On the perception of similarity among talkers

Robert E. Remez[a)]
Department of Psychology, Barnard College, 3009 Broadway, New York, New York 10027

Jennifer M. Fellowes
Department of Psychiatry, New York Presbyterian Hospital, 180 Ft. Washington Avenue, New York, New York 10032

Dalia S. Nagel
Department of Ophthalmology, Mount Sinai School of Medicine, One Gustave L. Levy Place, New York, New York 10029

(Received 14 November 2006; revised 24 September 2007; accepted 25 September 2007)

A listener who recognizes a talker notices characteristic attributes of the talker's speech despite the novelty of each utterance. Accounts of talker perception have often presumed that consistent aspects of an individual's speech, termed *indexical* properties, are ascribable to a talker's unique anatomy or consistent vocal posture distinct from acoustic correlates of phonetic contrasts. Accordingly, the perception of a talker is acknowledged to occur independently of the perception of a linguistic message. Alternatively, some studies suggest that attention to attributes of a talker includes indexical linguistic attributes conveyed in the articulation of consonants and vowels. This investigation sought direct evidence of attention to phonetic attributes of speech in perceiving talkers. Natural samples and sinewave replicas derived from them were used in three experiments assessing the perceptual properties of natural and sine-wave sentences; of temporally veridical and reversed natural and sine-wave sentences; and of an acoustic correlate of vocal tract scale to judgments of sine-wave talker similarity. The results revealed that the subjective similarity of individual talkers is preserved in the absence of natural vocal quality; and that local phonetic segmental attributes as well as global characteristics of speech can be exploited when listeners notice characteristics of talkers.
 [DOI: 10.1121/1.2799903]

PACS number(s): 43.71.Bp, 43.71.Sy, 43.71.An [MSS] Pages: 3688–3696

I. ON THE PERCEPTION OF SIMILARITY AMONG TALKERS

The perception of characteristic attributes of an individual talker is customarily considered to occur independently of the resolution of the linguistic properties of utterances (Abercrombie, 1967; Bricker and Pruzansky, 1976; Halle, 1985). Conceptualizations of linguistic perception have typically appealed to a function by which a perceiver discovers linguistic form despite the variation in correspondence between a linguistic property and its acoustic correlates, whether the variation is attributed to coarticulation of phonetic segments or to affective or anatomical properties of a talker. Concurrently, the perception of a talker from a speech sample is said to be a kind of second message carried by an utterance, and perception of talker-specific attributes to depend on the resolution of *indexical* acoustic properties common to all of a talker's utterances, special to none. Classic approaches to indexical perception have offered evidence of perceptual sensitivity to acoustic correlates of size variation in vocal anatomy across talkers (Fant, 1966; Ladefoged and Broadbent, 1957; reviewed by Kreiman, 1997; and Pisoni, 1997). Others have appealed to persistent characteristics of use rather than a direct effect of anatomy as a reliable cause of long-term qualitative consistency in a speaker's speech, such as a tendency to whispery voice, or heavy nasality (Laver, 1980; Nolan, 1983). Research on qualitative variation of voices similarly has implicated basic auditory attributes, such as the pitch and pitch range of phonatory frequency, the timbre of the voice, as well as more abstract attributes, such as vocal strength or melodiousness, in the identification of talkers by ear (for example, Krauss *et al.*, 2002).

Recent refinements of this approach have included specification of the acoustic attributes unique to female voices (Klatt and Klatt, 1990) independent of linguistic contrasts, and estimation of the effects of variation in vowel spectra on phoneme quality (Frieda *et al.*, 1999) independent of specific individual talkers. A dissociation of linguistic and indexical perception in the effects of brain activity and injury (Belin *et al.*, 2004; Neuner and Schweinberger, 2000; Stevens, 2004; Van Lancker *et al.*, 1988) has also granted a biological license to the speculation that linguistic and indexical perception are distinct functions fed by different sensory attributes: Short-term elements pertain to symbolic contrasts, and long-term characteristics pertain to distinctions among talkers.

A. Identifying a sine-wave talker

Despite evidence of this dissociation of linguistic and indexical perceptual functions, each devoted to its kind of acoustic correlate and each promoting the perception of dif-

[a)]Electronic mail: remez@columbia.edu

ferent attributes of speech, some findings oppose the hypothetical independence of linguistic perceptual analysis and the perception of individual characteristics. Common among these is the premise that some grammatically regulated subphonemic properties of phonetic expression are nested within dialect and idiolect, and if perceivers notice, track and remember these linguistic attributes they can serve as indexical properties even though they do not stem from anatomical or physiological differences among talkers. Such findings have shown that perceivers require a phonetically diverse sample in order to resolve some distinctive aspects of individual talkers (Pollack *et al.*, 1954), and that familiarity with a specific talker promotes linguistic resolution of novel utterances (Lieberman, 1963; Nygaard *et al.*, 1994; Smith, 2004; also, see Hawkins, 2003; and a review by Pardo and Remez, 2006). In some tests of this approach to individual and linguistic identification, sine-wave replicas of natural speech (Fellowes *et al.*, 1997; Remez *et al.*, 1997; Sheffert *et al.*, 2002; see, also, Brungart *et al.*, 2006) were used to assess the conditions in which linguistically regulated attributes of speech promote the perception of talkers as well as words. These projects revealed that both strangers and long-time acquaintances could identify talkers under listening conditions in which the acoustic correlates of natural voice quality were eliminated from the speech samples. The use of sine-wave replicas of natural samples permitted such tests, in which a tone was set equal in frequency and amplitude to each of the three lowest oral resonances. A three-tone replica of a natural utterance fails to evoke an impression of natural vocal timbre, though it preserves the spectrotemporal variation sufficient to elicit impressions of detailed phonetic characteristics of consonants and vowels. Because sine-wave replicas of speech spectra are produced without setting a tone analog to match the fundamental frequency of the natural model, listeners in these tests identified talkers without relying on veridical impressions of vocal pitch (Remez and Rubin, 1984, 1993), although this attribute of speech has been favored as a perceptually and forensically useful acoustic correlate of individual differences (Bricker and Pruzansky, 1976; Hollien, 2002; von Dommelen, 1987). Indeed, Fellowes *et al.* (1997) eliminated the acoustic correlates of vocal tract scale variation as well, by transposing the tone analogs of the formants in frequency. Although this condition imposed an acoustic constraint impairing the perceptual identification of a talker's sex, listeners were able to identify most individuals nonetheless (Fellowes *et al.*, 1997). These findings show that many of the acoustic correlates of voice quality that had been candidates for use as indexical information are unnecessary for individual identification by listeners.

Overall, this pattern of results encouraged the hypothesis that listeners can exploit phonetic attributes in remembering and identifying individual talkers. Because habits of phonetic expression are presumably consistent in the speech of an individual and are distinct among individuals whether by dialect or idiolect, an intelligible utterance, even one that is distorted or anomalous in timbre, presents a potential for distinguishing talkers perceptually (Remez, in press). Of course, linguistically regulated subphonemic phonetic attributes are also useful in recognizing spoken words (Luce *et al.*, 2000).

This claim that talkers are identifiable by idiolect without access to qualitative attributes of the voice was based on indirect and opportunistic evidence. Performance levels in tests of individual identification from sine-wave samples were high, especially when listeners were personally acquainted with the talkers to be identified (Remez *et al.*, 1997; cf. Sheffert *et al.*, 2002). The distribution of errors of identification was used to estimate the perceptual similarity among the ten individuals whose utterances composed the set of test materials. The basic finding can be formulated succinctly: No single acoustic feature or dimension of variation of the sine-wave samples matched the perceptual similarity estimated from errors of identification. Neither similarity in duration, nor in central spectral tendency, nor in the spectral distribution of stressed vowels, nor in the rate of syllable production explained the similarity analyses based on misidentification of sine-wave talkers.

Although this pattern of results encouraged it, the hypothesis of an idiolectal cause for the patterns of similarity was an especially risky speculation because it was proposed on the evidence of very few data. Errors were rare given such good performance. The present project sought to improve the quality of this evidence about the causes of subjective similarity of talkers without natural vocal timbre. The three experiments of this report use direct assessments of subjective similarity and a variety of utterance types in tests that aimed to calibrate the perceived similarity of talkers without impressions of natural vocal quality.

B. The present tests

Three experiments investigated the apparent similarity among a set of talkers, in an attempt to explain prior findings about the identification of talkers from sine-wave samples. In Experiment 1, simple and direct reports of similarity were used to estimate the apparent similarity among the set of ten talkers used in the empirical precedents (Fellowes *et al.*, 1997; Remez *et al.*, 1997; Sheffert *et al.*, 2002) with two kinds of acoustic signals, natural samples and sine-wave replicas, and two different sentences. The finding of this study is that the pattern of perceptual similarity among a group of talkers is largely preserved over an acoustic transformation from natural sample to sine-wave signal. In Experiment 2, the contribution of vocal quality and phonetic inventory was estimated in tests that used temporally reversed natural and sine-wave samples. The results of this study showed that the loss of veridical phonetic properties affected impressions of similarity of natural and sine-wave samples alike, though a core of subjective similarity remained. Experiment 3 aimed to ascertain the role of the tone analog of the first formant, an index of the anatomical scale of a talker (Goldstein, 1980), in the perception of similarity among sine-wave talkers. These results showed that this single component of the tone complex of a sine-wave sentence does contribute to impressions of similarity of sine-wave talkers, as if it is used perceptually to index consistent qualitative differences among them.

TABLE I. Mean frequency and dispersion measures of the three lowest vocalic formants of the sentences used as natural samples or as models for sine-wave synthesis. Ten talkers each contributed two sentences. Male talkers are designated by M, female by F. The two sentences are: Yell: The drowning man let out a yell. Scarves: The scarves were made of shiny silk.

		First formant				Second formant				Third formant			
Talker	Sentence	Mean	s.d.	High	Low	Mean	s.d.	High	Low	Mean	s.d.	High	Low
M_1	Yell	480.1	133.4	693	280	1496.3	336.0	2330	914	2438.9	233.0	2817	1651
	Scarves	486.8	98.5	607	277	1428.8	289.6	2013	937	2186.4	233.2	2482	1666
M_2	Yell	488.1	175.1	784	192	1690.1	362.0	2352	961	2623.2	204.3	3165	2085
	Scarves	475.4	123.9	635	244	1570.9	415.2	2377	732	2394.5	405.6	3061	1726
M_3	Yell	488.9	170.5	783	160	1619.1	323.6	2269	1004	2410.8	206.9	2831	2068
	Scarves	434.4	119.7	608	197	1454.3	312.5	2023	871	2188.2	185.3	2746	1907
M_4	Yell	493.8	153.8	802	203	1583.6	350.9	2376	910	2580.5	294.9	3101	1620
	Scarves	500.1	128.7	706	305	1453.0	349.3	2220	822	2297.9	377.7	2830	1360
M_5	Yell	450.5	141.4	707	168	1390.1	278.0	2087	858	2277.5	256.9	2760	1430
	Scarves	439.6	112.3	616	264	1372.7	297.9	1848	510	2162.3	272.7	2517	1549
F_1	Yell	511.0	218.0	945	259	1926.0	414.7	2720	1243	3017.1	337.8	3549	1878
	Scarves	485.7	121.1	678	271	1750.3	432.6	2655	1162	2734.7	415.8	3197	1802
F_2	Yell	488.9	166.5	864	263	1939.9	399.2	2650	1165	2906.8	230.8	3458	2199
	Scarves	648.3	111.6	932	424	1764.5	407.6	2760	1216	2841.7	287.8	3281	2203
F_3	Yell	634.0	220.1	1024	245	1734.6	392.1	2650	969	2663.6	342.4	3442	1584
	Scarves	633.5	101.9	798	361	1622.6	484.9	2566	855	2431.8	361.1	2984	1787
F_4	Yell	602.3	203.6	996	187	1995.4	479.9	2875	1240	2976.1	241.3	3458	2443
	Scarves	662.2	101.7	803	381	1814.1	361.1	2590	1084	2773.5	266.9	3132	2188
F_5	Yell	504.7	159.1	794	281	1820.8	294.5	2450	1225	2982.7	309.9	3774	2218
	Scarves	634.9	146.2	812	375	1722.1	412.9	2520	1104	2766.3	127.2	3062	2458

Nonetheless, a test of partial correlation showed that the similarity of sine-wave talkers cannot be attributed solely to impressions of vocal pitch or scale evoked by the tone analog of the first formant. Taken together, the results indicate a role of global acoustic properties of speech as well as local phonetic segmental attributes in the perceptual similarity observed among these talkers.

II. EXPERIMENT 1

A. Simple reports of similarity

1. Method

a. Test materials. This test used natural samples of spoken sentences and sine-wave replicas modeled on them. The natural samples of two sentences, "The drowning man let out a yell," and "The scarves were made of shiny silk," were spoken within a list of sentences by each of ten talkers, five males and five females. The talker ensemble was relatively heterogeneous, representing different American English and British English dialects. Each talker practiced speaking the sentences in the list, and the test used samples of utterances that were produced comfortably and without dysfluency. The sentences were recorded in a sound-attenuating chamber, low-pass filtered at 4.5 kHz, digitally sampled at 10 kHz, equated for amplitude, and stored as sampled data with 12-bit amplitude resolution.

Sine-wave replicas of the 20 natural speech tokens were created by estimating the frequencies and amplitudes at 5-ms intervals of the three oral formants and the intermittent nasal and fricative formants, relying interactively on two representations of the spectrum: (1) Linear predictive coding and (2) the discrete Fourier transform. The derivation of these estimates was performed by hand, with multiple passes for correction of errors. Three time-varying sinusoids were then synthesized from these estimates to replicate the oral and nasal formant pattern, and a fourth sinusoid to replicate an intermittent fricative pole, based on the center frequencies and amplitudes obtained in the acoustic analysis (Rubin, 1980; two sine-wave synthesizers are available at http://www.haskins.yale.edu/featured/sws/information.html). (Table I presents a summary of the measures of each talker and sentence.) The sine-wave synthesis procedure preserved patterns of spectrotemporal change of the vocal resonances, while eliminating the fundamental frequency, harmonic relations, and fine-grained spectral details of natural speech. Subjectively, sine-wave sentences are intelligible though unnatural in vocal quality (Remez *et al.*, 1981, 2002). The 20 natural samples and 20 sine-wave sentences were stored in sampled data format; they were sequenced for use in testing and converted from digital records to analog signals delivered to each listener at a nominal level of 65 dB SPL over a calibrated Telephonics TDH-39 headset.

b. Procedure. Each volunteer listener participated in a single test appraising the likeness of the ten talkers in the set. There were four tests overall, two natural (Natural Yell and Natural Scarves) and two sine-wave (Sine-wave Yell and Sine-wave Scarves). Participation was blocked by test, and each listener was assigned randomly to a single acoustic type and sentence. Within each test, the trial structure was identical. A pair of sentences each produced by a different talker of the ten was presented separated by 1 s of silence. The listener then indicated the subjective likeness of the two talkers, using a five-point scale. Along a Similarity scale, the listener reported that the two talkers on the trial were *very similar* (a value of 1) or *not very similar* (a value of 5). Along a Dissimilarity scale, the listener reported that the two talkers on the trial were *very dissimilar* (a value of 1) or *not very dissimilar* (a value of 5). A participant used a single

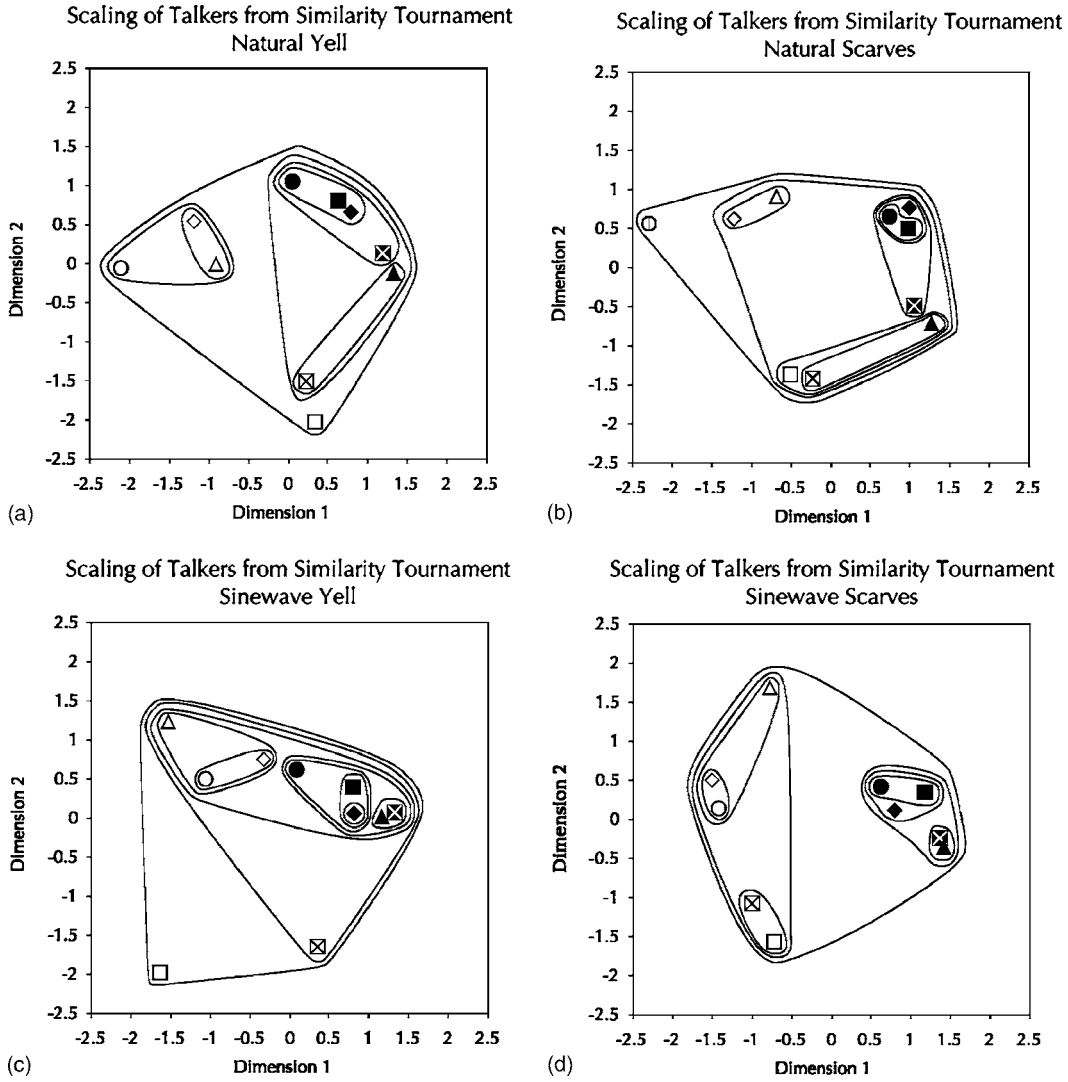

FIG. 1. Multidimensional scaling and hierarchical clustering analysis of the group data for four tests of similarity in Experiment 1: (a) With natural samples of the sentence, "The drowning man let out a yell." (b) With natural samples of the sentence, "The scarves were made of shiny silk." (c) With sine-wave replicas of "The drowning man…" (d) With sine-wave replicas of "The scarves were made…," Tests of hypotheses use ranked similarities derived from the hierarchical clustering analyses. In each panel, the placement of the bullets represents the solution of the scaling analysis; the placement of the enclosing curves represents the solution of the clustering analysis. Black bullets are male talkers; white bullets are female talkers.

scale, Similarity or Dissimilarity, for the entire procedure. There were 3 s of silence between successive trials, and 6 s of silence after every tenth trial.

To compose a sequence in which each of the ten talkers was presented with every other required 90 trials, 45 in each order. Each of these 90 trials occurred twice in a test session, composing a procedure of 180 trials, overall.

c. Subjects. One hundred and eleven students registered in the Barnard College Subject Pool participated in the listening tests of Experiment 1. Each was a right-handed native speaker of English reporting no history of language disorder. None was familiar with sine-wave replicas of speech. Course credit was granted for participation.

B. Results and discussion

With so few reports per subject, the measures were pooled within condition, and statistical analyses were performed solely on the group performance. The reports of three participants who failed to follow instructions were excluded from the data set.

Across the four conditions, the data treatment was the same. The results of the tests using a Dissimilarity scale (1=*very dissimilar*, 5=*not very dissimilar*) did not differ from those of the tests using a Similarity scale, and were reflected to coincide with the Similarity scale, in which the more dissimilar a pair was judged to be, the greater the value of the index. A square matrix was created for each of the four test conditions, Natural Yell, Natural Scarves, Sine-wave Yell, and Sine-wave Scarves, representing the accumulated reports of all 27 subjects in a group for each of the 90 ordered pairs of talkers. Each group matrix was analyzed using the technique of multidimensional scaling (Kruskal, 1964) to produce a two-dimensional representation of each of the four similarity tournaments. The solutions of the scaling analyses are shown in the four panels of Fig. 1.

Each group matrix was also used to derive a hierarchical clustering (Johnson, 1967) of the ten talkers in each condition, producing a likeness classification implicit in the similarity reports of the subjects. The dendrogram of the cluster

analysis was used to produce a similarity ranking in each of the four tests, for testing the principal hypotheses of this study.

Overall, the subjects produced consistent and differentiated reports of similarity, as shown by the indices of variance accounted for in the scaling solutions: For Natural Yell, $r^2=0.76$; for Natural Silk, $r^2=0.78$; for Sine-wave Yell, $r^2=0.87$; for Sine-wave Silk, $r^2=0.96$. To determine the role of natural vocal timbre in the assessment of similarity, we took the performance of the natural version of each sentence as a standard, and compared it to the sine-wave version. This was achieved by testing for correlated similarity ranks obtained in the cluster analyses for natural and sine-wave variants of Yell (Spearman's $r=0.85$, $p<0.01$) and Silk (Spearman's $r=0.92$, $p<0.01$). This outcome shows that listeners largely judged the ten talkers to be similar in the same way whether their judgments were concurrent with impressions of natural vocal quality or not.

Evidence of an influence of vocal timbre on the judgment of similarity among the talkers would be seen in a differential contribution of the phonetic sample to natural and sine-wave signals. To estimate this, we compared the performance of the two natural sentences to each other (Spearman's $r=0.92$, $p<0.01$) and the two sine-wave sentences to each other (Spearman's $r=0.76$, $p<0.01$). The stronger correlation of the natural sentences is arguably an indication that some talker attributes driving perceptual similarity in natural samples are preserved despite a difference in phonetic inventory distinguishing the sentences Yell and Scarves. Sine-wave sentences do not evoke impressions of natural vocal quality, in contrast to natural utterances. Accordingly, we expected to find a smaller correlation between the two sine-wave sentences, an effect of their different segmental inventory, than between the two natural sentences, which share qualitative attributes of vocal sound production, however the segmental inventory varies.

Two attributes of each talker that are plausibly preserved over variation in phonetic inventory are the vocal quality and the pitch height and range of phonation. Neither is present in sine-wave sentences, and the listener is obliged to resolve a talker's characteristics by attending to other attributes. The phonetic samples differ in each sentence, and because variation in phonetic samples is known to influence assessments of a talker's identity (Lieberman, 1963; Pollack *et al.*, 1954) it is reasonable to expect the similarity classification to follow the variation in the phonetic sample in the absence of other sustaining attributes. Nonetheless, this relative difference between natural and sine-wave similarity does not negate the clear indication that listeners were consistent in their overall performance in natural and sine-wave conditions. Whether this can be attributed to consistency in individual expression of segmental phonetic properties or to other more global aspects of vocal sound production (Laver, 1980) can only be resolved by additional tests.

In order to determine the contribution of qualitative attributes and of phonetic form to the natural and sine-wave performance in Experiment 1, Experiment 2 used a test with temporally reversed natural samples of Yell of each acoustic type. Although temporal reversal abolishes lexical access and harshly distorts the apprehension of many consonants, it preserves vowels at the syllable nuclei, and retains the glottal spectrum associated with impressions of vocal timbre along with the range of frequency variation of phonation. To the extent that assessments of similarity depend on these attributes alone, the performance for temporally veridical and reversed natural samples should not differ (cf. Van Lancker *et al.*, 1985). In the case of sine-wave replicas of speech, though, performance would be especially affected by temporal reversal if the appraisal of similarity rests on the phonetic details disrupted by reversal.

III. EXPERIMENT 2

A. Similarity of temporally reversed signals

1. Method

a. Test materials. Temporally reversed versions of the ten natural and ten sine-wave sentences in the Natural Yell and Sine-wave Yell sets were created by inverting the order of the digital records composing the files used in the temporally veridical conditions of Experiment 1. The result of this method applied to natural samples produced test items that preserved the glottal period and the spectrum variation of the natural and sine-wave sentences, arguably preserving quality while abolishing lexical access and perceptual resolution of many of the segmental phonetic attributes. The reversed sine-wave sentences were no less natural in vocal quality than the temporally veridical versions. However, temporal inversion affected the time-critical acoustic correlates of many consonants and gliding vowels, while sparing the tonal replicas of quasi-steady-state vowels at syllable nuclei and consonants that evolve more slowly, such as those of nasal and fricative manner.

b. Procedure. Two similarity tournaments were prepared from the reversed sentences, Reversed Natural Yell and Reversed Sine-wave Yell. Participation was blocked by test, and each listener was assigned randomly to a single acoustic type. Within each test, the structure of the trials and of the tests paralleled the procedure of Experiment 1.

c. Subjects. Thirty-five volunteers registered in the Barnard College Subject Pool participated in the listening tests of Experiment 2. Each was a right-handed native speaker of English reporting no clinical history of language disorder. No volunteer had participated in Experiment 1, nor was any familiar with sine-wave replicas of speech. Course credit was granted for participation.

B. Results and discussion

A single participant was excluded from the group data for declining to complete the test session. This left 17 subjects in each group. The individual reports were pooled within condition, and statistical analyses were performed solely on the group performance, repeating the practice of the first experiment of this report.

Consistent with the observations in Experiment 1, the judgments of similarity in these two tests were consistent and differentiated, and the multidimensional scaling solutions were again effective in two dimensions (for Reversed Natural Yell, $r^2=0.93$; for Reversed Sine-wave Yell, $r^2=0.80$). To test the hypothesis that a portion of the similarity relations among the natural talkers was based on phonetic properties,

we compared the ranked similarities of the Natural Yell performance of the first experiment (derived from a hierarchical clustering analysis) with the ranked similarities of the Reversed Natural Yell items assessed in this procedure. The performance based on reversed samples was highly correlated with performance in the first experiment based on veridical samples (Spearman's $r=0.78$, $p<0.01$), indicating an effect of qualitative attributes evoked by natural speech samples in judgments of perceptual similarity. However, this picture of perceptual similarity is complicated by the correlated performance that was observed for temporally veridical Sine-wave Yell signals in Experiment 1 and the temporally reversed signals in this procedure (Spearman's $r=0.72$, $p<0.01$). A principle stating that the perceived similarity of sine-wave talkers derives from the accumulation of phonetic segmental details cannot explain the parallel effect of temporal reversal on apparent similarity with and without natural vocal quality.

If temporal reversal preserves timbre at the expense of time-critical phonetic segments, then the effects that were observed here on talker perception with sinewave sentences warrant a similar explanation, at least initially. Perhaps listeners assessed talker similarity by attending to qualitative and phonetic attributes of the samples, natural and sine-wave alike. Although sine-wave replicas of speech do not conserve natural vocal timbre, the time-varying tones differ in average frequency and in frequency excursion over the test set. A listener accustomed to relying on multiple attributes—some long term and others short term—to appraise the similarity of talkers might do the same with the sine-wave cases as the natural cases, in contrast to our claim that phonetic attributes govern the perception of talkers when veridical qualitative attributes are unavailable (Fellowes *et al.* 1997; Remez *et al.*, 1997; Sheffert *et al.*, 2002). Although sine-wave replicas lack natural vocal quality, the frequency differences in tone components might evoke consistent differences in impressions of auditory form, affecting perceived similarity.

Empirical precedents found that variation in qualitative attributes was not required for perceptual identification of talkers from sine-wave replicas. For one, Fellowes *et al.* (1997) had demonstrated the identification of sine-wave talkers when the tone components had all been transposed in frequency to exhibit the same average values across the set, and with such items there can be little qualitative difference among the talkers, leaving phonetic properties as the likely grain for differentiation and identification of individuals. For another, Brungart *et al.* (2006) found that sine-wave sentences exhibit sufficient talker-specific characteristics to resist some effects of masking by a concurrent speech signal. In the present test, the sine-wave items replicated the natural frequency values estimated for vocal resonance, and across the test set each sentence evoked a potentially unique spectrum pitch despite the absence of a fundamental frequency, harmonic excitation, and broadband formants. Indeed, in a circumstance similar to this experiment, listeners reported an apparent vocal pitch of sine-wave sentences despite the absence of the familiar auditory excitation correlated with glottal pulsing. Tests revealed that the tone analog of the first formant supplied the acoustic correlate of sine-wave sentence intonation (Remez and Rubin, 1984, 1993). This attribute of a sine-wave replica survives temporal reversal, and could be responsible in part for the finding of similar perceptual scaling of temporally veridical and reversed sine-wave sentences in the absence of natural vocal quality.

The third experiment of this series aimed to account for the consistency of sine-wave talker similarity by testing the role of the tone analog of the first formant, all other things being equal. In this study, a similarity tournament was conducted using this single constituent of the sine-wave replica of each of the ten talkers. These single time-varying sinusoids were not intelligible, and did not evoke a perceptual impression of vocalization (cf. Remez *et al.*, 1981). Nonetheless, in intelligible sine-wave utterances, this tone component plays a complex role, as the surrogate for a vocal resonance associated with consonant manner and voicing and with the height of vowels. This tone analog of the first formant also is responsible for impressions of the weird intonation that accompanies sine-wave utterances (Remez and Rubin, 1984, 1993). As a replica of the first formant, it is arguably an acoustic marker of variation in the pharyngeal cavity, which differs in scale among talkers (Goldstein, 1980; Stevens, 2004). One way to determine whether the auditory form of this tone contributed a qualitative impression of a sine-wave talker in our prior tests is to compare similarity judgments of this isolated tone to judgments of intact sentences. However, because isolated tones do not evoke phonetic impressions, an auditory form similarity judgment was used as the task. The outcome of this test held the potential to explain the ability of listeners to treat sinewave replicas of utterances in two ways, as phonetic effects of idiolect of specific talkers, and as qualitative effects devoid of linguistic significance, much as the conventional view dichotomizing talker identification and phonetic perception warrants.

IV. EXPERIMENT 3

A. A test of the tone analog of the first formant

1. Method

a. Test materials. A set of new test items was developed from the ten sentences of the Sine-wave Yell set. This new set consisted of the tone analog of the first formant (T_1) of each of the ten sentences, synthesized as a single tone without concurrent sinusoids replicating the higher frequency formants. This was accomplished by taking the synthesis parameters for the tone analog of the first formant of each sentence replica and converting it to an analog signal composed of a single time-varying sinusoid. Each tone analog of a natural first formant realized the frequency and amplitude estimates of the original. Other aspects of preparation and delivery of acoustic test materials to listeners followed the practice of the first and second experiments.

b. Procedure. A single similarity tournament was used in this experiment, Sine-wave Yell T_1. Within each test, the structure of the trials and of the tests paralleled the procedure of Experiment 1, although the instructions described the test items as electronic melodies. We asked the listener to compare two electronic melodies presented in each trial and then to appraise their likeness. Continuing the plan of the prior studies, some subjects reported similarity and others reported dissimilarity.

c. Subjects. Nineteen students registered in the Barnard College Subject Pool participated in the listening tests of Experiment 3. No volunteer had been tested in any other condition of this set of experiments, and none had previously encountered sine-wave replicas of speech. Each was a right-handed native speaker of English reporting no history of language disorder. Course credit was granted for participation.

B. Results and discussion

The reports of two participants who did not complete the procedure were excluded from the compiled data, leaving 17 subjects in the group. The individual reports were pooled and statistical analyses were performed solely on the group performance.

Multidimensional scaling solutions were effective in two dimensions (Sine-wave Yell T_1, $r^2=0.89$). To test the hypothesis that the subjective similarity relations among sine-wave talkers reduces to pitch impressions or vocal tract scaling deriving from attention to T_1, the tone analog of the first formant, we compared the ranked similarities of the Sine-wave Yell T_1 performance (determined by a hierarchical clustering analysis) in Experiment 3 with the ranked similarities of the Sine-wave Yell items assessed in Experiment 1. The classifications were highly correlated (Spearman's $r=0.84$, $p<0.01$), encouraging an account of the similarity among sine-wave talkers, to a first approximation, as due to similarity of the lowest frequency tone components. This outcome was surprising, because an analysis of the physical properties of the samples (Remez *et al.*, 1997) had previously shown that the talkers were not well distinguished acoustically by the average frequency, or frequency range, or dispersion of frequency variation of the first formant. Prior ventures in similarity scaling of talkers based on error data (Fellowes *et al.*, 1997; Remez *et al.*, 1997) had also disconfirmed a reliance on criterial auditory properties of the samples to determine the perceived similarities. To explain the discrepancy of the precedents and present findings, it is useful to consider the possibility that the apparent similarity of two sine-wave talkers is influenced by the frequency variation in T_1, with additional attributes of the sine-wave samples contributing concurrently to the perception of the properties of talkers.

To estimate the extent to which the first formant and its tone analog determined the similarity scaling, we applied the findings of Experiments 1 and 3 in a test of partial correlation. This permitted us to assess the degree of agreement of similarity judgments evoked by natural and by sine-wave samples of the talkers while excluding the contribution of the tone analog of the lowest resonance, at least insofar as the perceptual tests in Experiment 3 calibrated it. A test of partial rank correlation based on the hierarchical clustering analyses found that the performance of the Natural Yell and the Sine-wave Yell sets remained correlated (Kendall's $\tau=0.31$) when the contribution of the tone analog of the first formant, as assessed in the Sine-wave Yell T_1 data, was neutralized.

This analytical strategy revealed that the congruent classification of sine-wave and natural talkers does not stem solely from the sensory appraisal of the first formant as a global property by which to scale a talker's vocal size. Instead, the finding of a substantial degree of correspondence without this factor exposes the room for other properties to influence the impressions of indexical attributes of individuals. Overall, Experiment 3 provides evidence of the relative contributions of global, qualitative factors and local, perhaps idiolectal factors in the perception of a talker from a speech sample. It seems fair to conclude that a variety of impressions determined the resolution of indexical attributes when a listener encountered a talker on a trial in our tests. Among these is an impression of scale evoked by global variation in the lowest resonance, as this test shows, as well as an impression of dialect or idiolect that is not reducible to the perception of vocal tract scale.

V. GENERAL DISCUSSION

How do listeners perceive similarity among talkers? This question was motivated by an unusual outcome in a set of tests of talker identification from sine-wave signals. Although it has become customary to refer to global acoustic properties of vocal sound production in explaining the attributes of speech on which talker identification is based, such an appeal was frustrated by the anomalous quality of sine-wave utterances. Lacking fundamental frequency, harmonic spectra, and broadband resonances, sine-wave signals are described as unnatural by listeners (Remez *et al.*, 1981) and are considered to be unnatural in quality even when they are intelligible (Remez *et al.*, 2002). The present findings from direct assessment of subjective similarity corroborate the estimates based on mistaken identity in studies by Fellowes *et al.* (1997) and by Remez *et al.* (1997), namely, that characteristic properties of a talker survive an acoustic transformation from natural sample to time-varying sinusoids. In the present project, Experiment 1 revealed that the perceived similarity among this set of talkers was much the same whether the judgments were based on natural samples or on sine-wave replicas. Experiments 2 and 3 together revealed that a combination of global and local properties is likely to contribute to the assessment of talker attributes when a listener encounters a sine-wave utterance.

The converging assessment of natural and sine-wave similarity performed in this study shows that a stable pattern of perceived similarity among sine-wave talkers is attributable to the robust properties of the talkers conveyed acoustically, and is not an artifact of the method employed by Remez *et al.* (1997) and Fellowes *et al.* (1997) to assess identification. These precedents directed a listener to choose which member of a pair of unidentified sine-wave samples had been spoken by a specific individual indicated by a printed name or exemplified in a brief natural sample. The prior findings encouraged the hypothesis that talkers are perceived from sine-wave patterns as if these synthetic utterances were speech samples with anomalous timbre—perhaps no more unusual than a sample of a familiar talker speaking with laryngitis—although the evidence to secure this conjecture was indirect. To test the hypothesis required perceptual measures with two critical controls.

The first control necessary to assess the claim that listeners perceived sine-wave talkers as if they were hearing the

natural models was a test of apparent similarity using both kinds of sample. In comparing the tests in these two acoustic conditions here, it was possible to determine that the pattern of performance manifest in each acoustic type resembled the pattern observed with the other. It appears from these tests that listeners do indeed express detailed sensitivity to properties of the speech of these individuals, and the pattern is highly correlated across the acoustic types.

The second control required to decide the claim was an experimental manipulation to test the perceptual role of a prominent global property, the frequency variation of the first formant in natural and in sine-wave samples. Our prior investigations had employed a statistical technique (cf. Walden *et al.*, 1978) in assessing the proximate causes of perceived similarity. This approach had proved fruitful in generating hypotheses about the phonetic properties available for indexical perception, chiefly because idiolectal attributes seemed far more distinctive and characteristic than did the global acoustic properties featured in the talker identification literature. The perceptual test was performed in Experiment 3 here. Although the result indicated a clear contribution of the tone analog of the first formant to a listener's perceptual appraisal of sine-wave talkers, the finding overall is consistent with the notion that indexical properties of talkers include linguistically governed properties like dialect and idiolect and rather more simple scalable properties like vocal tract size.

Limits to the generality of the findings. Although the tests reported here arguably expose the relative salience of phonetic details in perceptual appraisals of talker similarity, the findings show the availability of these attributes in the perception of individuals without defining the role that such properties play in ordinary listening. A test of direct magnitude estimation of the likeness of two speech samples in a controlled listening environment differs greatly from a challenge to identify a talker under ordinary circumstances. Indeed, the stability of recognition even in controlled tests can depend on the number and variety of other talkers in the test set (Bricker and Pruzansky, 1976; Hecker, 1971; Papcun *et al.*, 1989). New tests are required to determine whether the attributes of an individual's speech that are promoted to prominence by judgments of likeness have a legitimate function in perceptual identification. In this regard, it will also be necessary to calibrate the perceptual effect of aspects of speech that distinguish talkers by dialect and by idiolect (for instance, Remez *et al.*, 2004).

A second limit to consider derives from the kind of speech samples that were used in the present tests and in the immediate empirical precedents (Fellowes *et al.*, 1997; Remez *et al.*, 1997). All of the utterances were taken from sentence lists read aloud by the talkers. As such, the listeners in these tests heard speech produced in a normative register, which might have been less characteristic of the individual talkers than casually produced speech despite the instruction to speak comfortably. Additional tests will be required to determine the relation between the assortment of phonetic attributes in fluent reading and in spontaneous registers, indeed, among the differing formal and informal registers of spontaneous speech (for example, Labov, 1986).

This study took an experimental approach to the question of the perceptual effect of the tone analog of the first formant, and in comparison we must acknowledge that the opportunity to take an experimental approach to dialect and idiolect will be difficult. Certainly, the phonetic attributes at play among these ten talkers are not obscure: Some individuals spirantized the coronal stop releases while others released them with a clean, brief burst; some geminated an intervocalic stop hold (VCCV), while others more nearly approximated a VCV; some raised or diphthongized low vowels, while others more nearly produced a clear singleton. Our hypothesis is that, in aggregate, these phonetic aspects of an individual's speech contribute to individual identification when vocal quality is ordinary as well as anomalous. But, the problem of treating such properties as discriminanda in a factorial design requires identifying individual talkers whose natural dialect and idiolect vary in the manner required by a research method of parametric variation. New studies will determine the resolution with which empirical tactics can address this intriguing question about speech perception.

ACKNOWLEDGMENTS

The authors thank Philip E. Rubin, Michael Studdert-Kennedy, Jennifer Van Dyk, and Cynthia Y. Yang for their guidance and criticism of this project and report; and, we thank Floye Sumida, Robert Boruchowitz, and Harriet Greisser for advice and encouragement about the prospects of understanding this difficult and subtle problem. The research was sponsored by an award to Barnard College (DC00308) from the National Institute on Deafness and Other Communication Disorders.

Abercrombie, D. (**1967**). *Elements of General Phonetics* (Aldine, Chicago).
Belin, P., Fecteau, S., and Bédard, C. (**2004**). "Thinking the voice: Neural correlates of voice perception," Trends Cogn. Sci. **8**, 129–135.
Bricker, P. D., and Pruzansky, S. (**1976**). "Speaker recognition," in *Contemporary Issues in Experimental Phonetics*, edited by N. J. Lass, (Academic, New York), pp. 295–326.
Brungart, D. S., Iyer, N., and Simpson, B. D. (**2006**). "Monaural speech segregation using synthetic speech signals," J. Acoust. Soc. Am. **119**, 2327–2333.
Fant, C. G. M. (**1966**). "A note on vocal tract size factors and nonuniform *F*-pattern scalings," Speech Transmission Laboratory Quarterly Progress and Status Report 4, Royal Institute of Technology, Stockholm, Sweden, pp. 22–30.
Fellowes, J. M., Remez, R. E., and Rubin, P. E. (**1997**). "Perceiving the sex and identity of a talker without natural vocal timbre," Percept. Psychophys. **59**, 839–849.
Frieda, E., Walley, A., Flege, J., and Sloane, M. (**1999**). "Adults' perception of native and nonnative vowels: Implications for the perceptual magnet effect," Percept. Psychophys. **61**, 561–577.
Goldstein, U. G. (**1980**). "An articulatory model for the vocal tracts of growing children," Doctoral dissertation, Massachusetts Institute of Technology, Cambridge, MA.
Halle, M. (**1985**). "Speculations about the representation of words in memory," in *Phonetic Linguistics: Essays in Honor of Peter Ladefoged*, edited by V. A. Fromkin (Academic, New York), pp. 101–114.
Hawkins, S. (**2003**). "Roles and representations of systematic fine phonetic detail in speech understanding," J. Phonetics **31**, 373–405.
Hecker, M. H. L. (**1971**). "Speaker recognition: An interpretive survey of the literature," ASHA Monogr. **16**, 1–103.
Hollien, H. (**2002**). *Forensic Voice Identification* (Academic, San Diego).
Johnson, S. C. (**1967**). "Hierarchical clustering schemes," Psychometrika **32**, 241–254.
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**, 820–857.
Krauss, R. M., Freyberg, R., and Morsella, E. (**2002**). "Inferring speakers' physical attributes from their voice," J. Exp. Soc. Psychol. **38**, 618–625.
Kreiman, J. (**1997**). "Listening to voices: Theory and practice in voice perception research," in *Talker Variability in Speech Processing*, edited by K. Johnson and J. W. Mullennix (Academic, San Diego), pp. 85–108.
Kruskal, J. B. (**1964**). "Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis," Psychometrika **29**, 1–27.
Labov, W. (**1986**). "Sources of inherent variation in the speech process," in *Invariance and Variability in Speech Processes* edited by J. S. Perkell and D. H. Klatt (Erlbaum, Hillsdale, NJ), pp. 402–425.
Ladefoged, P., and Broadbent, D. E. (**1957**). "Information conveyed by vowels," J. Acoust. Soc. Am. **29**, 98–104.
Laver, J. (**1980**). *The Phonetic Description of Voice Quality* (Cambridge University Press, Cambridge, UK).
Lieberman, P. (**1963**). "Some effects of semantic and grammatical context on the production and perception of speech," Lang Speech **6**, 172–187.
Luce, P. A., Goldinger, S. D., Auer, E. T., Jr., and Vitevitch, M. S. (**2000**). "Phonetic priming, neighborhood activation, and PARSYN," Percept. Psychophys. **62**, 615–625.
Neuner, F., and Schweinberger, S. R. (**2000**). "Neuropsychological impairments in the recognition of faces, voices and personal names," Brain Cogn **44**, 342–366.
Nolan, F. (**1983**). *The Phonetic Bases of Speaker Recognition* (Cambridge University Press, Cambridge, UK).
Nygaard, L., Sommers, M., and Pisoni, D. (**1994**). "Speech perception as a talker-contingent process," Psychol. Sci. **5**, 42–46.
Pardo, J. S., and Remez, R. E. (**2006**). "The perception of speech," in *Handbook of Psycholinguistics*, 2nd Edition, edited by M. Traxler and M. A. Gernsbacher (Academic Press, San Diego), pp. 201–248.
Papcun, G., Kreiman, J., and Davis, A. (**1989**). "Long-term memory for unfamiliar voices," J. Acoust. Soc. Am. **85**, 913–925.
Pisoni, D. B. (**1997**). "Some thoughts on 'normalization' in speech perception," in *Talker Variability in Speech Processing*, edited by K. Johnson and J. W. Mullennix (Academic, San Diego), pp. 9–32.
Pollack, I., Pickett, J. M., and Sumby, W. H. (**1954**). "On the identification of speakers by voice," J. Acoust. Soc. Am. **26**, 403–406.
Remez, R. E., "Spoken expression of individual identity and the listener," in *Expressing Oneself/Expressing One's Self: A Festschrift in Honor of Robert M. Krauss*, edited by E. Morsella (Taylor & Francis, London), in press.
Remez, R. E., Fellowes, J. M., and Rubin, P. E. (**1997**). "Talker identification based on phonetic information," J. Exp. Psychol. Hum. Percept. Perform. **23**, 651–666.
Remez, R. E., and Rubin, P. E. (**1984**). "On the perception of intonation from sinusoidal sentences," Percept. Psychophys. **35**, 429–440.
Remez, R. E., and Rubin, P. E. (**1993**). "On the intonation of sinusoidal sentences: Contour and pitch height," J. Acoust. Soc. Am. **94**, 1983–1988.
Remez, R. E., Rubin, P. E., Pisoni, D. B., and Carrell, T. D. (**1981**). "Speech perception without traditional speech cues," Science **212**, 947–950.
Remez, R. E., Wissig, S. C., Ferro, D. F., Liberman, K., and Landau, C. (**2004**). "A search for listener differences in the perception of talker identity," J. Acoust. Soc. Am. **116**, 2544.
Remez, R. E., Yang, C. Y., Piorkowski, R. L., Wissig, S., Batchelder, A., and Nam, H. (**2002**). "The effect of variation in naturalness on phonetic perceptual identification," J. Acoust. Soc. Am. **111**, 2432.
Rubin, P. E. (**1980**). "Sinewave synthesis," Technical report, Haskins Laboratories, New Haven, CT.
Sheffert, S. M., Pisoni, D. B., Fellowes, J. M., and Remez, R. E. (**2002**). "Learning to recognize talkers from natural, sinewave and reversed speech samples," J. Exp. Psychol. Hum. Percept. Perform. **28**, 1447–1469.
Smith, R. (**2004**). "The role of fine phonetic detail in word segmentation," Doctoral dissertation, University of Cambridge, Cambridge, UK.
Stevens, A. A. (**2004**). "Dissociating the cortical basis of memory for voices, words and tones," Cognit. Brain Res. **18**, 162–171.
Van Lancker, D., Cummings, J. L., Kreiman, J., and Dobkin, B. H. (**1988**). "Phonagnosia: A dissociation between familiar and unfamiliar voices," Cortex **24**, 195–209.
Van Lancker, D., Kreiman, J., and Emmorey, K. (**1985**). "Familiar voice recognition: Patterns and parameters. I. Recognition of backward voices," J. Phonetics **13**, 19–38.
von Dommelen, W. A. (**1987**). "The contribution of speech rhythm and pitch to speaker recognition," Lang Speech **30**, 325–338.
Walden, B. E., Montgomery, A. A., Gibeily, G. J., Prosek, R. A., and Schwartz, D. M. (**1978**). "Correlates of psychological dimensions in talker similarity," J. Speech Hear. Res. **21**, 265–275.