The effect of exposure to a single vowel on talker normalization for vowels

John R. Morton[a)] and Mitchell S. Sommers
Department of Psychology, Washington University in St. Louis, Campus Box 1125, St. Louis, Missouri 63130

Steven M. Lulich
Department of Speech and Hearing Sciences, Indiana University, 200 South Jordan Avenue, Bloomington, Indiana 47405

(Received 4 March 2014; revised 13 January 2015; accepted 10 February 2015)

The current work investigated the role of single vowels in talker normalization. Following initial training to identify six talkers from the isolated vowel /i/, participants were asked to identify vowels in three different conditions. In the *blocked-talker* conditions, the vowels were blocked by talker. In the *mixed-talker* conditions, vowels from all six talkers were presented in random order. The *precursor mixed-talker* conditions were identical to the mixed-talker conditions except that participants were provided with either a sample vowel or just the written name of a talker before target-vowel presentation. In experiment 1, the precursor vowel was always spoken by the same talker as the target vowel. Identification accuracy did not differ significantly for the blocked and precursor mixed-talker conditions and both were better than the pure mixed-talker condition. In experiment 2, half of the trials had a precursor spoken by the same talker as the target and half had a different talker. For the same-talker precursor condition, the results replicated those in experiment 1. In the different-talker precursor, no benefit was observed relative to the pure-mixed condition. In experiment 3, only the written name was presented as a precursor and no benefits were observed relative to the pure-mixed condition. © *2015 Acoustical Society of America*.
[http://dx.doi.org/10.1121/1.4913456]

[CGC] Pages: 1443–1451

I. INTRODUCTION

One of the central issues in research on speech perception has been to establish how listeners obtain invariant linguistic information from a highly variable acoustic signal. Sources of acoustic-phonetic variability include phonetic context (Liberman *et al.*, 1967), speaking rate (Miller, 1981; Miller and Dexter, 1988; Wayland *et al.*, 1994), and talker characteristics (Mullennix *et al.*, 1989; Pisoni, 1996), among others (Pisoni and Remez, 2008). Thus, listeners are faced with a significant perceptual challenge when trying to understand spoken language; namely, there is both a many-to-one and a one-to-many mapping between acoustic speech signals and phonetic perception. For the most part, however, listeners have relatively little difficulty maintaining perceptual constancy and can usually recognize target items correctly despite the extensive acoustic-phonetic variability present in speech signals.

The focus of the current work is on how listeners' are able to accommodate one source of acoustic-phonetic variability—differences in the size and shape of vocal tracts across talkers (talker variability). Specifically, we will examine how listeners maintain consistency in the perception of vowel quality, despite talker-dependent variation in acoustic features known to be important for vowel discrimination.

In their seminal paper on vowel identification, Peterson and Barney (1952) demonstrated large areas of overlap for vowel categories as defined by the first (F1) and second (F2) resonances (formants) of the vocal tract. Although a replication of the original Peterson and Barney study (Hillenbrand *et al.*, 1995) reported somewhat less overlap between adjacent vowel categories than reported by Peterson and Barney, with just a few exceptions, it was still not possible to identify a unique and invariant mapping between vowel quality and specific values of F1 and F2 frequencies for men, women, and children.

The most common account of how listeners maintain perceptual constancy across talkers is generally referred to as talker normalization, and suggests that speech signals or their representations undergo one or more transformations that "normalize" them to a standardized representation and these representations are then compared to canonical forms stored in long-term memory.[1] Broadly, theories of talker normalization can be divided into two (not mutually exclusive) categories. *Intrinsic* theories (Syrdal and Gopal, 1986; Traunmüller, 1981) suggest that the information needed for normalization is syllable internal, including fundamental frequency (F0) and the frequencies of the higher formants, primarily F3–F5 which generally remain stable within a talker. In contrast, *extrinsic* theories of perceptual normalization (Ladefoged and Broadbent, 1957; Nearey, 1989) propose that properties external to the vowel, such as the range of formant frequencies, are the primary basis for perceptual

[a)] Author to whom correspondence should be addressed. Electronic mail: jrmhvc333@yahoo.com normalization because they provide a frame of reference with which to consider specific F1 and F2 values.

The current investigation is directed at examining extrinsic talker normalization during vowel identification. Extrinsic theories of talker normalization differ considerably with respect to the nature of the transformation and the amount and type of information needed for successful normalization (see Johnson, 2008, for a review). Joos (1948), for example, proposed that both articulatory patterns and acoustic information could serve as a basis for perceptual normalization. On the other hand, both Nearey (1989) and Ladefoged and Broadbent (1957) suggested that normalization operates exclusively on acoustic information (especially the formants and F0). Differences also exist within acoustic-based theories of normalization as to whether scaling of acoustic parameters during normalization is uniform (Nordstrom and Lindblom, 1975) or non-uniform (Nearey, 1989).

Despite these important differences, the general proposal of extrinsic-based theories of talker normalization is that listeners use a sample of a talker's productions to gain information about vocal-tract (and perhaps articulatory) properties specific to that individual. The normalization process then uses the individual's vocal-tract characteristics to transform the incoming speech signal into a standardized representation that serves as the basis for perceptual identification. Joos (1948) elegantly described the general idea of talker normalization as follows: "On first meeting a person, the listener hears a few vowel phones, and on the basis of this small but apparently sufficient evidence he swiftly constructs a fairly complete vowel pattern to serve as a background (coordinate system) upon which he correctly locates new phones as fast as he hears them…" (p. 61). Nusbaum and Morin (1992) proposed a similar idea in their contextual tuning theory of talker normalization. They suggested that the normalization process is triggered immediately after encountering a new talker and serves to map talker-specific acoustic values onto internal phonetic categories. An important aspect of the normalization system is that it appears to maintain a given set of mapping parameters until a change in talker is detected. Upon detecting the new talker, the normalization system is reset to map the new talker's productions onto canonical representations.

One line of evidence that has been used to support a talker normalization mechanism are studies in which the perception of a target item varies as a consequence of changes in a precursor stimulus. In one of the earliest and most influential studies (Ladefoged and Broadbent, 1957) participants were presented with a precursor carrier phrase in which vowel formant frequencies were manipulated to have either relatively high or low values and were then asked to judge the quality of a subsequent target vowel. The critical finding from this study was that vowel quality for the identical stimulus varied depending on the frequency of formants in the precursor phrase. Using the terminology suggested by Joos (1948), the precursor phrase provided a coordinate system and listeners judged subsequent vowel quality of the target according to the specific "system" derived from formant frequency values in the precursor phrase. More recently, Sjerps *et al.* (2011) found that vowel identification could be altered by changing the long-term average spectrum (LTAS) of a precursor phrase. Participants were asked to identify items on a [pɪt]–[pɛt] (low–high F1) continuum immediately following a precursor phrase in which the frequency of F1 was either low or high. Their findings indicated that more items were labeled as [pɪt] following a high-F1 precursor than following a precursor with a low F1. Interestingly, similar findings were obtained when the stimuli were spectrally rotated, producing nonspeech sounds that had speech-like acoustic properties (see also, Huang and Holt, 2012). Whalen *et al.* (1995) also found that vowel identification could be altered by changes in a nonspeech precursor—in this case natural inspiration sounds produced by a male and a female.

Other studies of vowel normalization using precursors have produced somewhat mixed results. Kato and Kakehi (1988) found significant and progressive improvements in vowel identification when vowels were repeated from 1 to 5 times consecutively (no additional improvements were observed after five consecutive repetitions). From the perspective of extrinsic talker normalization, the improvements in vowel identification across five repetitions was, at least in part, a consequence of listeners reducing normalization demands by gaining additional knowledge of vocal-tract properties across the five repetitions. In contrast, Verbrugge *et al.* (1976) reported no benefit for vowel identification from having any one of six vowels (/i/, /ɑ/, /a/, /u/, /ɪ/, /ʌ/) as precursors.

A second line of research that has been used as support for a talker normalization mechanism is research examining the effects of talker variability on speech perception (Mullennix *et al.*, 1989; Pisoni, 1996). In these studies, performance for lists of items spoken by a single individual (blocked-talker lists) is compared to performance for the same items spoken by a number of different talkers (mixed-talker lists). The rationale for these studies is that talker normalization is considered a perceptually demanding process that engages resources that could otherwise be used for perceptual identification. Thus, the prediction is that despite identical linguistic content, performance for the mixed-talker lists will be significantly poorer than for the blocked-talker lists because the trial-to-trial variations in talker will require listeners to redirect perceptual resources that could otherwise be used for perceptual identification for use in the perceptual normalization process. From the perspective of contextual tuning theory (Nusbaum and Morin, 1992), normalization demands are minimal in the blocked-talker condition because once listeners develop a stable representation of the talker's vocal tract properties they can use this information as a basis for identifying subsequent stimuli produced by that talker. In the mixed-talker condition, however, the trial-to-trial changes in talker require listeners to continually engage in the talker normalization process to acquire vocal-tract properties of the new speaker. It is the additional normalization demands in the mixed-talker condition that is hypothesized to produce decrements in identification performance relative to a single-talker condition.

Consistent with these proposals, Mullennix *et al.* (1989) found that identification accuracy for lists of CVC words was 10%–15% poorer when list items were spoken by multiple talkers than when the same stimuli were spoken by a single talker. Subsequently, similar negative effects of talker variability have been reported for infants (Jusczyk *et al.*, 1989), older adults (Sommers *et al.*, 1995), individuals with Alzheimer's disease (Sommers, 1998), and listeners with hearing loss (Sommers, 1997).

Other studies examining the perceptual costs of talker variability (Magnuson and Nusbaum, 2007; Nusbaum and Morin, 1994) provide converging evidence that decrements in identification performance for mixed-talker compared with blocked-talker conditions is a consequence of additional normalization requirements when list items are spoken by multiple talkers. Nusbaum and Morin (1994), for example, used a dual-task paradigm in which participants were asked to remember either 1 (low-load condition) or 3 (high-load condition) digits prior to a speeded vowel identification task. During the identification task participants were asked to name vowels as quickly and accurately as possible for conditions in which the stimuli were spoken either by a single talker or by a mix of four talkers (2 males and 2 females). In the low-load conditions, similar response latencies were observed for the single- and mixed-talker presentations. Latencies in the high-load single-talker condition also did not differ from either of the two low-load conditions (single and multiple talkers). Only in the most difficult condition, high-load with multiple talkers, was there a significant increase in response latencies. Nusbaum and Morin suggested that the additional processing required to remember 3 items, rather than 1 reduced resources available for talker normalization. Wong *et al.* (2004) used fMRI to compare activation levels in several cortical regions for single- versus mixed-talker conditions. The regions examined, superior temporal and superior parietal areas, are sensitive to task difficulty and Wong *et al.* reported that activation levels in these areas were significantly greater for the mixed- than for the blocked-talker condition.

In the present set of experiments, we examined whether providing listeners with a precursor vowel from a target talker would attenuate (or eliminate) the negative effects of talker variability for vowel identification in mixed-talker lists. The rationale for this approach is that presenting a precursor vowel immediately prior to a target vowel spoken by the same talker should provide listeners with information about that talker's vocal-tract properties, thereby reducing normalization demands and increasing correct identification of target vowels. Thus, in these experiments we compared vowel identification for (1) a blocked-talker condition, (2) a mixed-talker condition, and (3) a precursor condition. This last (precursor) condition was a mixed-talker condition in which listeners were presented with information about the talker who would produce a subsequent target vowel. If, as predicted by contextual tuning and other theories of talker normalization, hearing a new talker initiates a resource-demanding normalization process, then we would expect poorer identification scores for mixed-, compared with blocked-talker conditions as has been shown previously for both words (Mullennix *et al.*, 1989) and vowels (Verbrugge *et al.*, 1976). Of particular importance, we would also expect that identification scores in the mixed-talker condition with a precursor would be significantly better than the mixed-talker condition without precursor and would not differ from a blocked talker condition.

II. EXPERIMENT 1

A. Method

1. Participants

A total of 12 participants (8 male) took part in experiment 1. All participants were young adults between the ages of 18 and 24 and all were native speakers of English. None of the participants reported a history of speech or language disorders. Participants were paid $10/h for participation. All procedures were approved by the Washington University Institutional Review Board.

2. Stimulus materials

Stimuli were taken from the WashU-UCLA Corpus of speech recordings (Lulich *et al.*, 2012). This corpus consists of CVC words embedded in the carrier phrase "I said a ____ again," where the CVCs were hVd, bVb, dVb, or gVb. Target stimuli for the current study were six of the vowels recorded in the hVd context [ɛ, æ, ɑ, ʌ, o, and ʊ]. During recording, the carrier phrase and target word were presented on a computer monitor in a sound-attenuated booth ten times each in random order. Recordings were made simultaneously with a SHURE PG27 microphone and a K&K Sound HotSpot accelerometer from 50 adult native speakers of American English (ages 18–25; 25 females). The start, steady-state, and end times of the vowel in each CVC were manually labeled using PRAAT (Boersma, 2001) and the vowels were then excised and saved as separate waveforms.

For the experiments described in this paper, three male and three female speakers' recordings were used (labeled as speakers s12, s13, s14, s15, s16, and s18 in the Corpus). Table I displays means and standard deviations for fundamental frequency (F0) and the first three formant frequencies (F1, F2, F3) for all vowels used in the current experiments.

The target vowels in all experiments were excised from the hVd words of each of the six speakers. Vowels were excised from their manually labeled start and end points. As there were 10 repetitions of each vowel, there were a total of 360 target stimuli (10 repetitions × 6 vowels × 6 talkers). A set of precursor stimuli was also used in these experiments. These stimuli were tokens of the vowel /i/ excised from their manually labeled start and end points in the bVb words of each speaker. There were a total of 60 precursor stimuli (10 from each of the 6 talkers).

3. Procedures common to all experiments

In all experiments, participants completed a total of four tasks, three of which were common to all experiments and one that was experiment specific. The tasks common to all experiments were: name-voice learning, blocked-talker condition, and mixed-talker condition. Each of these is described below, followed by the task specific to experiment 1. All testing was conducted individually with participants seated in a double-walled sound-attenuated booth.

TABLE I. Means and standard deviations (shown in parentheses) for the fundamental (F0) and first three formants (F1, F2, F3) of the ten tokens of each vowel presented to listeners. ID refers to the speaker number as labeled in the WashU-UCLA corpus (Lulich *et al.*, 2012).

ID	Sex	Measure	[ɛ]	[æ]	[ɑ]	[ʌ]	[o]	[ʊ]	[i]
12	M	F0	103.71 (2.71)	100.64 (4.83)	100.52 (5.78)	103.98 (5.86)	104.22 (4.21)	109.3 (7.4)	109.68 (3.26)
		F1	599.75 (15.24)	734.83 (29.22)	691.22 (23.99)	643.24 (12.58)	448.12 (18.77)	449.13 (10.67)	278.79 (18.27)
		F2	1876.47 (324.73	1675.82 (104.9)	1086.16 (21.25)	1262.59 (34.79)	982.56 (53.03)	1173.02 (33.75)	2288.24 (49.2)
		F3	2772.63 (435.11)	2732.43 (204.89)	2741.78 (187.09)	2597.61 (187.09)	2542.56 (98.86)	2613.43 (179.5)	3034.97 (54.14)
13	M	F0	164.47 (5.3)	164.54 (7.3)	174.05 (6.3)	170.42 (4.36)	173.18 (2.94)	179.5 (9.47)	180.58 (9.41)
		F1	633.17 (21.93)	750.98 (29.93)	734.06 (49.83)	680.31 (25.76)	510.64 (20.39)	509.26 (26.77)	278.47 (23.38)
		F2	1825.15 (61.02)	1664.01 (97.33)	1202.62 (47.68)	1346.05 (62.19)	1103.12 (49.48)	1270.39 (44.84)	2333.82 (46.68)
		F3	2513.61 (40.92)	2505.1 (45.16)	2699.75 (61.85)	2650.8 (102.31)	2668.19 (85.03)	2589.4 (67.51)	3166.22 (77.21)
14	F	F0	192.58 (4.51)	175.24 (29.88)	184.82 (3.96)	190.61 (4.88)	195.89 (4.49)	200.07 (4.12)	187.11 (31.16)
		F1	626.26 (16.67)	828.47 (27.99)	785.94 (23.89)	650.49 (33.21)	561.09 (32.28)	592.15 (11.96)	380.55 (18.53)
		F2	1942.22 (36.4)	1785.39 (39.92)	1438.58 (32.77)	1631.85 (32.23)	1425.5 (67.93)	1633.19 (66.08)	2663.36 (32.01)
		F3	2720.38 (38.7)	2571.49 (73.78)	2484.28 (115.23)	2570.92 (74.25)	2582.94 (55.3)	2626.63 (51.29)	2996.75 (147.33)
15	M	F0	112.19 (5.48)	108.99 (5.59)	104.99 (2.69)	112.51 (2.41)	116.36 (3.3)	120.41 (2.99)	118.9 (3.04)
		F1	581.44 (21.69)	731.1 (16.9)	743.21 (26.49)	623.65 (21.55)	432.01 (14.27)	477.52 (17.16)	273.76 (11.07)
		F2	1925.95 (82.41)	1720.79 (81.32)	1269.69 (38.33)	1424.02 (33.33)	969.26 (34.17)	1251.89 (42.61)	2640.89 (32.77)
		F3	2874.44 (117.04)	2751.86 (76.49)	2491.05 (86.37)	2828.33 (98.82)	2748.97 (100.63)	2736.89 (48.84)	3470.55 (158.76)
16	F	F0	187.02 (12.63)	180.73 (13.2)	187.57 (4.27)	191.38 (5.27)	198.53 (5.23)	194.19 (4.55)	197.68 (16.61)
		F1	740.33 (28.56)	979.74 (31.69)	839.49 (30.71)	708.04 (25.19)	529.67 (11.98)	579.99 (18.49)	386.26 (4.79)
		F2	1936.34 (60.38)	1750.78 (86.03)	1366.94 (40.17)	1571.45 (27.01)	1134.27 (72.6)	1555.35 (58.88)	2697.09 (50.98)
		F3	3040.22 (65.44)	2802.19 (186.79)	2971.53 (54.15)	2998.73 (32.8)	2909.99 (47.94)	2904.28 (34.45)	3295.47 (61.86)
18	F	F0	136.37 (28.13)	135.48 (23.5)	145.52 (16.19)	131.13 (43.19)	168.13 (14.74)	157.66 (30.65)	126.4 (43.53)
		F1	763.3 (48.43)	1010.82 (70.06)	816.35 (35.46)	721.31 (64.5)	429.23 (27.04)	540.65 (26.8)	303.91 (14.37)
		F2	1891.14 (54.85)	1730.87 (76.35)	1303.84 (42.14)	1488.46 (45.61)	1036.67 (91.78)	1348.02 (19.45)	2829.55 (64.44)
		F3	2937.86 (60.46)	2917.23 (61.11)	3051.49 (108.69)	2972.21 (46.16)	2821.27 (50.43)	2971.85 (54.43)	3558.43 (170.74)

a. Name learning. This task was always the first one completed and required participants to learn associations between printed names and spoken voices. The main goal of this phase was to provide participants with sufficient experience with the different talkers to enable correct voice identification of each of the six talkers. On each trial, a fictitious name of one of the six speakers was presented on the computer monitor followed by the vowel /i/ from one of the three same-gendered talkers (e.g., if a female name was shown, then participants heard an /i/ by one of the three female talkers). The name and voice matched on half the trials and were mismatched on the remaining half (divided evenly between the two remaining non-target talkers). After hearing the vowel, participants saw the question: "Did the speaker name match the speaker voice?" presented on the computer monitor. Participants could repeat the vowel stimulus on any given trial a maximum of two additional times before answering. If they answered correctly, they could initiate the next trial by pressing the mouse button after a 750-ms inter-trial-interval (ITI). If they answered incorrectly, they were given feedback via the computer monitor indicating the actual name of the speaker. Following feedback and a 750-ms ITI, participants pressed the mouse button to continue to the next trial. Participants heard a minimum of 60 trials (10 different instances of the vowel /i/ by each of the 6 talkers). If participants achieved an accuracy rate of 80% or better after the initial 60 trials, this phase of the experiment was terminated. If accuracy rates did not exceed 80% after the initial 60 trials, testing continued until the participant reached the 80% criterion or exceeded 420 trials. If the participants still had not met the 80% criterion after 420 trials, they were excluded from further testing. The number of participants who failed to reach the 80% criterion was 6, 5, and 7 for experiments 1, 2, and 3, respectively. These individuals were replaced to obtain the sample size listed. Although this represents a relatively large number of young adults with normal hearing who were unable to learn name-voice associations, similar levels of individual variability in voice learning have been reported previously (Nygaard and Pisoni, 1998).

b. Blocked-talker vowel identification. In the blocked-talker condition, participants heard 10 different versions of each of the 6 target vowels from each of the six talkers for a total of 360 presentations. The 60 vowels from each talker were presented consecutively (i.e., blocked) with the order of vowel presentation randomized for each talker. Furthermore, the order in which talkers were presented was also selected randomly for each participant.

Each trial began with the word "START" presented on the computer monitor and participants were required to press the mouse button to initiate a trial. Stimuli were presented binaurally over headphones (Beyerdaynamic DT 801) in a background of white noise at a signal-to-noise ratio of −8 dB. Each sample of the noise was generated independently and came on 100 ms prior to target onset and was terminated 100 ms after target offset. After the target stimulus was presented, participants saw a response screen with six vowel alternatives and an example of a common English word containing each of the target vowel sounds. Participants responded by selecting the response alternative that they thought had been presented and this was followed by a 1-s ITI. At the beginning of this task, participants received 30 practice trials presented in noise that included five vowel presentations from each of the 6 talkers.

c. Mixed-talker vowel identification. This condition was identical to the blocked-talker task except that the talker for each presentation was selected pseudo-randomly (without replacement) on each trial. Thus, participants heard the same 360 target vowels in the blocked- and mixed-talker conditions, but talker was varied from trial to trial in the mixed-talker condition.

d. Vowel identification with auditory precursor and printed name. This task was specific to experiment 1 and was identical to the mixed-talker condition except that prior to target vowel presentation participants saw the name of one of the six talkers on the computer monitor and simultaneously heard the vowel /i/ as a precursor (the precursor /i/ was never used as a target vowel) spoken by that same talker. The name was presented in the center of the computer monitor in black letters with a white background. The name appeared on the screen 100 ms prior to the precursor vowel and stayed visible throughout the trial. The mean duration for the precursor vowel across the six talkers was 118 ms (standard deviation = 27 ms). The auditory precursor (/i/) and the target vowel were always spoken by the same talker. In experiment 1, both the precursor and target were presented in a background of white noise at an SNR of -8 dB. Independent samples of white noise were generated for both the precursor and target. Precursor and target were separated by a 75-ms inter-stimulus-interval.

The order in which subjects completed tasks 2, 3, and 4 (blocked, mixed, precursor, respectively) was counterbalanced, such that across all participants each condition was presented second, third, or fourth (recall that the name-voice association task was always completed first) an equivalent number of times.

B. Results and discussion

Figure 1 displays means and standard errors for proportion correct vowel identification for the three conditions tested in experiment 1. To examine whether identification performance differed across the three conditions we conducted a one-way analysis of variance (ANOVA) with condition as a repeated measures variable and identification performance as the dependent measure. The analysis indicated significant differences across the three conditions, $F(2, 11) = 8.7$, $p < 0.01$, $\eta_p^2 = 0.44$. Pairwise *post hoc* comparisons[2] indicated significant differences between the blocked and mixed conditions ($p < 0.01$) and between the mixed and precursor conditions ($p < 0.05$), but no significant difference between the blocked and precursor conditions ($p = 0.46$). To examine whether benefits of a precursor might have differed for male and female talkers, we repeated the *post hoc* comparisons for the male and female talkers separately. Overall, the pattern of results was identical to the combined results: significant differences between the blocked and mixed conditions ($p < 0.01$ for both males and females); significant differences between the mixed and precursor conditions ($p < 0.05$ for both males and females), and no difference between the blocked and precursor conditions ($p = 0.35$ for males, $p = 0.58$ for females).

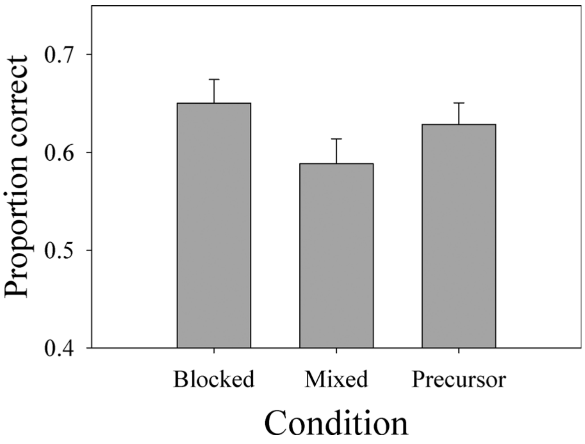

FIG. 1. Mean proportion correct vowel identification for the three conditions tested in experiment 1. In the blocked-talked condition, participants heard consecutive stimuli spoken by the same talker. In the mixed-talker condition, talkers were randomized from trial to trial. The precursor condition was identical to the mixed-talker condition except that participants heard a sample of the vowel /i/ spoken by the upcoming talker immediately prior to hearing the target vowel. Error bars represent standard error of the mean.

Taken together the results of experiment 1 suggest that providing listeners with a sample vowel prior to target identification reduces demands on talker normalization by providing an opportunity to gain information about the vocal tract characteristics of the target talker. The absence of a significant difference between the precursor and blocked conditions suggests that exposure to a single vowel from a target talker can produce performance for a mixed-talker condition that is not significantly different than that observed for a blocked-talker condition. Conversely, the significant difference between the mixed-talker and precursor conditions suggest that without prior exposure to a sample vowel, listeners incur a significant cost in a mixed-talker, compared with a blocked-talker or precursor condition.

If this account is correct, then it should be possible to modulate the effects of the precursor by varying whether the precursor and target vowel are spoken by the same talker. Thus, in experiment 2 on half the precursor trials the precursor and target vowel were spoken by the same talker. On the other half, however, the precursor and target were spoken by individuals of the same gender, but by different talkers. Based on the findings from experiment 1, we would expect to replicate the advantage for same-talker precursor trials as was observed in the first experiment. Furthermore, we would not expect differences between mixed-talker conditions with and without a precursor if the precursor and target vowels are spoken by different individuals.

III. EXPERIMENT 2

A. Methods

1. Participants

Twenty-four participants (18 female) from the same participant population used in experiment 1 were recruited for experiment 2. Inclusion and exclusion criteria were also the same as in experiment 1; participants had to be between 18 and 25 years old, native speakers of English, with no history of speech or hearing disorders.

2. Stimuli and procedures

With the following exceptions, the stimuli and procedures were identical to experiment 1. First, in the precursor condition, half the trials had the precursor and target vowel spoken by the same talker and half had a different, but gender-consistent talker for the precursor and target stimuli. These mismatch trials were divided evenly between the two remaining gender-consistent talkers. Same-talker and different-talker precursor trials were presented in random order within the same block. Second, to provide listeners with the clearest talker and vocal tract information about the precursor, the precursor was presented without any background noise (the target was still presented at a −8 signal-to-noise ratio). Finally, in experiment 2 we did not present the name of the talker in the precursor condition, as we wanted to establish whether we could obtain the effect using the precursor vowel alone.

B. Results

Figure 2 displays performance for the conditions tested in experiment 2. Overall, the pattern of results was similar to that observed in experiment 1. Specifically, a repeated measures ANOVA indicated that performance differed significantly across the four conditions $F(3, 69) = 6.8$, $p < 0.01$, $\eta_p^2 = 0.21$. *Post hoc* pairwise comparisons indicated that performance was significantly better in the blocked than in the mixed-talker condition ($p < 0.01$), significantly better in the matching-precursor than in the mixed-talker condition ($p < 0.01$), significantly better in the blocked than in the non-matching precursor condition ($p < 0.01$), and significantly better in the matching-precursor than in the non-matching precursor condition ($p < 0.01$). Identification did not differ between the blocked and matching precursor condition ($p > 0.6$) nor between the mixed and non-matching precursor conditions ($p > 0.8$).

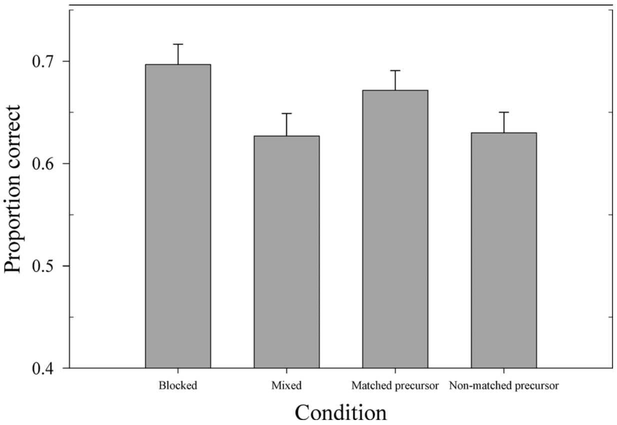

FIG. 2. Mean correct vowel identification for the four conditions tested in experiment 2. In the matching-precursor condition, the precursor and target vowel were spoken by the same talker and in the non-matching precursor condition, the talker for the precursor and target vowel were different (but same gender). Error bars represent standard error of the mean.

The findings from experiment 2 indicate that the improved performance for the mixed-talker with precursor condition, relative to the mixed-talker without precursor in experiment 1 was not simply a consequence of having a precursor. When precursor and target vowel were spoken by the same talker, there was a benefit (relative to the no precursor condition) for the mixed-talker condition such that performance did not differ significantly from the blocked-talker condition. However, when the precursor and target were spoken by different (but gender consistent) talkers, the addition of the precursor did not improve vowel identification relative to the mixed-talker condition.

IV. EXPERIMENT 3

The results of the first two experiments suggest that presenting listeners with a precursor vowel spoken by the same talker who produces a subsequent target vowel for identification significantly reduced differences in identification performance between blocked- and mixed-talker presentations. Recall, however, that in both experiments listeners initially learned voice-name associations in a learning phase prior to the vowel identification tasks. Thus, the acoustic precursor not only provided listeners with information about the upcoming talker's vocal tract but also about talker identity. In experiment 3, we wanted to establish whether providing listeners with just the orthographic name of a familiar talker as a precursor would also reduce normalization demands and produce similar identification performance for blocked- and mixed-talker with precursor conditions. If so, it would suggest that knowledge of a talker's identity is sufficient to reduce differences between blocked- and mixed-talker conditions in the absence of acoustic information about talkers' vocal tracts.

Although most current theories of talker normalization suggest that listeners require an acoustic signal (from a new talker) to initiate the normalization process, there are several findings that support the proposal that talker information other than that provided by the acoustic signal can affect talker normalization (Johnson *et al.*, 1999; Nygaard and Pisoni, 1998; Strand, 1999). Johnson *et al.* (1999), for example, found that providing a picture of either a male or female talker (without an accompanying acoustic signal) significantly altered listeners' categorization of items on a hood-hud continuum. In addition, they also asked participants to categorize stimuli whose F0 had been altered to a value intermediate between typical male and female values and found that instructions to imagine that the talker was male or female significantly altered the categorization function. These findings suggest that vowel quality can be altered based on knowledge about the gender (and perhaps identity) of a talker, even in the absence of an acoustic signal from that talker. In experiment 3, we investigated whether providing listeners with the name of a talker whose voice they had previously learned to recognize would reduce normalization demands and thereby lead to reduced differences between blocked- and mixed-talker lists.

A. Method

1. Participants

Twenty-four participants (11 female) were recruited from the same participant population as in the previous experiments. Inclusion and exclusion criteria were also the same as in the first two experiments.

2. Procedure

The stimuli and procedures were identical to those of experiments 1 and 2, with the exception that rather than both seeing the name of a talker and hearing a precursor vowel (/i/) spoken by that talker, participants only saw the talker's name. The name was presented in the center of the computer monitor in black letters with a white background. The name appeared on the screen 100 ms prior to the target vowel and stayed visible throughout the trial.

B. Results

Figure 3 displays mean identification scores for the three conditions tested in experiment 3 (blocked-talker, mixed-talker, name-precursor). A repeated measures ANOVA with condition as a repeated-measures variable and identification scores as the dependent variable revealed that identification scores differed significantly across the three conditions $F(2, 46) = 7.2$, $p < 0.001$, $\eta_p^2 = 0.25$. *Post hoc* pairwise comparisons indicated that correct identification was significantly higher in the blocked-talker condition than in the mixed-talker condition ($p < 0.002$). Performance was also significantly better in the blocked-talker than in the name-precursor condition ($p < 0.05$). Finally, performance in the mixed-talker and name-precursor conditions did not differ significantly ($p = 0.11$).

The results of experiment 3 suggest that knowledge of talker identity alone is not sufficient to reduce the costs of talker normalization. Listeners in experiment 3 completed the same voice familiarization task as in experiments 1 and 2 and were therefore familiar with the name-voice associations for the six talkers used in the experiment. However, presentation of the written name alone as the precursor did not reduce differences between mixed-talker and blocked-talker conditions.

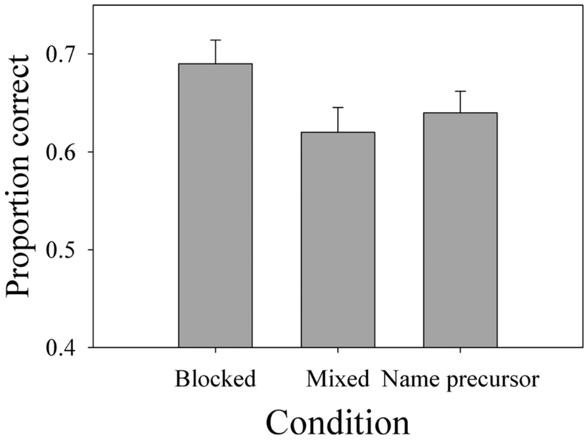

FIG. 3. Mean percent correct vowel identification for experiment 3. The blocked and mixed-talker conditions were the same as described in Fig. 1. The name precursor condition presented the written name only without a corresponding spoken sample. Error bars represent standard error of the mean.

V. GENERAL DISCUSSION

The findings from the current series of experiments advance our understanding of how listeners accommodate acoustic-phonetic variability due to differences across talkers in several ways. First, the findings demonstrate that listeners use vocal-tract information from a current talker to normalize subsequent speech tokens by that talker. Vowel identification for mixed-talker conditions was statistically indistinguishable from the corresponding blocked-talker conditions, but only when listeners heard a precursor vowel spoken by the target vowel talker. In contrast, performance in mixed-talker conditions without an auditory precursor or with a precursor spoken by a different talker was significantly lower than the corresponding blocked-talker conditions. Second, the same pattern of results (benefits of same-talker precursors, but no benefit for different-talker or no auditory precursor conditions) was observed whether the precursor contained both the name of a familiarized talker and a sample vowel from that talker or just the sample vowel. Finally, the results from experiment 3 suggest that the benefits of the precursor are not due to knowing the identity of the target vowel talker. Participants in experiment 3 had the same familiarization phase as in the first two experiments, but were presented with a written name of the target talker rather than an acoustic vowel sample. Without the sample vowel as a precursor, performance in the mixed and precursor conditions were not significantly different, but both were poorer than the blocked talker condition.

One potential concern in the present study is that the effects of adding a same-talker auditory precursor were relatively small, with differences between the pure-mixed and same-talker precursor conditions averaging approximately 5%. However, it is important to consider the benefits of providing the precursor relative to differences between the pure-blocked and pure-mixed conditions. As noted, Kato and Kakehi (1988) reported incremental improvements in vowel identification across five consecutive repetitions of a vowel by a given speaker, with no further improvements resulting from additional successive presentations. Thus, performance in the current blocked condition (where participants received 60 consecutive presentations by the same talker) likely represents close to asymptotic performance for isolated vowel identification under the present SNR conditions. Average differences between this pure-blocked and the pure-mixed (no precursor) conditions were between 6% and 7%. Thus, the approximately 5% improvement from a same-talker precursor relative to the pure-mixed condition represents close to ceiling level benefits. That is, maximum improvement from adding the precursor would equate performance in the blocked and mixed talker conditions and it came quite close to providing this optimal advantage.

In considering the overall benefits from the same-talker precursor, it is also important to note that the current study only examined one aspect of the talker normalization process—the ability to extract vocal tract information from an isolated vowel and use that information to identify subsequent (but different) vowels from that same talker. It is almost certainly the case that additional mechanisms contribute to listeners' ability to accommodate acoustic-phonetic variability arising from differences across talkers. Several researchers (Nearey, 1989), for instance, have suggested that information within an individual vowel (intrinsic normalization) may also serve as a basis for talker normalization. Nordstrom and Lindblom (1975) proposed that because F3 is causally linked to vocal tract length, listeners may use F3 information as a basis for constructing a scaling factor to shift vowel formants into a talker-independent coordinate system. Syrdal and Gopal (1986) incorporated both F3 and F0 information as part of a formula used to place vowels within a common perceptual system that can be used for talker-independent vowel identification. What these proposals share is the idea that information within a target vowel makes an important contribution to accommodating talker-based acoustic-phonetic variability. This type of vowel-intrinsic normalization is quite distinct from the vowel-extrinsic mechanisms investigated in the current proposal—the former relying on information within a single utterance and the latter on information outside that utterance—and suggests that multiple mechanisms may contribute to the process of talker normalization.

Consistent with this proposal, Nusbaum and Morin (1992) investigated the role of structural estimation (intrinsic normalization) and contextual tuning (extrinsic normalization) on the latency of vowel detection. Listeners were required to respond as quickly and accurately as possible following the presentation of four target vowels in a sequence of 16 vowels consisting of both target and non-target vowels. In one condition (blocked talker), vowels were all spoken by the same talker and in a second condition (mixed talker) they were spoken by four different talkers, with a different talker heard on successive presentations. Overall accuracy exceeded 95% for both the mixed- and blocked-talker conditions. This level of accuracy for the mixed-talker condition provides strong support for an intrinsic (structural estimation) mechanism of talker normalization because contextual tuning (extrinsic normalization) was not possible in the mixed-talker condition (because the talker changed on each presentation). Reaction times were significantly faster, however, for the blocked-talker condition, supporting an extrinsic basis for talker normalization; the initial few presentations in the blocked-talker condition likely depended on structural estimation, but consistency in talker across the 16 vowels allowed listeners to develop a representation of the talker's vocal tract properties and thereby reduce normalization demands relative to the mixed-talker condition. The combined use of both intrinsic and extrinsic talker normalization mechanisms is by no means surprising—typical listening conditions include ones in which a single talker produces speech over extended periods of time (e.g., lectures) as well as ones in which there are rapid changes between talkers (e.g., conversations). The operation of both intrinsic and extrinsic normalization mechanisms would account for findings that listeners generally have little or no difficulty with speech perception in either of these situations.

One limitation of the current work is that although the findings suggest that extrinsic (i.e., syllable external) normalization can provide a mechanism for obtaining invariant vowel quality despite acoustic variability due to talker differences, they do not identify the specific acoustic properties nor the nature of the transformations that mediate talker normalization. For example, uniform scaling of formant frequencies can reduce differences between talkers (Nordstrom and Lindblom, 1975), but there is also support for non-uniform scaling of formant frequencies. Fant (1966), for example, used separate scaling factors for the first three formants of each vowel (i.e., 30 different scaling factors for each talker in a 10-vowel system). Acoustic differences across talkers can also result from factors other than those associated with vocal-tract size and shape, including articulatory style (Henton, 1989; Henton and Bladon, 1985). Subsequent studies examining the basis of talker normalization should therefore focus on systematic manipulation of both individual and combined acoustic features as an approach for establishing the precise nature of the transformations that can lead to successful normalization.

A second limitation is that listeners were familiarized with the talkers producing both the target vowels and the precursors prior to the vowel identification task. Thus, it remains unclear whether the benefits of providing a same-talker precursor would produce similar benefits for normalizing unfamiliar talkers. In the absence of any degradation, listeners generally have little or no difficulty understanding speech produced by an unfamiliar talker even if that individual has a highly distinct accent or dialect. The findings from the current study suggest that one way listeners can accomplish such an impressive perceptual feat is by using quite short samples—in the current study a vowel lasting approximately 100 ms—of a talker's speech as a basis of normalization. Future research, however, will need to establish whether such a mechanism can operate as quickly and efficiently when normalizing speech produced by an unfamiliar talker.

VI. CONCLUSIONS

Considered together, the findings from the current experiments suggest that listeners can extract vocal tract information about a talker even from relatively short samples of the talker's speech (in this case a single isolated vowel). In addition, the present results suggest that presentation of a sample vowel provides general information about vocal tract properties rather than specific details about how individual vowels are produced; in all of the current experiments the precursor vowel was always one (/i/) that did not serve as a target stimulus. Presumably, the precursor provided the listener with at least partial information about the vowel coordinate system of individual talkers, and listeners were able to use that information as a basis for talker normalization. The goal for future research will be to establish how this information as well as other processes combine to produce a remarkably robust system for maintaining perceptual constancy despite extensive acoustic phonetic variability in the speech signal.

[1] A number of mechanisms other than talker normalization have been proposed to account for perceptual constancy in vowel perception, but a review of these is beyond the scope of the current work. The reader is referred to Johnson (1990) for a discussion of these alternative mechanisms.

[2] In this and all remaining *post hoc* comparisons, alpha levels were adjusted using a Bonferroni correction for multiple comparisons.

Boersma, P. (**2001**). "PRAAT, a system for doing phonetics by computer," Glot Int. **5**(9/10), 341–345.

Fant, G. (**1966**). "A note on vocal tract size factors and non-uniform F-pattern scalings," Speech Trans. Lab. Quart. Prog. Stat. Rep. **4**, 22–30.

Henton, C. G. (**1989**). "Fact and fiction in the description of female and male pitch," Lang. Commun. **9**, 299–311.

Henton, C. G., and Bladon, R. A. (**1985**). "Breathiness in normal female speech: Inefficiency versus desirability," Lang. Commun. **5**, 221–227.

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

Huang, J., and Holt, L. L. (**2012**). "Listening for the norm: Adaptive coding in speech categorization," Front. Psych. **3**, 10.

Johnson, K. (**1990**). "The role of perceived speaker identity in F0 normalization of vowels," J. Acoust. Soc. Am. **88**, 642–654.

Johnson, K. (**2008**). "Speaker normalization in speech perception," in *The Handbook of Speech Perception*, edited by D. B. Pisoni and R. E. Remez (Blackwell, Malden, MA), pp. 363–389.

Johnson, K., Strand, E. A., and D'Imperio, M. (**1999**). "Auditory-visual integration of talker gender in vowel perception," J. Phonetics **27**, 359–384.

Joos, M. (**1948**). "Acoustic phonetics," Lang. Suppl. **24**, 1–136.

Jusczyk, P. W., Pisoni, D. B., and Mullennix, J. (**1989**). "Effects of talker variability on speech perception by 2-month-old infants," Res. Speech Percept. **15**, 133–161.

Kato, K., and Kakehi, K. (**1988**). "Listener adaptability to individual speaker differences in monosyllabic speech perception," J. Acoust. Soc. Jpn. **44**, 180–186.

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

Liberman, A. M., Cooper, F. S., Shankweiler, D. P., and Studdert-Kennedy, M. (**1967**). "Perception of the speech code," Psych. Rev. **74**, 431–461.

Lulich, S. M., Morton, J. R., Arsikere, H., Sommers, M. S., Leung, G. K. F., and Alwan, A. (**2012**). "Subglottal resonances of adult male and female native speakers of American English," J. Acoust. Soc. Am. **132**, 2592–2602.

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**, 391–409.

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

Miller, J. L., and Dexter, E. R. (**1988**). "Effects of speaking rate and lexical status on phonetic perception," J. Exp. Psychol. Human Percept. Perform. **14**, 369–378.

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**, 365–378.

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

Nordstrom, P. E., and Lindblom, B. (**1975**). "A normalization procedure for vowel formant data," paper presented at the *Proceedings of the 8th International Congress of Phonetic Science*, Leeds, UK.

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 Morin, T. M. (**1994**). "Paying attention to differences among talkers," in *The Development of Speech Perception: The Transition from Speech Sounds to Spoken Words*, edited by J. C. Goodman and H. C. Nusbaum (MIT Press, Cambridge, MA), pp. 113–134.

Nygaard, L. C., and Pisoni, D. B. (**1998**). "Talker-specific learning in speech perception," Percept. Psychophys. **60**, 355–376.

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

Pisoni, D. B. (**1996**). "Some thoughts on 'normalization' in speech perception," in *Talker Variability in Speech Processing*, edited by K. Johnson and J. W. Mullennix (Academic Press, San Diego).

Pisoni, D. B., and Remez, R. E. (**2008**). *The Handbook of Speech Perception* (Blackwell, Malden, MA), pp. 311–390.

Sjerps, M. J., Mitterer, H., and McQueen, J. M. (**2011**). "Constraints on the processes responsible for the extrinsic normalization of vowels," Atten. Percept. Psycho. **73**, 1195–1215.

Sommers, M. S. (**1997**). "Stimulus variability and spoken word recognition. II. The effects of age and hearing impairment," J. Acoust. Soc. Am. **101**, 2278–2288.

Sommers, M. S. (**1998**). "Spoken word recognition in individuals with dementia of Alzheimer's type: Changes in talker normalization and lexical discrimination," Psychol. Aging **13**, 631–646.

Sommers, M. S., Humes, L. E., and Pisoni, D. B. (**1995**). "The effects of speaking rate and stimulus variability on spoken word recognition by young and elderly listeners," Res. Spoken Lang. Process. **19**, 91–100.

Strand, E. A. (**1999**). "Uncovering the role of gender stereotypes in speech perception," J. Lang. Soc. Psych. **18**, 86–100.

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**, 1086–1100.

Traunmüller, H. (**1981**). "Perceptual dimension of openness in vowels," J. Acoust. Soc. Am. **69**(5), 1465–1475.

Verbrugge, R. R., Strange, W., Shankweiler, D. P., and Edman, T. R. (**1976**). "What information enables a listener to map a talker's vowel space?," J. Acoust. Soc. Am. **60**, 198–221.

Wayland, S. C., Miller, J. L., and Volaitis, L. E. (**1994**). "The influence of sentential speaking rate on the internal structure of phonetic categories," J. Acoust. Soc. Am. **95**, 2694–2701.

Whalen, D. H., Hoequist, C. E., and Sheffert, S. M. (**1995**). "The effects of breath sounds on the perception of synthetic speech," J. Acoust. Soc. Am. **97**, 3147–3153.

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