is correct, then one would expect that in situations that facilitate the detection of speaker changes, the decrease in accuracy related to formant-space differences between speakers might be minimized.

Although not explicitly stated by Nusbaum and colleagues, contextual tuning seems to imply a rather complex relationship between reaction times, identification accuracy and the detection of speaker changes. In general, phonetically ambiguous stimuli, or more difficult mixed-voice lists, might be expected to result in a decrease in accuracy and an increase in reaction times so that identification accuracy and average reaction times may be negatively correlated across blocks (see Whalen *et al.*, 1993). Independently of this relationship, the detection of speaker changes and the re-initiation of the normalization process may also result in an increase in reaction times. However, since the re-initiation of the normalization process resulting from a detected speaker change should result in a more accurate estimation of the speaker's vowel space, it should result in relatively higher identification accuracy by reducing ambiguity. Consequently, if contextual tuning is correct, one would expect that when the listener detects a speaker change, reaction times will increase without necessarily being associated with lower accuracy.

2. Differential predictions of alternative accounts

This version of contextual tuning may be contrasted with two alternative views of normalization in which the detection of speaker changes does not play an important role. In "pure-intrinsic" normalization theories, the detection of speaker changes is irrelevant because extrinsic information does not play an important role in vowel perception (Syrdal and Gopal, 1986; Smith *et al.*, 2005). According to these views, each vowel token is essentially "self-normalizing" in that it carries all the information necessary for its interpretation. If this were the case, we would expect that identification rates for vowels for a given voice should not vary based on whether they were presented in a mixed-or single-voice condition. With respect to reaction times, although listeners may take more or less time to identify a given vowel produced by a certain voice, there is no clear reason why the reaction times associated with the identification of a set of stimuli should vary systematically based on whether they are presented in a mixed-or single-voice condition. Furthermore, although there may be a positive relationship between average reaction times and identification accuracy in a block, this relationship should not be mediated in any way by the detection of speaker changes.

A second possibility is that extrinsic information is important, but that listeners use information related to the spectral properties of the last n tokens (or the last k seconds of speech) in order to estimate the speaker-dependent formant-space, with no role for the detection of speaker changes. This might be expected if normalization were primarily driven by mechanisms such as those reported in Watkins and Makin (1994) and Watkins and Makin (1996), in which listeners were demonstrated to compensate for the long-term spectral characteristics of a signal when identifying vowel sounds. In a series of experiments, Watkins and Makin presented listeners with a carrier phrase followed by a word containing a vowel token, and asked listeners to identify the word that followed the carrier phrase. Several experiments were carried out, and several different filters were applied to the carrier phrases.

Results indicate that the perceived identity of the vowel following that carrier phrase was predictable based on the long-term average spectral characteristics of the carrier phrase. The authors suggested that some of the perceptual shifts observed in experiments which manipulate carrier phrases to affect the perceived identity of a following target may be caused by accommodation to the long-term average spectral characteristics of the carrier phrase, and not the result of the listener adapting to the formant-space suggested by the carrier phrase. Although there are no clear examples of a normalization method that relies solely on a mechanism like this in the literature, a formant-space normalization system that utilizes statistics such as formant means or ranges over given intervals might have generally similar properties.

A normalization method which worked solely by mechanisms of this kind might be termed "passive-extrinsic,"

since the extrinsic information involved in the process is not variable based on perceived speaker changes or listener expectations, but only on the recent history of stimulus properties (in contrast to this, contextual tuning might be thought of as an "active-extrinsic"[2] model of normalization). If the estimate of the speaker-dependent formant-space involved the joint consideration of information from a fixed number of previous tokens, identification errors would be correlated with the difference between the formant-spaces of the two voices, since the estimated formant-space would be somewhere between these two. Reaction times might be expected to vary based on the phonetic ambiguity of the vowels being presented, but again, there should not be systematic variation in the relationship between reaction times and identification accuracy resulting from whether the listener thought the round contained one, or more than one speaker.

3. Testing contextual tuning theory in Magnuson and Nusbaum (2007)

Magnuson and Nusbaum (2007) present the results of an experiment (Experiment 1) meant to offer explicit support for contextual tuning theory.[3] The stimuli consisted of isolated vowels produced by four natural voices; those of two adult males and two adult females. The average F1 and F2 values for the vowels of the two female speakers differed by only 0.3%, while the average F1 and F2 values for the two male speakers differed by 5.4%. Although within-gender differences were somewhat larger for the males than for the females, both were small compared to the 20% differences between male and female speakers.

Vowels were presented in blocks of 16 vowels produced by either a single voice, or two different voices. Each listener heard vowels presented in both single- and mixed-voice conditions, where one group of listeners was always presented with mixed-voice blocks in which speakers were of the same gender, and another group was presented with mixed-voice blocks in which speakers were of different genders. Within each block, the target vowel was one of /i ɪ u ʊ/, while distractors were chosen from the vowels /e æ ʌ ɛ/, plus any of the four target vowels that were not acting as targets for that particular block. Each block contained a total of four targets inserted randomly into the sequence, with the constraint that no two targets appear in a row. Listeners performed a speeded-monitoring task where they had to push a computer key as soon as they heard the target vowel (indicated to them on a monitor), and ignore all non-target distractor vowels. Response times were measured from stimulus onset, and hit rates (responses registered following targets) and false alarms (responses registered following distractors) were collected.

Magnuson and Nusbaum report a significant decrease in hit rates for mixed-voice blocks relative to single-voice conditions. Hit rates were slightly higher for different-gender blocks relative to same-gender blocks overall, but the main effect for gender homogeneity did not reach significance. There was a nearly significant ($p = 0.072$) interaction between talker condition (mixed-speaker vs single-speaker) and gender homogeneity. Reaction times were significantly higher in all mixed-voice blocks relative to the single-voice blocks, save for the female-female mixed-voice blocks which did not differ significantly from single-voice blocks.

According to contextual tuning, performance may be higher in different-gender mixed-voice blocks than in the same-gender mixed-voice block because listeners are aware that these blocks contain multiple speakers. This realization may partly counteract the negative effect of the much larger formant-space differences between speakers of different genders compared to speakers of the same gender.

On the other hand, although there were relatively smaller differences between the formant-spaces of different speakers of the same gender, listeners may not have realized that the blocks involved multiple speakers; or, even if they did, they may have missed exactly when speaker changes were occurring. As a result, the same-gender mixed-speaker blocks manifested a trend toward slightly lower performance than the different-gender mixed-voice blocks. This is true despite the fact that formant-space differences between voices are smaller in same-gender cases. Finally, although the female-female mixed-voice blocks objectively consisted of vowels from two different voices, reaction times did not differ significantly from those of single-voice blocks, suggesting that listeners may not have realized that the blocks contained more than one speaker. This highlights the fact the detection of speaker changes is an imperfect, non-deterministic process.

Although the trends in the pattern of results are generally consistent with contextual tuning theory, many effects tested in Magnuson and Nusbaum (2007) are generally weak or non-significant and thus do not offer strong support for contextual tuning. However, some aspects of the experimental design may have contributed to the limited size of the effects. First of all, the target vowels used (/i ɪ u ʊ/) may not be very confusable with each other in mixed-voice conditions. These four vowels were identified correctly in 97% of cases in data presented by Peterson and Barney (1952) and in 98% of cases in Hillenbrand *et al.* (1995). Furthermore, the vowels which are most spectrally similar /u ʊ/ and /i ɪ/ may be distinguishable on the basis of durational differences or because of vowel inherent spectral change when produced by natural voices (Hillenbrand *et al.*, 1995; Nearey and Assmann, 1986). Perhaps as a result of this, hit rates hovered around 93% in all listening conditions. This leaves very little room to model variation in performance as a result of different voice pairs. Furthermore, because natural voices were used, it is difficult to know which aspect of the speakers' voices listeners were using to detect speaker changes, or under what conditions they were likely to detect speaker changes.

B. Rationale for current experiment

The experiment to be described below adopts the same basic design used in Experiment 1 of Magnuson and Nusbaum (2007) with some modifications which may enhance and clarify the effects reported for that experiment. A series of synthetic voices was created which differed in their formant-spaces and/or their source characteristics, and the four vowels /æ ʌ ʊ ɑ/ were synthesized for each voice. As

opposed to the vowels used in Magnuson and Nusbaum (2007), these vowels are generally more difficult to identify: in data presented by Peterson and Barney (1952) they were identified correctly in 93% of cases, while they were identified correctly in 93.7% of cases in Hillenbrand *et al*. (1995). This was expected to result in lower performance overall. Synthetic voices were used in order to control for random variation in the production of vowels and to eliminate idiosyncratic differences in source characteristics between voices. Furthermore, each block contained a higher number of total vowel tokens (30) and target tokens (12), in order to allow for more variation in hit rates.

Differences in source characteristics between voices in a block were expressly intended to facilitate the detection of speaker changes in a block, thereby potentially mitigating the decrease in hit rates associated with mixed-voice listening conditions by strongly encouraging the detection of speaker changes when the voices had different sources. The formant-space differences between speakers were intended to result in decreased performance (i.e., the mixed-voice effect) when listeners were unlikely to detect speaker changes in a block (e.g., in the absence of source differences between voices). If a version of contextual tuning theory adequately describes the process of normalization, three general results are expected.

(A) The decrease in identification rates associated with formant-space differences in mixed-voice conditions will be mitigated in situations in which the detection of speaker changes is facilitated.
(B) In situations where speaker changes are not detected, performance should improve in blocks where voices have similar formant-spaces. When listeners are likely to detect speaker changes (e.g., in blocks with heterogeneous sources), their ability to refine their speaker-dependent formant-space estimate may be limited. This may result in a lack of improvement throughout a block or in lower performance overall.
(C) Although average reaction times may co-vary negatively with hit rates for blocks, blocks in which speaker changes are likely to be detected may exhibit an increase in average reaction times without a concomitant decreases in hit rates.

II. METHODOLOGY

A. Participants

Participants were 71 native speakers of Canadian English, drawn from the linguistics participant pool at the University of Alberta. Participants received partial course credit for taking part in the experiment. Participants were randomly assigned to a target vowel group and each participant only monitored for a single vowel. There were 18 participants in each of the target vowel groups, except for the /æ/ group which had only 17 participants.

B. Stimuli

The vowels used in the experiment were /æ ʌ ʊ ɑ/, where one of the four acted as the target and the others acted as distractors. A series of 6 synthetic voices were created which differed in terms of their vowels spaces and/or f0 and source characteristics. Formant-space differences were manipulated by using three formant frequency (FF) scaling levels: a baseline level with FFs appropriate for an adult male, a second level with a 10% increase to all FFs (F1–F10) and a third level with a 20% increase to all FFs (F1–F10) relative to baseline. The baseline FF values used are presented in Table I, and these were based on production values collected from native-speakers of Edmonton English. Baseline F4 values were set at 3500 for all vowels with subsequent FFs set to 1050 Hz greater than the previous FF. Formants above F3 were scaled by the same factor as F1 to F3 for the other conditions.

The two voice source levels consisted of an f0 of 120 Hz with modal source characteristics and an f0 of 240 Hz with breathy source characteristics. The breathy source characteristics were simulated by setting the source bandwidth to 75 Hz and using 10 dB of negative spectral tilt at 3000 Hz (Klatt and Klatt, 1990). Since f0 level and source characteristics were perfectly correlated, the different f0 and source levels will simply be referred to as voice source characteristics. All vowels had steady-state formants, were 200 ms in duration and were synthesised at a sampling rate of 22 050 Hz.

C. Procedure

The general design of the task is an extension of experiments outlined in Nusbaum and Morin (1992) and Magnuson and Nusbaum (2007). Listeners were asked to perform a speeded monitoring task where they had to respond only when they heard a specific target vowel and ignore all distractor vowels. Each listener monitored for a single target vowel so that the designation of a vowel as either target or distractor is listener-specific. All listeners were told which vowel they would be targeting and which vowels would serve as distractors.

Listeners were presented with all combinations of voice pairs, presented in blocks. There were 21 unique voice pair combinations and listeners heard each combination twice resulting in 42 blocks per participant. Listeners were told that any given block might contain vowels from a single voice or from more than one voice. Thirty vowels were presented within each block, consisting of six targets and nine distractors from each voice (three instances of each non-target vowel). Vowels were randomized within a block subject to the constraint that no two targets appear in a row. The onset of each vowel within a block occurred one second after the onset of the previous vowel, meaning that each block of vowels was roughly 30 s in duration. When a block was completed, there was a self-timed pause, which ended when the participant pressed a button. Reaction times (measured from stimulus onset) and accuracy for responses to targets (hits) and distractors (false alarms) were recorded within a block. The hit rate for a block was calculated by dividing the number of correct identification of targets by the total number of targets in the block. False alarm rates were calculated by dividing the number of responses to non-target distractor vowels by the total number of non-target distractors in the

TABLE I. Formant frequencies for the vowels of the baseline voice.

	Baseline FF values (in Hz)			
Vowel	æ	ʌ	ʊ	ɑ
F1	717	665	483	651
F2	1497	1283	1093	1055
F3	2319	2318	2272	2323

block. The experiment was carried out using DMDX (Forster and Forster, 2003), and responses were collected using a USB gamepad.

Although the relatively large source differences between voices were intended to strongly suggest to listeners that there were multiple voices in a block, while conducting the experiment it was realized that it would be beneficial to ask participants how many voices they thought they heard in a given round. The last 14 participants performed an additional task where at the end of each block they were asked to report whether they thought the block contained one or more than one voice and whether they were confident or uncertain of the number of voices in the block. Participants were told that they would be asked to perform this task at the end of each block prior to the beginning of the experiment. The results from this secondary task strongly met expectations regarding the expected relationship between source differences between voices and the detection of speaker changes. The results of this secondary task, in addition to a summary of tests of heterogeneity of results between participants who completed the secondary task and those who did not, are presented in the Appendix.

III. RESULTS

Since the task was designed to be difficult, participants were screened to ensure that they were completing the task to a minimally satisfactory level. This was done by removing any participant who had more false alarms than correct identifications of targets. This resulted in the removal of six of 71 participants, 5 from the /ʌ/ target group, and one from the /ɑ/ target group. All further discussion will be based on the results of the remaining 65 participants.

Each participant heard a total of 1260 vowels across all 42 blocks for a total of 81 900 trials across all participants. Since the software used only registered one response per stimulus, very fast responses were ambiguous. For example, in some cases responses were registered only 10 ms after stimulus onset, making it more likely that it was a very late response to the previous stimulus than a very fast response to the current one. As a result of this, when a reaction time under 200 ms (the duration of the vowel stimuli) was registered, both the stimulus that was responded to and the stimulus that immediately preceded it were discarded. Participants responded in less than 200 ms in 533 cases, resulting in 1065 discarded responses (1.3% of total responses) and 80 835 useable trials. An average of 16.4 responses were lost from each participant (SD = 14.2) with the most lost from any participant being 64 trials, 5% of total trials for that participant.

The predictions made by the contextual tuning hypothesis (outlined at the end of Sec. I B) relate directly to the formant-space and source differences between voices in a block. To test these predictions more directly, all blocks were classified into one of six voice-pair types based on the acoustic differences between the voices in the block –i.e., formant-space differences of 0, 10, or 20 % between voices, and either homogeneous or heterogeneous voice sources for each formant-space difference. Hit rates, false alarm rates and average reaction times (for correct identifications) were calculated for each block, independently for each listener. The average of each of these values was then found for each voice-pair type for each participant, resulting in 18 measurements per listener: an average hit rate, an average false alarm rate, and an average reaction time for each of the six voice-pair types. Unless otherwise specified, the remaining discussion will involve average performance, within-participant, between voice-pair types. Each of the predictions to be tested will be dealt with in turn in the following three subsections (III A through III C).

A. Vowel identification performance

A series of repeated-measures analyses of variance was conducted on hit rates, false alarm rates and average reaction times for the two factors used to differentiate voice-pair types: formant-space difference between voices (0, 10, 20%) and voice source homogeneity. The average within-participant hit rate, averaged across all voice-pair types, was 76% (sd = 15%) with a minimum of 34% and a maximum of 95% across participants. The distribution of hit rates, organized by voice-pair type, is presented in Fig. 1. The main effects for voice source homogeneity [$F(1,64) = 4.74$, $p = 0.0331$], and formant-space difference [$F(2,128) = 70.83$, $p < 0.0001$] were both significant, as was the interaction of the two [$F(2,128) = 31.83$, $p < 0.0001$].

The nature of the interaction effect was explored by simple main-effects analysis of hit-rates. When voices in a block had homogenous source characteristics, there was a very strong effect for formant-space differences in hit rates [$F(2,128) = 87.22$, $p < 0.0001$]. As seen in Fig. 1, the interaction pattern suggests that formant-space differences between

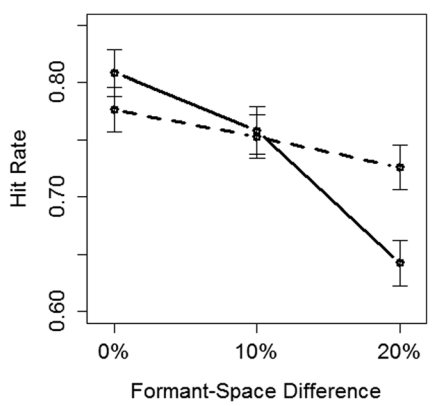

FIG. 1. Average within-participant hit rate, presented by voice-pair type. Blocks with homogeneous source characteristics are indicate by a solid line, blocks with heterogeneous source characteristics are indicated by the broken line. Error bars indicate standard errors for each mean.

voices appear to affect hit-rates less for heterogeneous-source blocks. Despite this reduction, the simple main effect of formant-space differences for heterogeneous source blocks is still significant [$F(2,128) = 10.59$, $p < 0.0001$].

Consider now the simple main effects of source-heterogeneity within levels of formant-space difference. Voice source heterogeneity between voices in a block is associated with a 3.2% decrease in hit rates for the 0% formant-space difference [$t(64) = 3.33$, $p = 0.0014$], however, when the formant-spaces of voices differ by 10%, source differences between voices have no significant effect on hit rates [$t(64) = 0.54$, $p = 0.56$]. When the formant-spaces of voices differ by 20%, hit rates are 8.3% higher in cases where source characteristics are heterogeneous [$t(64) = 5.9$, $p < 0.0001$]. Note that in this case, the effects of source heterogeneity are in the opposite direction from those in the 0% formant-space case, resulting in the crossing lines in Fig. 1.

Turning now to false alarms, the average within-participant rate was 8.2% (sd = 7.3%) with a minimum of 0.1% and a maximum of 27.3% across participants. A significant main effect for both voice source [$F(1,64) = 16.85$, $p = 0.0001$] and formant-space differences was found [$F(2,128) = 5.76$, $p = 0.004$]. Unlike the analysis of hit rates, however, the interaction between the two did not reach significance [$F(2,128) = 1.94$, $p = 0.1473$]. On average, source differences between voices in a block resulted in 1.6% more false alarms [$t(64) = 4.1$, $p = 0.0001$]. Formant-space differences of 10% did not significantly increase the number of false alarms relative to blocks in which voices had the same formant-space [$t(64) = 1.1$, $p = 0.26$]; but, when formant-spaces differed by 20% false alarms increased by 1.2% [$t(64) = 3.14$, $p = 0.0025$].

A pattern similar to the false-alarms results was found for reaction times. There was a significant main effect for voice source homogeneity [$F(1,64) = 75.43$, $p < 0.0001$] and formant-space difference [$F(2,128) = 10.59$, $p < 0.0001$], but there was not even a hint of a significant interaction between the two [$F(2,128) = 1.4$, $p = 0.2496$]. The average, within-participant reaction time for the voice-pair type in which voices had the same formant-space and source characteristics was 516 ms (sd = 62 ms), with voice source heterogeneity resulting in an average delay of 27 ms [$t(64) = 8.7$, $p < 0.0001$]. Compared against the control 0% formant difference case, formant-space differences of 10% resulted in an added delay of 10.9 ms [$t(64) = 4.1$, $p = 0.0001$] relative to blocks with no formant-space differences, while formant-space differences of 20% resulted in an added delay of 12.4 ms [$t(64) = 4.3$, $p < 0.0001$] relative to blocks with no formant-space differences. There was no significant difference in response times between blocks with 10% and blocks with 20% formant-space differences [$t(64) = 0.46$, $p = 0.64$].

B. Improvement within a block

According to contextual tuning (at least as elaborated in Sec. I A 1), in blocks where listeners do not detect speaker changes, they are expected to refine their estimate of the speaker-dependent formant-space throughout the block. When voices in a block share a formant-space, this should lead to an improvement in performance from the beginning to the end of the block, as every consecutive token provides the listener with information which may be used to accurately refine their estimate. On the other hand, in cases where the listener is likely to detect speaker changes, they are expected to re-initialize the normalization process and avoid the use of inappropriate extrinsic information in normalization. This is expected to mitigate some of the mixed-voice effect, by minimizing the inappropriate use of extrinsic information. However, it may also mean that listeners are not able to refine their estimate of the speaker-dependent formant-space as the block progresses to the extent that they would in the absence of detected speaker changes.

An analysis was devised to summarize the nature of change of identification accuracy during the course of a block and to relate patterns of such change to voice-pair type. Each block contained a total of 30 vowels, 12 of which were targets. Although the targets within a block were presented in a random order (with the constraint that no two targets appear in succession), targets can be considered in terms of the order in which they appeared in a block. On average, in cases where the performance of listeners improves in a block, hit rates for target n_i is expected to be lower than performance for target n_{i+1}. In cases where performance decreases throughout a block, performance for target n_i is expected to be higher than performance for target n_{i+1}. When the performance of a listener is stable within a block, there should be no relationship between target position within a block and expected performance for that target. As a result, the slope coefficient relating hit rates to within-block target number should give an indication of how performance varies within a block, with a positive coefficient indicating improvement, a negative coefficient indicating worsening performance and a coefficient of zero indicating stability.

To investigate how performance within a block varies by voice-pair type, all blocks were sorted by voice-pair type, according the acoustic differences between the voices in the block. Targets were assigned a number from 0 to 11, based on the relative position in which they appeared within the block. This target number was then divided by eleven so that target numbers corresponded to equal fractional increments from 0 to 1. In this way, estimated coefficients have a straightforward interpretation as the expected increase in hit rates (measured in percentage points) from the first target in the block to the last target in the block. Within-participant hit rates were calculated for each target position within each voice-pair type. A regression was then carried out independently for each voice-pair type and individually for each participant, predicting hit rates by relative target position. This resulted in six estimated coefficients for each participant (one for each voice-pair type). The distribution of these coefficients, organized by voice-pair type, is presented in Fig. 2.

A repeated-measures analysis of variance was carried out on these estimated coefficients, with voice source homogeneity and formant-space differences (0, 10, 20%) between voices in a block acting as within-subjects factors. A significant main effect was found for formant-space differences between voices [$F(2,128) = 6.67$, $p = 0.0017$]. The main

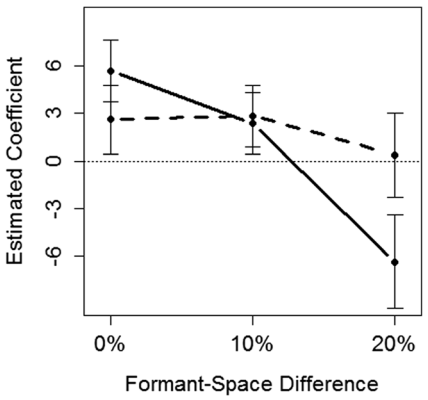

FIG. 2. Average coefficient relating within-block target number, and expected hit rates for that target within a block. Blocks with homogeneous source characteristics are indicate by a solid line, blocks with heterogeneous source characteristics are indicated by the broken line. Error bars indicate standard errors for each mean.

effect for voice source [$F(1,64) = 0.81$, $p = 0.3718$] was not significant. Although the interaction between formant-space differences and voice source [$F(2,128) = 2.93$, $p = 0.0573$] fell just short of the conventional .05 significance level, it seemed reasonable to investigate it further. Accordingly, simple main effects tests were performed.

Consider the simple main effect of formant space within source condition. When source characteristics in a block were homogenous, there was a strong effect for formant-space differences on improvement in a block [$F(2,128) = 8.41$, $p = 0.0004$]. However, when voices in a block had heterogeneous source characteristics, there was no significant effect for formant-space differences on improvement [$F(2,128) = 0.49$, $p = 0.6148$]. In these cases, coefficients did not differ significantly from zero in any case, regardless of formant-space differences between voices, although in all three cases they were slightly positive.

Consider now the case of homogeneous source characteristics. In cases where voices had the same source characteristics and formant-spaces, listeners showed a significant improvement as blocks progressed [$m = 5.7$, $t(64) = 2.96$, $p = 0.0043$], while in cases where voices in a block had the same source but formant-spaces differed by 20%, listeners performed significantly worse as blocks progressed [$m = -6.3$, $t(64) = -2.16$, $p = 0.0345$]. When voices had homogenous source characteristics and a 10% formant-space difference, there was no significant change in hit rates as the block progressed [$m = 2.4$, $t(64) = 1.23$, $p = 0.22$].

C. The relationship between reaction times, hit rates and the detection of speaker changes

As mentioned in the Introduction, phonetically-ambiguous stimuli may take longer to identify in general than less ambiguous stimuli. Since ambiguous vowels should be less accurately perceived, this should by itself result in a negative relationship between the average reaction times in a block and the hit rate for that block. There is in fact a negative relationship between the hit rates and average reaction time in a block. Correlation coefficients between these two measures were calculated for each participant. A between-participants t-test conducted on these correlation coefficients reveals a highly significant negative correlation, averaging -0.18 [$t(64) = -8.5$, $p < 0.0001$].

However, contextual tuning posits that when a speaker change is detected, processes related to the more accurate identification of vowels (e.g., the re-initiation of normalization) are also expected to result in an increase in reaction times. As a result, in situations where listeners are likely to detect speaker changes in a block, reactions times should be higher overall without necessarily being associated with a decrease in hit rates.

To explore how the relationship between acoustic differences and reaction times may be mediated by the detection of speaker changes, the following procedure was carried out individually for each participant. The average reaction time for each block was regressed on the hit rate for that block, resulting in a reaction-time residual for each block. This residual represents variation in average reaction times that cannot be accounted for by the hit rate for that block. A positive residual indicates that a listener responded slower than expected given their average accuracy, while a negative residual indicates that listeners tended to respond faster than expected given their average accuracy. The mean reaction time residual for each voice-pair type was found, resulting in six measurements for each of the 65 participants. The distribution of average within-participant residuals, grouped by voice-pair type, are presented in Fig. 3.

Since heterogeneous source characteristics between voices in a block are strongly associated with the detection of speaker changes, it is expected that average reaction times for blocks in which voices have heterogeneous source characteristics should be longer than expected given the hit rate for the block. This suggests that if contextual tuning is correct, the average residual resulting from the analysis presented above should be positive when there are source differences in a block, indicating delays not explicable by ambiguity as indexed by decreased hit rates. The results presented in Fig. 3 confirm this expectation.

To test for the significance of this effect, a two-way, repeated measures analysis of variance was carried out on

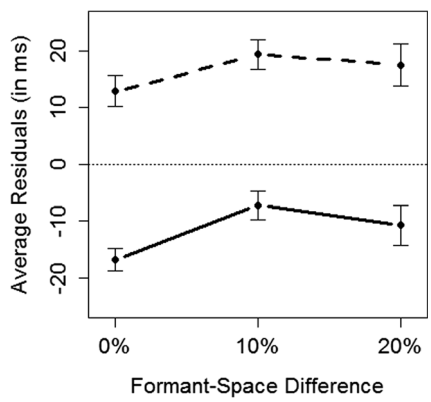

FIG. 3. Average, within-participant residual resulting from regressing reaction time on hit rates, presented by voice-pair type. Blocks with homogeneous source characteristics are indicate by a solid line, blocks with heterogeneous source characteristics are indicated by the broken line. Error bars indicate standard errors for each mean.

the average reaction-time residual, with voice source homogeneity and formant-space difference between voices (0, 10, 20%) acting as within-participant factors. A significant main effect was found for voice source [$F(1,64) = 88.02$, $p < 0.0001$], with the average absolute difference in residuals resulting from voice source heterogeneity being 28 ms. The main effect for formant-space differences [$F(2,128) = 3.5$, $p = 0.0331$] was also significant, however, the interaction between voice source and formant-space difference was not significant [$F(2,128) = 0.15$, $p = 0.8588$]. Although formant-space differences affect the reaction time residuals (likely reflecting the fact that these alone were sometimes sufficient to trigger the detection of speaker changes), on average, listeners respond faster than expected given their hit rates when there is voice source homogeneity in a block.

IV. DISCUSSION

In the introduction, contextual tuning theory was outlined and contrasted with two alternate views of normalization. Rather than focus on the specific processes involved in normalization, these theories were framed in terms of how normalization is controlled, and specifically, how extrinsic information is used in normalization. The two types of theories considered in alternative to contextual tuning theory were pure-intrinsic theories, in which extrinsic information plays no important role in normalization, and passive-extrinsic theories, in which extrinsic information is used by the normalization process in a rigid way. Although they differ in terms of the role played by extrinsic information, both of these alternatives are cognitively passive, in that they do not necessarily require active cognitive control to be carried out (Magnuson and Nusbaum, 2007). Furthermore, neither of these alternatives involves the detection of speaker changes in any way. Thus, they cannot predict any relationship between hit rates, reaction times and the detection of speaker changes.

In contrast, contextual tuning theory posits that the detection of speaker changes plays a critical role in guiding listeners' use of extrinsic information in normalization. In a sense, contextual tuning might be thought of as consisting of two "modes," one being more similar to pure-intrinsic normalization and the other being more similar to passive-extrinsic normalization.

In the absence of a detected speaker change, the listener is in a passive-extrinsic normalization mode and extrinsic information from previous tokens is accumulated and used to identify subsequent vowel tokens. If the formant-spaces of the voices in the block are the same or similar, this refinement will facilitate identification. If the voices in a block have substantially different formant-spaces, the joint consideration of information from different voices may negatively affect hit rates. On the other hand, when a speaker change is detected, the listener shifts to a strategy similar to pure-intrinsic normalization. Previous extrinsic evidence may be discarded as inappropriate and the hit rates associated with a particular vowel token may be closer to those that would be predicted based on the intrinsic properties of the vowel sound.

The experiment described above relied on source differences between voices in a block to give listeners the impression that a block contained multiple voices. The results presented in the Appendix confirm this expectation; when voices in a block had homogenous source characteristics listeners were very likely to hear a single voice in a block. As a result, when voices in a block had homogenous source characteristics, listeners may have been in a passive-extrinsic normalization mode. This resulted in good performance when voices in a block had the same formant-space, and poor performance when voices in a block had very different formant-spaces (these two situations are presented in the extreme points on the solid line in Fig. 1). In addition, when voices in a block had homogenous source characteristics and the same formant-space, hit rates improves significantly within a block. This suggests that listeners were, in fact, refining their formant-space estimates on the basis of additional extrinsic information in order to arrive at more accurate estimates. On the other hand, when the formant spaces of voices differed by 20%, hit rates declined significantly within blocks, suggesting that identification accuracy suffered from the incorrect combination of extrinsic information from multiple voices.

The variation in hit rates within a block may be explained by the amount of extrinsic information available to a listener for each consecutive vowel target in a block. For example, the average ordinal position of the first target in a block was 1.6 (out of 30), while the average ordinal position of the final target in a block was 29.1. Clearly, in blocks where voices have different formant-spaces, the chances that a target has been preceded by inappropriate extrinsic information is fairly low for the first target in a block, while it is a certainty for the final target in the block. As a result, in situations in which listeners were unlikely to detect speaker changes, the incorrect use of extrinsic information may increase or become more likely as a block progresses, and identification accuracy may suffer. Conversely, in situations in which voices had the same formant-spaces, listeners would have been provided with increasing amounts of appropriate extrinsic information as a block progressed and the lack of detected worked in their favor.

In blocks in which voices had heterogeneous source characteristics, listeners overwhelmingly reported hearing multiple voices in a block. This greatly diminished the negative effect of formant space differences between voices in a block, as demonstrated by the relative lack of change in hit rates when voices in a block had heterogeneous source characteristics (represented by the broken line in Fig. 1). As opposed to blocks where voices had homogenous source characteristics, hit rates were relatively stable, with no significant increase or decrease in hit rates within a block regardless of the formant-space differences between voices. These results support the notion that, in the presence of detected speaker changes, listeners were likely to be operating in something more similar to a pure-intrinsic normalization mode in which extrinsic information plays a diminished role.

Contextual tuning also suggests a complicated relationship between reaction times, hit rates and the detection of speaker changes. The results presented in Sec. III C indicate

that although reaction times are negatively correlated with hit rates, blocks in which voices had heterogeneous source characteristics tended to feature slower reaction times without being associated with decreased hit rates. When considered together with the fact that source heterogeneity resulted in the detection of multiple speakers, a decreased negative effect of formant-space differences between voices, and stability in identification rates within blocks, this is considered to be strong support of the claim that the detection of speaker changes results in additional processing associated with the more accurate perception of speech, and the avoidance of the incorrect use of previously heard extrinsic information.

Magnuson and Nusbaum (2007) report an increase in reaction times of 29 ms in mixed-voice blocks relative to single-voice blocks for a task very similar to the one reported here (Experiment 1). This is very close to the 27 ms average increase in reaction times resulting from source differences between voices in a block, presented in Sec. III A. This suggests that source differences between synthetic voices used here resulted in remarkably similar processing costs to those incurred when listeners are presented with mixed-voice lists consisting of vowels produced by different human speakers in a similar task. Furthermore, this increase in average response times is very close in magnitude to the 28 ms difference in average residuals after controlling for hit-rate resulting from voice source heterogeneity between voices in a block, reported in Sec. III C. Since these residuals represent variation in reaction times that cannot be accounted for by the phonetic ambiguity of tokens in a block, this suggests that increases in reaction times resulting from source differences between voices in a block may primarily result from additional processing associated with the detection of speaker changes.

V. CONCLUSION

Taken together, the results outlined in the previous section offer strong evidence for a version of contextual tuning theory as the mechanism that controls the normalization process. Source differences between voices in a block resulted in the impression that there were multiple voices in a block. These differences also resulted in increased reaction times that cannot be fully explained by increased phonetic ambiguity (as indexed by lower hit rates). This is consistent with the hypothesis that the additional processing in blocks with heterogeneous voice sources is actually related to the more accurate perception of many of the vowels. For homogeneous source blocks, the absence of the additional processing associated with a detected speaker change resulted in good accuracy (with improvement within a block) when voices had similar formant-spaces, and poor accuracy (with decline within a block) when voices had dissimilar vowel spaces. In heterogeneous source blocks, when the listener was more likely to be aware of speaker changes in a block, performance was relatively stable within a block and the negative effect of formant-space incongruences between voices was greatly reduced.

To sum up, the complex pattern of results for hit-rates and reaction time differences outlined above cannot be explained either: (a) by a pure-intrinsic normalization process where extrinsic information plays no role whatsoever or (b) by an extrinsic normalization theory in which information is used in a rigid, automatic, fully stimulus-driven manner. By contrast, all the results are reasonably explained by the contextual tuning hypothesis as elaborated in Sec. I A 1 and in the discussion. This is a version of contextual tuning that includes a switch between two processing modes guided by the presence or absence of the detection of a change in speaker. The first mode is operative when a new trial is detected as originating from a speaker that is different from that of an immediately preceding trial. It is viewed here as a form intrinsic normalization, where the current speaker's formant-space is estimated only from information in the current utterance and where that fresh estimate is used in the identification process. The second mode applies when a new trial is perceived as having been produced by the same speaker as an immediately preceding trial. It is viewed as a form of extrinsic normalization, in which a listener's estimate of the formant-space is refined from the estimate used in the previous trial and applied to the identification of the current stimulus. Although a full account of the details of this process will require substantial additional research, the broad outlines seem rather clear.

ACKNOWLEDGMENTS

I wish to thank Terry Nearey for many valuable comments during the production of this manuscript and Pat Bolger for advice regarding the software and hardware used to carry out this experiment.

APPENDIX: EXPLICIT DETECTION OF VOICE CHANGES

The source differences between voices in a block were intended to result in the detection of speaker changes. To confirm this, the final 14 participants performed an additional task at the end of each block. Although these participants were not randomly interspersed among all participants, they still represent a random sample of participants in that they were not selected because of any particular quality they possessed. It is important to note that this additional task was not meant to establish a firm connection between acoustic differences between voices in a block and the detection of speaker changes, but only to confirm that, within the context of this experiment, source heterogeneity would strongly signal a likely speaker change. Participants were instructed that at the end of each block they would have to answer two additional questions:

(1) How many voices did you hear in the block?
(2) How confident are you in that assessment?

At the end of each block, participants were asked to select from two options to answer question one: "one voice" or "more than one voice." After they answered this question, they were asked to select from the following options to answer question two: "confident" or "unsure." These options were presented in successive screens so that answering the

first question brought up the second question. After answering the second question, participants had a self-timed pause after which they continued on to the next block. Answers to these two questions were analyzed separately as described below.

Since the participants who performed this additional task had their attention explicitly drawn to the number of voices in a block, their performance may have varied in some way from that of the 51 participants who did not perform the secondary task. To test for this, participants were divided according to whether or not they performed the secondary task, and their hit rates, false alarm rates, and reaction times were sorted according to voice-pair type (as for the analyses presented in Sec. III A). A series of independent-sample *t*-tests was then carried out on hit rates, false alarm rate, and reaction times for each voice-pair type, where performance of the secondary task served as the grouping factor. The results of the 18 individual *t*-tests revealed no significant differences between any of the measurements for any of the voice-pair types, even at an uncorrected *p*-value of 0.05. The lack of a difference in performance between the two groups may be a result of the fact that, although this secondary task drew explicit attention to the number of voices in a block, it was stated clearly in the instructions given to all participants before commencing the experiment that each block could potentially contain more than one voice, and that this would change from block to block in an unpredictable manner.

1. Number of voices per block

The results for this question are presented in Table II. A two-way, repeated-measures analysis of variance was carried out on the rate at which speakers thought a block contained more than one voice. Because of the extreme values for some conditions, an arcsine transform was carried out on the dependent variable. There were two within-subject factors: the formant-space difference between the two voices (0, 10, 20%), and voice source homogeneity. A significant main effect was found for both formant-space difference [$F(2,26) = 6.51$, $p = 0.0051$] and voice source homogeneity [$F(1,13) = 258.53$, $p < 0.0001$], as well as a significant interaction between the two [$F(2,26) = 9.86$, $p = 0.0006$]. When voice sources were heterogeneous, listeners indicated hearing more than one voice in a block in 96.5% of cases, and there is no significant effect for formant-space difference [$F(2,26) = 1.58$, $p = 0.2246$]. When voice sources were homogenous, listeners reported hearing more than one voice in 16.8% of cases and the effect of formant-space difference is significant [$F(2,26) = 11.54$, $p = 0.0040$].

TABLE II. Percent of rounds in which listeners reported hearing more than one voice in a block, presented by voice-pair type. Numbers in parentheses are the standard errors of each mean.

	More than 1 voice in block		
	Formant-space difference		
Voice source	0%	10%	20%
Homogeneous	4.8 (2.1)	9.8 (4.9)	35.7 (9)
Heterogeneous	97.6 (1.6)	99.1 (0.9)	92.9 (3.1)

TABLE III. Percent of rounds in which listeners reported being unsure of the number of voices in a block, presented by voice-pair type. Numbers in parentheses are the standard errors of each mean.

	Unsure of number of voices in block		
	Formant-space difference		
Voice source	0%	10%	20%
Homogeneous	14.3 (4.4)	25.9 (5)	39.3 (5.7)
Heterogeneous	4.8 (2)	4.5 (2.5)	3.6 (2.4)

2. Confidence in number of voices per block

A similar analysis of variance was applied to the rate at which listeners were sure of the number of voices in a block, revealing the same pattern of results, presented in Table III. A significant main effect was found for both formant-space difference [$F(2,26) = 3.72$, $p = 0.038$] and voice source homogeneity [$F(1,13) = 35.78$, $p < 0.0001$], as well as a significant interaction between the two [$F(2,26) = 5.17$, $p = 0.0129$]. When voice sources were heterogeneous, listeners indicated being unsure of the number of voices in the block in only 4.3% of cases and there is no significant effect for formant-space difference [$F(2,26) = 1.58$, $p = 0.2246$]. When voice sources were homogenous, listeners indicated being unsure of the number of voices in the block in 26.5% of cases and the effect of formant-space difference is significant [$F(2,26) = 11.54$, $p = 0.0040$].

3. Summary of results

When voices in a block had heterogeneous source characteristics, listeners were very likely to hear multiple voices and were confident of this assessment, regardless of the difference in the formant-spaces of the voices. When voices in a block had homogenous source characteristics, listeners were most likely to think that there is a single voice in the block. Even in cases where the formant-spaces of voices differed by 20%, listeners only reported hearing more than one voice in 35.7% of cases. Voice source homogeneity also led to uncertainty regarding the number of voices in the block, and this uncertainty was increased by formant-space differences between voices. Finally, in cases where voices shared source and formant-space characteristics (effectively a single-voice condition), listeners reported being unsure of the number of voices in the block in 14.3% of cases, indicating that the experimental design may have led to a hyper-awareness of speaker-changes.

[1]Extrinsic information is information which is not carried by a vowel sound itself, while intrinsic information is carried within the vowel (Ainsworth, 1975, Nearey 1989). For example, the average pitch or formant frequencies of a carrier phrase that precedes a vowel is extrinsic to the vowel, while the formant frequencies and pitch of the vowel are intrinsic to it.

[2]The distinction between active and passive control structures, and their implications for theories of normalization is discussed in detail in Magnuson and Nusbaum (2007). In short, active control structures allow for the same input to result in different outputs based on the specific listening situation, while passive control structures feature a predictable and rigid relationship between input and output regardless of context.

[3]This experiment is a replication of Experiment 4 in Nusbaum and Magnuson (1992). The pattern of results reported for that experiment are

generally consistent with what is reported in Experiment 1 of Magnuson and Nusbaum (2007). Unfortunately, the authors do not provide a full accounting of results, nor do they provide a useful description of their vowel stimuli. For those reasons, the results of that experiment will not be discussed here.

Ainsworth, W. (**1975**). "Intrinsic and extrinsic factors in vowel judgments," in *Auditory Analysis and Perception of Speech*, edited by G. Fant and M. Tatham (Academic, London), pp. 103–113.

Assmann, P. F., Nearey, T. M., and Hogan, J. T. (**1982**). "Vowel identification: Orthographic, perceptual, and acoustic aspects," J. Acoust. Soc. Am. **71**(4), 975–989.

Creelman, C. D. (**1957**). "Case of the unknown talker," J. Acoust. Soc. Am. **29**, 655.

Forster, K. I., and J. C. Forster. (**2003**). "DMDX: A Windows display program with millisecond accuracy," Behav. Res. Methods Instrum. Comput. **35**(1), 116–124.

Gerstman, Louis. (**1968**). "Classification of self-normalized vowels," IEEE Trans. Audio Electroacoust. **AU-16**, 78–80.

Hillenbrand, J., Getty, L. A., Clark, M. J., and Wheeler, K. (**1995**). "Acoustic characteristics of American English vowels," J. Acoust. Soc. Am. **97**(5), 3099–3111.

Joos, M. (**1948**). "Acoustic phonetics," Language **24**(2), 5–136.

Klatt, D. H., and Klatt, L. C. (**1990**). "Analysis, synthesis, and perception of voice quality variations among female and male talkers," J. Acoust. Soc. Am. **87**(2), 820–857.

Ladefoged, P., and Broadbent, D. E. (**1957**). "Information conveyed by vowels," J. Acoust. Soc. Am. **29**(1), 98–104.

Macchi, M. J. (**1980**). "Identification of vowels spoken in isolation versus vowels spoken in consonantal context," J. Acoust. Soc. Am. **68**(6), 1636–1642.

Magnuson, J. S., and Nusbaum, H. C. (**2007**). "Acoustic differences, listener expectations, and the perceptual accommodation of talker variability," J. Exp. Psychol. Hum. Percept. Perform. **33**(2), 391–409.

Martin, C. S., Mullennix, J. W., Pisoni, D. B., and Summers, W. V. (**1989**). "Effects of talker variability on recall of spoken word lists," J. Exp. Psychol. Learn. Mem. Cogn. **15**, 676–684.

Mullennix, J. W., Pisoni, D. B., and Martin, C. S. (**1989**). "Some effects of talker variability on spoken word recognition," J. Acoust. Soc. Am. **85**(1), 365–378.

Nearey, T. (**1978**). *Phonetic feature systems for vowels*, (Indiana University Linguistics Club, Bloomington, IN).

Nearey, T. M. (**1989**). "Static, dynamic, and relational properties in vowel perception," J. Acoust. Soc. Am. **85**(5), 2088–2113.

Nearey, T. M., and Assmann, P. F. (**1986**). "Modeling the role of inherent spectral change in vowel identification," J. Acoust. Soc. Am. **80**(5), 1297–1308.

Nearey, T. M., and Assmann, P. F. (**2007**). "Probabilistic 'sliding template' models for indirect vowel normalization," in *Experimental Approaches to Phonology*, edited by M. J. Sole', P. S. Beddor, and M. Ohala (Oxford University Press, Oxford), pp. 246–269.

Nusbaum, H. C., and Magnuson, J. S. (**1997**). "Talker normalization: Phonetic constancy as a cognitive process," in *Talker Variability in Speech Processing*, edited by K. Johnson and J. W. Mullennix (Academic, San Diego, CA), pp. 109–132.

Nusbaum, H. C., and Morin, T. M. (**1992**). "Paying attention to differences among talkers," in *Speech Perception, Speech Production, and Linguistic Structure*, edited by Y. Tohkura, Y. Sagisaka, and E. Vatikiotis-Bateson (OHM, Tokyo), pp. 113–134.

Nusbaum, H. C., and Schwab, E. C. (**1986**). "The role of attention and active processing in speech perception," in *Pattern recognition by humans and machines: Vol. 1. Speech perception*, edited by E. C. Schwab, and H. C. Nusbaum, (Academic, San Diego, CA), pp. 113–157.

Peterson, G. E., and Barney, H. L. (**1952**). "Control methods used in a study of the vowels," J. Acoust. Soc. Am. **24**(2), 175–184.

Rakerd, B., Verbrugge, R. R., and Shankweiler, D. P. (**1984**). "Monitoring for vowels in isolation and in a consonantal context," J. Acoust. Soc. Am. **76**(1), 27–31.

Smith, D. R. R., Patterson, R. D., Turner, R., Kawahara, H., and Irino, T. (**2005**). "The processing and perception of size information in speech sounds," J. Acoust. Soc. Am. **117**(1), 305.

Syrdal, A. K., and Gopal, H. S. (**1986**). "A perceptual model of vowel recognition based on the auditory representation of American English vowels," J. Acoust. Soc. Am. **79**(4), 1086–1100.

Verbrugge, R., Strange, W., and Shankweiler, D. (**1974**). "What information enables a listener to map a talker's vowel space?," J. Acoust. Soc. Am. **55**(S1), S53–S54.

Watkins, A. J., and Makin, S. J. (**1996**). "Effects of spectral contrast on perceptual compensation for spectral-envelope distortion," J. Acoust. Soc. Am. **99**, 3749.

Watkins, Anthony, J., and Makin, S. J. (**1994**). "Perceptual compensation for speaker differences and for spectral-envelope distortion," J. Acoust. Soc. Am. **96**(3), 1263–1282.

Whalen, D. H., Abramson, A. S., Lisker, L., and Mody, M. (**1993**). "F0 gives voicing information even with unambiguous voice onset times," J. Acoust. Soc. Am. **93**(4), 2152–2159.

Wong, P. C. M., Nusbaum, H. C., and Small, S. L. (**2004**). "Neural bases of talker normalization," J. Cogn. Neurosci. **16**(7), 1173–1184.