Perceptual susceptibility to acoustic manipulations in speaker discrimination

Gregory Sell[a)]
Institute for Systems Research, Electrical and Computer Engineering Department, University of Maryland, College Park, Maryland 20742

Clara Suied
Institut de Recherche Biomédicale des Armées, Département Action et Cognition en Situation Opérationnelle, 91223 Brétigny sur Orge, France

Mounya Elhilali
Electrical Engineering Department, Johns Hopkins University, Baltimore, Maryland 21218

Shihab Shamma
Institute for Systems Research, Electrical and Computer Engineering Department, University of Maryland, College Park, Maryland 20742

(Received 14 October 2013; revised 11 September 2014; accepted 8 December 2014)

Listeners' ability to discriminate unfamiliar voices is often susceptible to the effects of manipulations of acoustic characteristics of the utterances. This vulnerability was quantified within a task in which participants determined if two utterances were spoken by the same or different speakers. Results of this task were analyzed in relation to a set of historical and novel parameters in order to hypothesize the role of those parameters in the decision process. Listener performance was first measured in a baseline task with unmodified stimuli, and then compared to responses with resynthesized stimuli under three conditions: (1) normalized mean-pitch; (2) normalized duration; and (3) normalized linear predictive coefficients (LPCs). The results of these experiments suggest that perceptual speaker discrimination is robust to acoustic changes, though mean-pitch and LPC modifications are more detrimental to a listener's ability to successfully identify same or different speaker pairings. However, this susceptibility was also found to be partially dependent on the specific speaker and utterances. © *2015 Acoustical Society of America*.
[http://dx.doi.org/10.1121/1.4906826]

[CYE] Pages: 911–922

I. INTRODUCTION

Human beings have a highly robust ability to recognize speakers based only on their voices in a wide variety of conditions. As a result, research into this process has a long history, dating back over 50 years, motivated at least in part by the belief that understanding how humans perform this task can help us better understand the reliability of the answers, increase the accuracy, and train computers to perform the same process on larger databases and for lower cost.

Numerous studies in the past have examined the ability to identify human speakers (Kreiman and Sidtis, 2011), though few have focused on specific acoustic metrics. The ability to directly control acoustic parameters was limited by available technology, and so, in some cases, only certain aspects could be tested, such as pitch range, duration of exposure, and voiced/non-voiced ratios (Pollack *et al.*, 1954), or frequency bandwidth (via filtering) and duration of exposure (Compton, 1963). In other cases, creative approaches were utilized to control aspects of the speaker's voice. For example, the absence of glottal source variation was examined by asking speakers to use an electronic larynx (Coleman, 1973), while using synthesized sine-wave speech (Remez *et al.*, 1997) only exposed listeners to the information contained in formant frequencies [this research was further examined (Remez *et al.*, 2007) with additional processing of the sine-wave speech]. Speech has also been played for listeners in reverse, distorting phonetic and temporal cues while preserving pitch and voice quality (Van Lancker *et al.*, 1985a).

Other studies have examined the role of higher level aspects such as phonetic content (Amino *et al.*, 2006; Amino and Arai, 2009), fluency in the spoken language (Thompson, 1987; Köster and Schiller, 1997; Maheshwari *et al.*, 2008; Perrachione *et al.*, 2011), or speaking rate (Van Lancker *et al.*, 1985b). Studies have also considered the effects of the listening conditions on speaker identification, such as duration of exposure (Compton, 1963; Bricker and Pruzansky, 1966), delay between exposure and identification (Kerstholt *et al.*, 2004), or familiarity with the unknown speaker (Van Lancker and Kreiman, 1987; Yarmey *et al.*, 2001).

It is also worth noting that the vast majority of modern automatic algorithms for speaker identification represent the incoming spoken signal with Mel-frequency cepstral coefficients (MFCCs), which encode the spectral envelope in the form of a lowpass-filtered power spectrum. While MFCCs

[a)]Author to whom correspondence should be addressed. Current address: Human Language Technology Center of Excellence, Johns Hopkins University, Baltimore, MD 21211. Electronic mail: gsell@jhu.edu

themselves have not been tested in a perceptual context, there is perceptual research to support the importance of spectral envelopes or related information (Coleman, 1973; Remez *et al.*, 1997; Gaudrain *et al.*, 2009; Amino and Arai, 2009).

Other research has looked into the perceptual cues utilized for decisions on speaker gender (Lass *et al.*, 1976; Bachorowski and Owren, 1999; Smith and Patterson, 2005; Skuk and Schweinberger, 2013), age (Hartman and Danhauer, 1976; Smith and Patterson, 2005), size (Smith and Patterson, 2005), or personality (Brown *et al.*, 1973, 1974), tasks that are potentially relevant to speaker recognition.

However, despite this rich history of research in perceptual speaker identification, the overall picture is still somewhat unclear. Past studies show that most, if not all, aspects of speech aid listeners to some degree in identifying an unknown speaker [for example, listeners have been successfully trained to recognize the same speakers with natural speech, sine-wave speech, or reversed speech (Sheffert *et al.*, 2002)], and that the magnitude of these effects is dependent on the speaker, the listener, and the listening conditions. It is with this continuing challenge in mind that the research presented here examined the effects of specific aspects of a speech signal through resynthesis of real recordings with only the selected feature modified. Through these manipulations of the acoustic characteristics of the signals, we were able to isolate the impact of removing specific cues on the ability of a listener to identify an unknown speaker.

In recent years, there have been a few examples of similar research directed at examining the effect of acoustic cues through direct manipulation of the signals.

One such study (Gaudrain *et al.*, 2009) examined the relationship between vocal tract length, glottal pulse rate, and speaker similarity by manipulating those parameters for consonant-vowel (CV) recordings from a single speaker. The experiment targeted the effects of these manipulations on speaker similarity rather than identity by asking participants to rate whether it was possible tokens with differing degrees of manipulation were uttered by the same speaker. The results showed that participants were more tolerant to glottal pulse rate differences than vocal tract length differences in assessing the similarity of the voices.

Another acoustics-focused study (Lavner *et al.*, 2000) manipulated the pitch, glottal waveform, and formants of recordings of the vowel /a/ then tested if participants could recognize familiar voices. Results showed that the recognizability of each voice was influenced differently by the manipulations, suggesting that the feature set utilized by listeners varies with the speaker.

Kuwabara and Takagi (1991) also used an analysis-synthesis method to manipulate formant frequencies, formant bandwidths, and fundamental pitch of two speakers uttering a nonsense word. The speakers of the manipulated utterances were then identified by three listeners who were familiar with both of the original speakers. The results suggested that formant shift is more closely tied to individuality than pitch or formant bandwidth.

Kreiman and Papcun (1991) also conducted a related study in which they extracted numerous metrics related to the pitch and formants of spoken stimuli and then correlated those values with a four-dimensional coordinate set derived from the perceptual results of a speaker discrimination task. Based on these correlations, the four dimensions were categorized as masculinity, creakiness, variability, and mood. Speaker subspaces had also been explored earlier as well (Voiers, 1964, 1979), though these subspaces were based on speaker ratings by trained listeners rather than discriminative decisions.

In the following article, we will discuss recent experiments designed to advance understanding in the acoustics of speaker discrimination. To accomplish these goals, we performed listening tests with human participants discriminating short stimuli consisting of spoken words. The first experiment provided a baseline with unmodified utterances, and the results of this experiment were used to analyze the role of a selected set of acoustic parameters, including several parameters that have not been previously considered in perceptual speaker identification (namely, MFCCs, spectro-temporal envelopes, and a new timbre metric called *raspiness*). On the basis of these analyses, we selected several featured parameters [mean pitch, phonetic duration, and linear predictive coefficients (LPCs)] for acoustic manipulation through resynthesis, and repeated the experiment with the modified stimuli, thus measuring the effect of the missing information on the listener's ability to identify unknown speakers. This experimental set-up shares some similarities with past work but also has important distinctions.

First, it is worth noting that we asked listeners to determine if two utterances were spoken by same or different speakers, as opposed to asking speakers to identify the speaker of a single utterance (Lavner *et al.*, 2000). That study also used speakers familiar to the listeners [as did Kuwabara and Takagi (1991)], while we utilized unfamiliar speakers. This distinction may seem nuanced, but familiar and unfamiliar speaker identification have been shown to be measurably different tasks (Yarmey *et al.*, 2001) that utilize different hemispheres of the brain (Van Lancker and Kreiman, 1987), and studies have not shown whether the features utilized by a listener are adjusted for familiar or unfamiliar speakers.

It is also important to note that our speech corpus was derived from speakers reading the same content in three separate sessions, which allowed us to pose listeners with separate utterances spoken by the same speaker that included identical phonetic content (and context) or utterances with completely different phonetic content. Using such a stimuli set helps control the role of phonetic content in the decision, which is important for examining acoustic parameters. A single vowel spoken in isolation was used by Lavner *et al.* (2000), while CV syllables spoken in isolation were used by Gaudrain *et al.* (2009), though in the experiment they were presented as two triplets with completely different syllables.

Results of our experiments showed that the modifications have a significant effect on listener accuracy, and that there were also significant effects within speakers and spoken words. Under these conditions, mean pitch and LPCs are

both important to perceptual speaker discrimination, but further analysis shows that the nature of these effects is different.

II. METHODS

A. Participants

Participants, recruited from the student population at the University of Maryland at College Park, MD, were fluent English speakers with no self-reported hearing impairments. Fifteen students participated in the baseline experiment (2 male and 13 female), ten in the mean-pitch experiment (all female), eight in the phonetic duration experiment (1 male and 7 female), and seven in the LPC experiment (all female). Across this entire set of participants, ages ranged from 18 to 23 (mean 20.2). The gender imbalance in this set was an unintentional product of the volunteer population.

Participants were permitted to take different experiments, though they were required to wait at least a full week between sessions.

All participants provided informed consent to participate in the study and were reimbursed for the hour-long experiment, in accordance with protocol 10–0411 approved by the University of Maryland Institutional Review Board.

B. Stimuli

The stimuli were selected from the *Mixer-6 Speech* corpus (Brandschain *et al.*, 2013), a large database collected and published at the Linguistic Data Consortium at the University of Pennsylvania. The database includes recorded interviews, telephone recordings, and clean read speech. For this experiment, only clean, read speech was used.

Samples were recorded over a single channel at a 16 kHz sampling rate. The database includes speakers reading the same transcript in three separate sessions on separate days. As a result, it was possible to create a database of the same speakers saying the same phrases on separate days.

To select the speakers, we started first by narrowing the group to male non-smokers from the Philadelphia region (arbitrary selections based on the demographics of the database). This group was narrowed to six speakers (ID 120346, 120863, 120552, 120664, 120749, 120537) through an informal experiment aimed at finding a set of voices that were separable but similar, since the discriminability of voices with drastic differences (such as male and female) is unlikely to be affected by adjusting a single acoustic parameter. For the remainder of this report, the six selected speakers will be referred to as SPK1, SPK2, etc.

For each of the six speakers, we isolated the word "know" from the sentence "I know; they are very frustrated after that," and the word "scam" from the sentence "It is a whole scam," resulting in a total of 36 stimuli (two words extracted from three different sessions for six total speakers).

These stimuli were used for the baseline experiment and were the basis for modifications for the follow-up experiments. The nature of these modifications will be explained in the respective experimental descriptions to follow.

It is worth mentioning that this database was used for the Speaker Recognition Evaluations SRE10, HASR10, SRE12, and HASR12, competitive evaluations in automatic speaker recognition. All six of these speakers used in the following experiments were included in the SRE10 evaluation, and SPK1 and SPK3 were identified as challenging enough to the automatic systems to also be included in the HASR10 subset (Greenberg *et al.*, 2011), which was comprised of only the most difficult trials. This is noteworthy because, to date, the automatic and perceptual research communities have stayed on largely separate tracks, and this data provides an opportunity to create a common set of results for better integration of research, both within the perceptual community and possibly even between the perceptual and automatic communities.

C. Protocol and apparatus

The experiment lasted approximately 1 h for each participant, and typically less. Participants were left alone in a sound booth for the duration of each experiment, though breaks were allowed if desired. The experiment utilized an iPad interface with audio playback through a Headamp Pico DAC/amplifier and Sennheiser HD600 headphones. Participants set their own playback volume level to a comfortable level and were permitted to adjust the level throughout the experiment.

In each trial, participants were asked to identify whether two recorded voices were uttered by the same speaker or different speakers, which is a speaker discrimination task. Participants initiated audio playback of the randomly selected stimuli by pushing one of two buttons (labeled "A" and "B"). Participants were permitted to listen to each stimulus any number of times and in whatever order desired. Participants were then able to select either "same" or "different" before proceeding to the next trial. Changing a response was permitted before moving to the next trial if desired, but participants could not skip or return to previous trials after moving on.

Participants trained for the task with a brief explanation followed by three sample trials under supervision using stimuli selected from outside the experiment database. Training occurred immediately prior to the experiment.

Participants were asked to respond for every pair of unique signals once (excluding comparisons of a signal to itself) which, for 36 stimuli, requires 630 total trials [$n(n-1)/2$, with $n=36$]. The total experiment broke down into 90 "same" trials and 540 "different trials." Furthermore, there were 153 trials each comparing only "know" utterances or only "scam" utterances and 324 trials comparing "know" utterances to "scam" utterances. Each speaker was presented in at least one of the utterances in 195 trials.

D. Statistical analysis

For each experiment, raw responses were plotted as a similarity matrix. For statistical analyses, an accuracy measure was derived based on the pair comparison responses of the participants. This accuracy (rate of correct responses) was computed for two conditions: Speaker and word-pairing. Speaker accuracy for a particular speaker is defined as the average accuracy on trials that involve that speaker. So, the

SPK1 accuracy measures the average listener accuracy on any trial involving SPK1 utterances compared to any speaker (including same trials comparing SPK1 to SPK1). Word-pairing accuracy separates trials into the three possible uttered word comparisons ("know"/"know," "scam"/"scam," or "know"/"scam") and measures the average accuracy for each condition.

For each experiment, a generalized linear mixed model was estimated, with speaker and word-pairing as fixed effects and listener as a random effect. For the binomial response data collected in the experiments, the model type is a logistic regression. Significance of effects was then assessed with Wald χ^2 tests.

III. BASELINE EXPERIMENT

A. Participants

Fifteen individuals (2 male and 13 female) participated in the baseline experiment.

B. Results

The full similarity matrix averaging participant responses is shown in Fig. 1. It corresponds to the aggregated raw data of pair comparisons between two stimuli.

The similarity matrix of perfect results would be block diagonal. Instead, there are several noteworthy errors visible in the similarity matrix. For example, the most frequently confused pair are SPK2 and SPK3, with SPK4 and SPK6 the second most confused.

Overall, participants performed the baseline task with 84.21% accuracy (defined as percent of trials answered correctly, regardless of speaker or word pairing). Figure 2 includes the participant accuracy for word-pairing and speaker conditions.

Statistical analysis found significant main effects of speaker [$\chi^2(5) = 29.86$; $p < 0.0001$] and word-pairing [$\chi^2(2) = 10.71$; $p < 0.005$]. The speaker main effect can be explained by a better recognition of SPK1 compared to every other speaker ($p < 0.01$). The word-pairing main effect was due to a difference between "scam"/"scam" and other word-pairings ($p < 0.05$).

No significant interaction was found [$\chi^2(10) = 14.17$; $p = 0.17$].

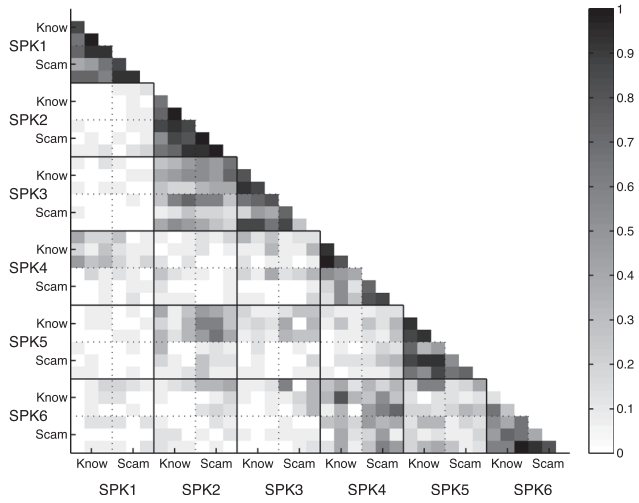

FIG. 1. Similarity matrix for the baseline experiment ("same" responses dark). Perfect accuracy would result in a block diagonal matrix.

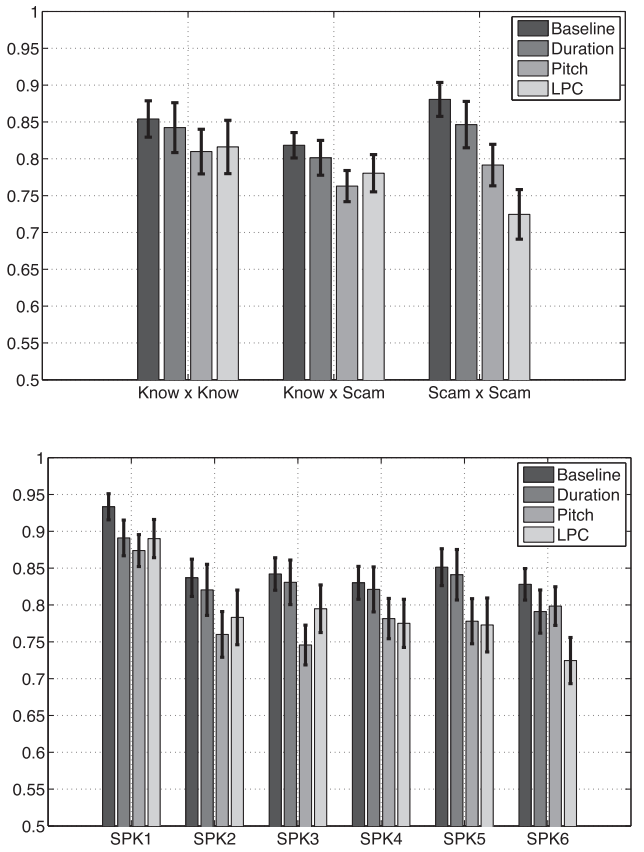

FIG. 2. Participant performance (a) for each word comparison and (b) for inclusion of each speaker for all experiments. Error bars show 95% confidence intervals.

C. Interim discussion

Overall, participants were very good at this speaker discrimination task. Interestingly, with the notable exception of SPK1, the average accuracies were very similar for all speakers. Also, while word-pairing had a measurable effect on accuracy, listeners were able to identify the necessary characteristics for speaker discrimination for all word-pairings, regardless of whether or not the utterances shared the same phonetic content.

IV. BASELINE EXPERIMENT PARAMETER ANALYSIS

In order to analyze the perceptual results in terms of the acoustic characteristics of the stimuli, we analyzed a series of parameters within the context of the perceptual results. Parameters calculated directly from the spoken speech were compared to both the perceptual results and the coordinates of a low-dimensional representation of the perceptual results. These comparisons show which parameters have the closest relationship to the behavior of the participants and, therefore, suggest themselves as potentially important acoustic cues.

It is important to note, though, that these comparisons only show that a relationship exists between acoustic

TABLE I. Parameters extracted from stimuli for analysis. The parameters are organized by whether the metric is single-dimensional or multi-dimensional, and whether it can be used as a meaningful comparison between all of the spoken words, or is only meaningful in comparing the same word.

Single dim, All	Single dim, Same words	Multi-dim, All	Multi-dim, Same words
Pitch mean	Duration	Cortical	MFCC mean
Pitch range	Formant means (F_1, F_2,…, F_5)		MFCC std
Pitch std	Formant differences (e.g., F_2-F_1)		LPC mean
Pitch slope	Formant ratios (e.g., F_2/F_1)		
Harmonic richness	Formant dispersion		
NHR			
Raspiness			
Spectral slope			

parameters and participant responses, but they are not sufficient to demonstrate causality or the extent to which the parameters are utilized by the participants.

A. Parameters

The full set of parameters examined is shown in Table I, separated into several categories. First, they are listed based on whether the parameter is relevant to all stimuli comparisons, or is only relevant in comparing the same word. For example, formants primarily encode phonetic information, and so comparing the formants of utterances of different words will be dominated by the phonetic content rather than speaker identity; they are more useful in this context for comparing different utterances of the same word, where formant variations indicate pronunciation or vocal tract differences. The parameters are also separated into single dimensional parameters and multi-dimensional parameters.

The parameter set combines those used in past research with more recently developed metrics.

(1) Pitch was estimated in Praat (Boersma, 2001), which estimates pitch by autocorrelation, and statistical metrics of pitch were calculated on linearly scaled hertz.
(2) Formants were also estimated in hertz with Praat (Boersma, 2001) using the Burg method.
(3) MFCCs and LPCs were calculated using the mfcc and proclpc functions, respectively, in the Auditory Toolbox (Slaney, 1998).
(4) Harmonic richness is a ratio of the sum of the energy of the partials as compared to the energy of the fundamental (Childers and Lee, 1991).
(5) The noise-to-harmonics ratio (NHR) is a ratio of the energy near the partials to the energy in the spectral regions between the partials (Childers and Lee, 1991).
(6) Formant dispersion is a metric for the average spectral spread of the formants of a signal, and has been shown to relate to physical attributes such as size and vocal tract length. It is calculated by summing the differences in frequency between adjoining formants and dividing by the number of formants (Fitch, 1997).
(7) Cortical features, which are based on measured neural responses, were extracted with the NSL MATLAB Toolbox (Neural Systems Laboratory, 2003). The features themselves are calculated by sliding spectro-temporal filters on an auditory spectrogram to track low-frequency variations in both time and frequency, then summing the response over time and frequency. This results in a scale-rate cortical representation with two dimensions defined by the time-frequency (rate) and spectral-frequency (scale) of the corresponding filter (Chi *et al.*, 1999).
(8) Raspiness is a novel parameter created for this study that measures the power ratio of the deterministic (sinusoidal) components to the stochastic (noisy) components in speech. This new metric locates the signal on the spectrum between pure tones and pure noise. This metric is aimed at the same voice characteristic as NHR, but is designed to be more stable and reliable to calculate than NHR. The signals were decomposed into sinusoidal and noisy components using the hpsmodel function in the Spectral Modeling Synthesis toolbox (Bonada *et al.*, 2011), and raspiness is defined as the ratio of the energy of those two components.

For each stimulus, all of the parameters in Table I were calculated for the entire word, the vowel, and the nasal, resulting in 72 total parameters. Many of these metrics (such as pitch) are time-varying, and in those cases, mean and standard deviation were calculated after extracting parameters for 25 ms hamming windowed segments every 10 ms. When compiling the mean or standard deviation for a parameter, measurements corresponding to non-speech (such as silence) were excluded.

B. Multi-dimensional scaling

One common analysis technique for similarity matrices in perceptual experiments is multi-dimensional scaling (MDS), (e.g., Kreiman and Papcun, 1991), which creates a low-dimensional set of coordinates for the stimuli. The distances between these coordinates relate to the similarities in the similarity matrix. For perceptual data like the results from this experiment, MDS creates a Euclidean space where stimuli that are close together are perceptually similar in the assigned task.

Performing MDS on the similarity matrix in Fig. 1 creates the four-dimensional space shown in Fig. 3. We chose four dimensions because that is where the quality of fit leveled out (stress = 0.073; $r^2 = 0.897$). Interestingly, Kreiman and Papcun also found four dimensions to be an appropriate choice for representing speaker discrimination results (Kreiman and Papcun, 1991) (recognition results were well represented in only three dimensions). The relative locations of the stimuli in these dimensions (especially the first two

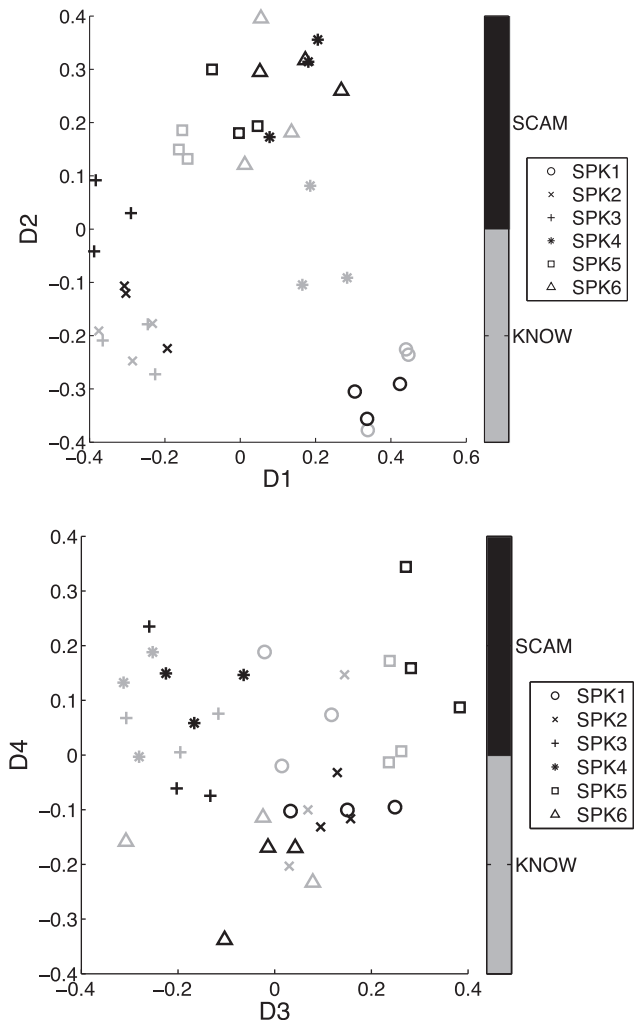

FIG. 3. The four-dimensional mapping derived from the baseline similarity matrix in Fig. 1 using multi-dimensional scaling (MDS). The distances between points in the first two dimensions (D1 and D2) show the dominant patterns in the perceptual results.

dimensions) represent the relationships described above from the similarity matrix, such as the discriminability of SPK1 or the confusability of SPK2 and SPK3.

We can correlate the single-dimensional extracted parameters directly with the MDS dimensions, and Table II shows the parameter that best correlates with each dimension along with the corresponding correlation coefficient (a value between 0 and 1). These correlations contextualize the dimensions of the MDS and try to redefine the space in terms of the parameters. The analysis is restricted to single-dimensional parameters, because multi-dimensional parameters cannot be collectively correlated with the single-dimensional coordinates.

TABLE II. Top correlates in each MDS dimension for the baseline experiment. N, V, and S correspond to nasal, vowel, and signal, respectively. The value in parenthesis is the correlation coefficient with an MDS dimension from Fig. 3.

	Know/Know	Scam/Scam	All
Dim 1	N F_3 mean (0.77)	V F_1 mean (0.57)	V Spectral slope (0.56)
Dim 2	S Raspiness (0.77)	V Raspiness (0.81)	V Raspiness (0.72)
Dim 3	S F_4 mean (0.89)	V Pitch mean (0.85)	S Pitch mean (0.79)
Dim 4	V Formant dispersion (0.41)	V F_3/F_2 (0.51)	S Pitch mean (0.26)

Note that these correlations were performed for three cases: only "know"/"know" trials, only "scam"/"scam" trials, and all trials. Trials comparing the same word were separately analyzed to include parameters that are only meaningful with common phonetic content, such as formant frequencies. In these cases, all of the single-dimensional parameters could be tested for correlation. When including all trials, where only those from the leftmost column of Table I were tested. This resulted in 19 parameters tested for all trials, and 55 tested for each of the same-word comparisons.

There are several conclusions from these correlations. First of all, the second dimension clearly correlates well with the raspiness parameter, as it is the top correlate in all three cases, and the correlation coefficient is reasonably high. Pitch and spectral shape information (such as slope and formants) are prominent in the other dimensions.

C. Logistic regression analysis

In Sec. III, we correlated parameters with derived low-dimensional coordinates to represent the stimuli based on the perceptual results. However, in order to calculate the correlations, we restricted the parameter list to only single-dimensional metrics. Using logistic regression analysis instead, it is possible to analyze the results in terms of the full set of parameters.

We used a logistic regression to model the results with parameter distances as input, and the rate of "same" responses as the output. Because we are using Euclidean distances in the parameter space rather than parameter values themselves, we can use the richer multi-dimensional parameters listed in Table I. Our goal in this analysis is to find the parameters that are best able to model the results, rather than to find the best combination of parameters, and so, as a result, a separate model was built for each parameter.

The smallest negative log likelihoods (and therefore largest likelihoods) of the observed data for the most accurate parameter models are shown in Fig. 4. When looking at all trials, the best parameters are almost identical to those determined by MDS analysis in Table II, with mean pitch, raspiness, and spectral slope all ranking at the top, though this time cortical features are also highly ranked. For "know" trials, the parameters identified by logistic regression analysis are quite different than those from MDS. Here, the first formant and pitch are both highly ranked, as are cortical features and MFCCs. For "scam" trials, there is some overlap with the MDS candidates, with raspiness identified by both, though, otherwise, multi-dimensional parameters mostly dominate the logistic rankings, with LPCs, MFCCs, and cortical features all performing well.

V. EXPERIMENTS WITH MODIFIED STIMULI

In the analysis of the baseline experiment, we identified several candidate parameters as well-related to participant responses. These candidates primarily relate to pitch, formants, spectral or temporal envelopes, and the ratio of noisy components to tonal components in the voice. However, these relationships do not ensure causality, and so we cannot

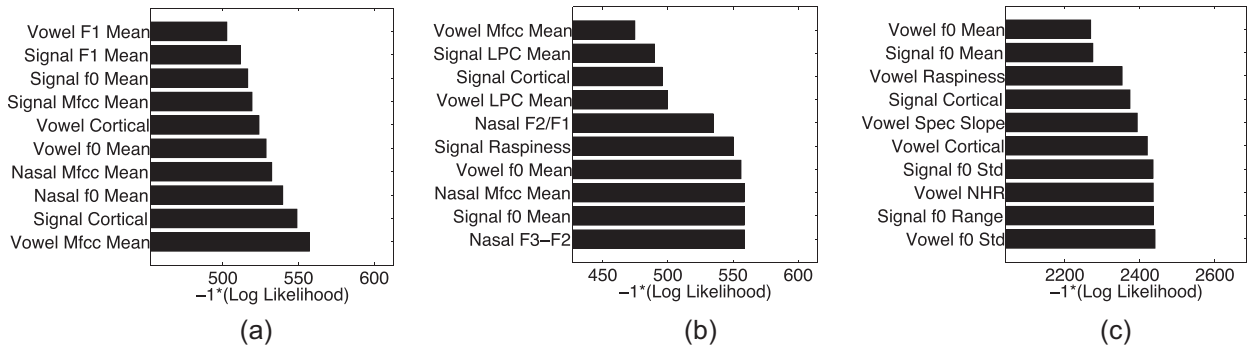

FIG. 4. Negative log likelihoods of the observed data for logistic regression models built on the distance within the listed parameter. Plots show results for (a) "KNOW" only; (b) "SCAM" only; and (c) all trials.

be sure how important these parameters are to the perceptual recognition process. Instead, we only know that they are related to the responses.

To further examine several of these candidate cues, we conducted follow-up experiments in which the stimuli were resynthesized with one of the identified parameters set to some common value. We selected mean pitch, duration, and LPCs for these follow-up experiments. Mean pitch and LPCs are intended to test the significance of pitch and formants, respectively, which were both prevalent in our analysis. LPCs were selected as the preferred parameter for formants as they fit cleanly into the source-filter speech production model, while mean formants and MFCCs are less easily incorporated into resynthesis. Though duration was not one of the most prominent parameters in the correlation analysis (except for MDS correlation on "know"-only trials), it was included because of its prevalence in past work, and also as a step in the process toward LPC normalization (explained below).

Ideally, raspiness would be tested as well, because it was highly rated in all of the correlation analyses, but resynthesis was not reliably stable enough to suffice for human participants in a high-level task like speaker discrimination. Cortical normalization also failed to produce sufficiently clean stimuli.

In each of the experiments to follow, participant recruitment and experiment protocol were the same as in the baseline experiment. The only change was the stimuli used in the experiment. Participants who had already taken the baseline experiment were permitted to participate in follow-up experiments as well, provided there was a minimum of one week separating the sessions.

A. Mean-pitch normalization

For this experiment, the baseline stimuli were resynthesized with the mean pitch shifted to the overall mean pitch (113.27 Hz) using pitch-synchronous overlap add (PSOLA) in Praat. Individual pitch trajectories were preserved.

1. Participants

A total of ten participants (all female) took part in the experiment with mean-pitch normalization. Five of these participants also took part in the baseline experiment.

2. Results

Overall participant accuracy for the mean-pitch-normalized experiment was 78.13%. At first sight, in the similarity matrix (Fig. 5), a small drop in accuracy from the baseline experiment seems to be reflected primarily in stronger similarity between SPK2 and SPK3, to the point that the two appear to be essentially indistinguishable to participants. Figure 2 also shows accuracies for the mean-pitch-normalized experiment for particular phrase comparisons or speakers. Comparisons between experiments will be analyzed in more detail in Sec. VI.

A significant main effect of speaker on the accuracy was revealed by statistical analysis [$\chi^2(5) = 25.09$; $p < 0.0005$]. This effect was due to a difference between SPK1 and both SPK2 and SPK3 ($p < 0.005$).

No main effect of word-pairing was found [$\chi^2(2) = 0.06$; $p = 0.97$], but a significant interaction between speaker and word-pairing was found [$\chi^2(10) = 20.28$; $p < 0.05$] due primarily to differences between SPK1 and all other speakers in trials where one or more speakers uttered "scam" ($p < 0.001$).

3. Multi-dimensional scaling

We repeated the MDS analysis for the mean-pitch-normalized results to derive a low-dimensional representation

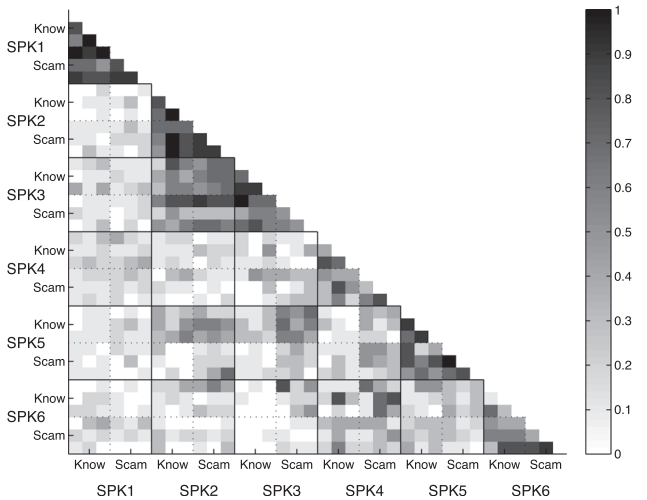

FIG. 5. Similarity matrix for the mean-pitch-normalized experiment ("same" responses dark). Perfect accuracy would lead to a block diagonal matrix.

(stress = 0.090; $r^2 = 0.940$). It is worth noting that, though we once again used a four-dimensional representation for consistency, three dimensions were sufficient for representing these results (stress = 0.123; $r^2 = 0.929$).

The largest parameter correlations with each of these dimensions are shown in Table III. The first dimension is now strongly associated with spectral slope (which was partially the case in the baseline experiment), and the second dimension is still primarily related to the raspiness. Since mean-pitch information has been removed from the stimuli, the mean-pitch correlations with the third and fourth dimensions are also no longer prominent, replaced by a variety of parameters pairing with the third dimension and duration most strongly connected to the fourth.

B. Duration normalization

Baseline stimuli were normalized in duration for each phoneme (and therefore for overall duration) within each word, once again using Praat. So, all utterances of "know" were resynthesized to the same duration, and the same is the case for "scam" utterances.

1. Participants

A total of eight participants (one male, seven female) took part in the experiment with duration-normalized stimuli. Two of these participants also took part in the baseline experiment.

2. Results

Participants identified duration-normalized speakers with 82.22% accuracy. In the similarity matrix (Fig. 6), SPK2 and SPK3 are still not well distinguished from each other. Accuracies for each phrase pair or speaker are shown in Fig. 2.

As in previous experiments, speaker had a significant effect on accuracy [$\chi^2(5) = 19.43$; $p < 0.005$]. This effect was primarily due to a difference between SPK1 and both SPK2 and SPK6 ($p < 0.05$).

No effect of word-pairing was found [$\chi^2(2) = 3.38$; $p = 0.18$], but a significant interaction was found between speaker and word pairing [$\chi^2(10) = 19.14$; $p < 0.05$], though this effect is not due to a clear pattern.

3. Multi-dimensional scaling

We repeated the MDS analysis to derive a four-dimensional representation for these results as well (stress = 0.076; $r^2 = 0.912$). In the parameter correlations, seen in Table IV, there are several differences from the baseline correlations, but many of the prominent patterns persevere. Raspiness is still strongly correlated to the second dimension (though spectral slope is a better match for "scam" trials), and pitch and formant statistics are consistently well correlated. These results would suggest that the prominent features used by listeners were not greatly changed by duration normalization, which is not surprising considering the similarity of the experiment results.

TABLE III. Top correlates in each MDS dimension for the mean-pitch-normalized experiment. N, V, and S correspond to nasal, vowel, and signal, respectively.

	Know/Know	Scam/Scam	All
Dim 1	S Spectral slope (0.86)	V Spectral slope (0.75)	V Spectral slope (0.68)
Dim 2	V F_2 mean (0.80)	V Raspiness (0.64)	V Raspiness (0.59)
Dim 3	V F_4 mean (0.65)	S F_3/F_2 (0.82)	S NHR (0.42)
Dim 4	V Duration (0.67)	V Duration (0.55)	S Harmonic richness (0.28)

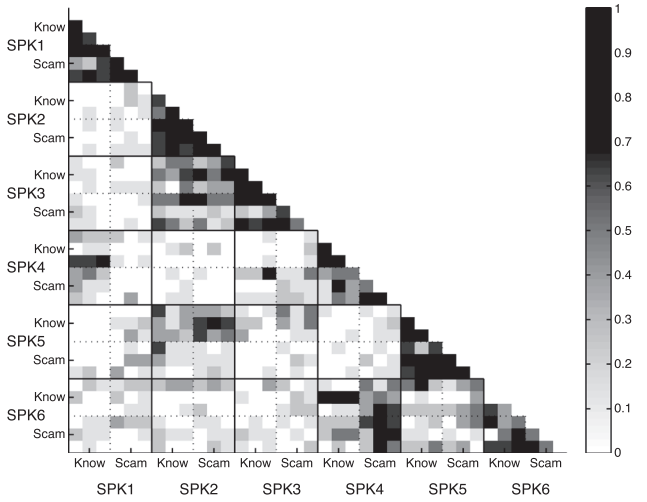

FIG. 6. Similarity matrix for the duration-normalized experiment ("same" responses dark). Perfect accuracy would result in a block diagonal matrix.

C. LPC normalization

The stimuli for the LPC-normalized experiment were resynthesized from the duration-normalized stimuli so that the source and filter in the model would align without any timing discrepancies. We used the Auditory Toolbox (Slaney, 1998) to recreate each signal with one of two sets of common LPCs (one set for each word).

Past research has shown that linear prediction is influenced by fundamental frequency, and so this is an experimental noise that should be considered. Fortunately, this effect is less significant for fundamental frequencies below 350 Hz (Monsen and Engebretson, 1983), which is a threshold well above the maximum pitch of the male speech used in this study (164.4 Hz). Furthermore, the expected error due to frequency quantization of the harmonic peaks in the spectrum is only 10% of pitch (Vallabha and Tuller, 2002), which, in our case, means an expected error of less than

TABLE IV. Top correlates in each MDS dimension for the duration-normalized experiment. N, V, and S correspond to nasal, vowel, and signal, respectively.

	Know/Know	Scam/Scam	All
Dim 1	V F_1 mean (0.85)	V Pitch range (0.68)	S Pitch mean (0.55)
Dim 2	S Raspiness (0.79)	V Spectral slope (0.76)	V Raspiness (0.62)
Dim 3	N F_4 mean (0.68)	V F_4 mean (0.79)	V Pitch mean (0.57)
Dim 4	V Pitch range (0.54)	V NHR (0.73)	S Pitch std (0.62)

15 Hz. So, while the effects are present, they are expected to be relatively minor.

1. Participants

A total of seven participants (all females) took part in the study with LPC-normalized stimuli. Four of these participants also took part in the baseline experiment.

2. Results

Participants performed the LPC-normalized experiment with an average accuracy of 77.55%. The similarity matrix for these results, shown in Fig. 7, appears noisier than the matrices for the other experiments, especially compared to the baseline. This point will be further discussed in Sec. VI. The accuracies for each phrase pair or speaker are shown in Fig. 2.

Again, speaker had a significant main effect on accuracy [$\chi^2(5) = 33.97$; $p < 0.0001$], explained by a difference between SPK1 and both SPK5 and SPK6, as well as a difference between SPK4 and SPK6 ($p < 0.005$).

There was not a significant effect from word-pairing [$\chi^2(2) = 0.17$; $p = 0.92$] or the interaction of speaker and word-pairing [$\chi^2(10) = 17.83$; $p = 0.06$].

3. Multi-dimensional scaling

A four-dimensional MDS space was derived for the LPC-normalized experimental results as well (stress = 0.85; $r^2 = 0.925$). The most prominent parameter correlations for each MDS dimension are shown in Table V. With the loss of formant information, the correlations are now predominated by pitch statistics and raspiness, the latter once again strongly pairing with the second dimension. Spectral slope also correlates well in a few cases, but this table primarily suggests that with the loss of formant information, pitch becomes an even more prominent feature in the decision process. However, it is also worth noting that many of the maximum correlation values in the third and fourth dimension are small, suggesting that the set of parameters examined in this study may not be sufficient for describing these results.

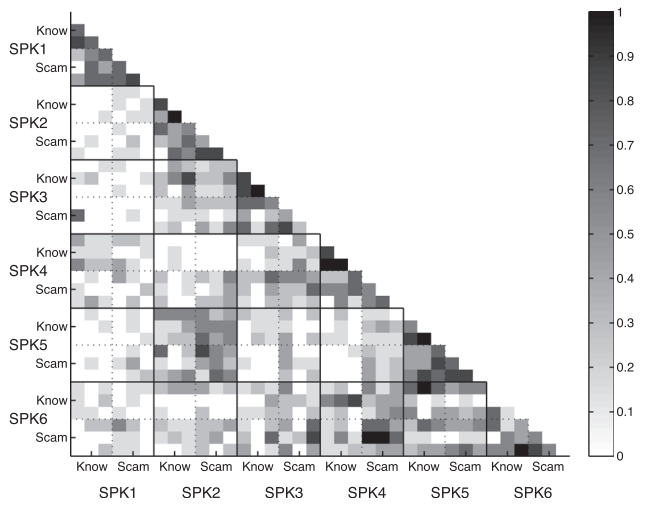

FIG. 7. Similarity matrix for the LPC-normalized experiment ("same" responses dark). Perfect accuracy would result in a block diagonal matrix.

TABLE V. Top correlates in each MDS dimension for the LPC-normalized experiment. N, V, and S correspond to nasal, vowel, and signal, respectively.

	Know/Know	Scam/Scam	All
Dim 1	V Pitch mean (0.89)	S Pitch std (0.77)	S Pitch mean (0.80)
Dim 2	S Raspiness (0.64)	S Raspiness (0.71)	S Raspiness (0.58)
Dim 3	S Raspiness (0.44)	N Pitch slope (0.43)	N Pitch slope (0.39)
Dim 4	N Pitch slope (0.35)	S Spectral slope (0.78)	S Spectral slope (0.44)

VI. INTER-EXPERIMENTAL RESULTS

A. Statistical analysis

Similarly to each individual experiment, the full set of experiments was fit with a generalized linear mixed model (logistic regression) treating experiment, speaker, and word-pairing as fixed effects and listener as a random effect, and significance of effects was tested using a Wald χ^2 test.

A main effect of the experiment was found [$\chi^2(3) = 45.05$; $p < 0.0001$], due primarily to the difference between the baseline and pitch-normalized experiment ($p < 0.05$), though the differences of the baseline and LPC-normalized experiment were only slightly less impactful. A significant main effect was also found for speaker [$\chi^2(5) = 71.78$; $p < 0.0001$], due to differences between SPK1 and all other speakers ($p < 0.01$), and a significant main effect was found for word-pairing [$\chi^2(2) = 30.95$; $p < 0.0001$] due to differences between "scam"/"scam" trials and all other trials.

A significant interaction was found between speaker and experiment [$\chi^2(15) = 68.79$; $p < 0.0001$], an effect that can be seen in Fig. 2(a). This interesting result is in line with our previous observations in the modification experiments that different modifications have unique effects on individual speakers.

A significant interaction was found between word-pairing and experiment [$\chi^2(6) = 128.97$; $p < 0.0001$]. The variations between word-pairings as a function of acoustic modification can easily be seen in Fig. 2(b), especially in the case of "scam"/"scam" trials versus other cases.

A significant interaction between speaker and word [$\chi^2(10) = 30.81$; $p < 0.001$] was also found, though it could not be explained by some specific or clear pattern in the results.

B. Speaker confusion

One especially noteworthy pattern that emerged within the speaker group was the consistent confusion of SPK2 and SPK3. The two speakers were identified as "same" disproportionately in the baseline experiment, the duration-normalized experiment, and the mean-pitch-normalized experiment. However, interestingly, the two were only very weakly paired in the LPC-normalized experiment. This suggests that listeners found SPK2 and SPK3 easier to distinguish after all LPC differences were eliminated. However, comparisons of the LPCs for each speaker pair do not show any distinction that would indicate why the SPK2/SPK3 pairing is affected differently than others. So, though we have empirical evidence that their confusion is related to the

LPCs, parameter analysis unfortunately does not give any further insight into this confusion.

SPK1, on the other hand, is significantly more easily discriminated by listeners than every other voice in all four experiments, indicating there must be some characteristic of SPK1 that aids listeners in the discrimination process (at least against these other five voices). At this point, however, all that we can say is that it is not mean pitch, duration, or LPC information that distinguishes SPK1, because, even after the removal of these features, listeners were still able to discriminate SPK1 with significantly greater accuracy. Analysis of the remaining parameters does not show any candidates where SPK1 is measurably different from the other speakers, and so beyond eliminating the tested parameters, we were unable to determine why listeners find SPK1 easier to discriminate.

C. Listener spread

Box plots of participant accuracy for each experiment are shown in Fig. 8, and comparing the statistics of each experiment suggests an interesting difference between mean-pitch normalization and other modifications. After LPC or duration normalization, the performance distribution is still relatively similar to the baseline distribution, but shifted downward and with a slightly wider spread. However, mean-pitch normalization leads to a much more skewed distribution with a denser set of performances in the range of baseline results, but a long tail with poor performances. This is especially evident in the median's relative location to the upper and lower quartiles. This observation is supported by a Levene test [$F(3, 36) = 2.77$; $p = 0.05$].

D. Discussion

Within the aggregation of the experimental results, there are several noteworthy trends and observations.

(1) Overall, there was a statistically significant difference in listener performance after modifications. The acoustic modifications of the stimuli between experiments significantly impacted the ability of listeners to discriminate speakers. This effect was most clearly measured after pitch normalization, but LPC normalization also had a large impact. However, it is worth noting that, despite these significant effects, listeners are able to discriminate speakers at a high rate in all cases, despite limited duration and phonetic variation. The persistence of high discriminability throughout the modification experiments, especially evident with SPK1, suggests both a robustness of perceptual discriminability, as well as a role for additional acoustic features in the process.

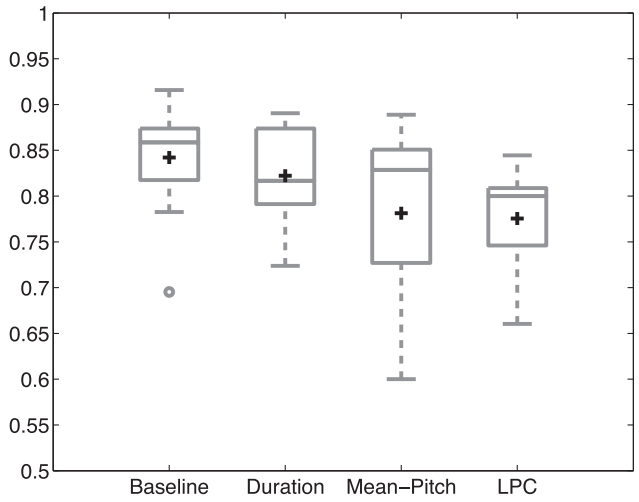

FIG. 8. Box plots derived from participant accuracy for each of the four experiments. The mean for each set is also displayed with a bold "plus."

(2) The loss of mean-pitch or LPC information damages discriminability differently. It is interesting that, though the drop in overall accuracy resulting from LPC or mean-pitch normalization is similar, the composition of those effects is not always the same. For example, participants found the "scam"/"scam" trials much more difficult after LPC normalization than for any other experiment, but the mean-pitch-normalized experiment was more difficult for the other two phrase pairs. Similarly, SPK3 and SPK6 were oppositely affected by mean-pitch and LPC normalization, while SPK4 and SPK5 rates are almost identical for the two experiments. In general, the results show that both parameters are important in the perceptual process, but that the effects are quite different. It is also worth noting that Gaudrain *et al.* (2009) found that listeners are much more tolerant to change in pitch than to change in vocal tract length (which affects the LPCs), but these results suggest that the normalization of either of the two parameters can have an effect, depending of speaker and phonetic content. Note, though, that the referenced study examined speaker similarity, rather than pure discrimination, and did so with many examples of varying degrees of pitch or vocal tract length differences, rather than the binary nature of the modifications in the stimuli used here. It is very possible that listeners are indeed more tolerant to changes in pitch, but that the manipulations seen in the present experiment are not equally spaced on the tolerance curves.

(3) Trials comparing only "scam" stimuli were more affected by modifications than other word pairings. It had already been clear when analyzing the individual experiment results that these trials were more adversely affected than those with "know" only or those comparing the different words, but it is not clear why "scam" trials were so much more susceptible to modifications (or, alternatively, why trials involving "know" were so much more robust). It is possible that the unvoiced phonemes that begin "scam" affect the quality of modification (particularly with regards to pitch), though no audible artifacts were detected for either word. This possibility would require testing with a larger set of words to discuss beyond conjecture.

(4) Timbral metrics like raspiness appear to be important and should be further analyzed. We introduced a metric called raspiness, based on the ratio of the sinusoidal energy to the stochastic energy in the signal, and the parameter consistently scored well as a candidate feature for perceptual speaker discrimination. This trend was

found even in the experiment with modified stimuli, with raspiness correlating well with the second MDS dimension for every experiment. Unfortunately, we were not able to synthesize sufficiently clean stimuli to test this parameter alongside the others studied above, but this parameter or others like it should be considered for future studies.

(5) SPK2 and SPK3 were perceived as similar in all cases except after LPC normalization. One of the most salient trends persistent through the set of experiments was the high degree of confusion between SPK2 and SPK3, except in the case of the LPC-normalized stimuli, when the two speakers were confused less often. The improved ability of participants to distinguish SPK2 and SPK3 after LPC-normalization is an especially interesting result. Typically, it would be expected that participants would find stimuli more similar after normalizing any parameter, since this normalization can only bring the stimuli closer together in the parameter space. It is possible that this is simply a result of unintended distortion in the resynthesis, but it could also indicate an adaptation of decision criteria by the listeners. In this scenario, participants would utilize LPC information in some form in the baseline, duration-normalized, and mean-pitch-normalized experiments, and that information contributes to the confusion of SPK2 and SPK3. However, after the LPC information is normalized, the participant relies on some other set of features in which the two speakers are not as perceptually close. The notion of adapted criteria is supported by the MDS correlations, where the best correlating parameters change somewhat for each experiment, though there also appear to be several persistent criteria, such as raspiness.

(6) MDS modeling after mean-pitch normalization needed one fewer dimension than all other experiments. All four experiment results were modeled with a four-dimensional MDS coordinate set, but only in the case of the mean-pitch-normalized experiment was the three-dimensional representation also sufficient. This observation supports the importance of mean pitch in the perceptual decision process, and it also suggests that, after the loss of mean-pitch information, listeners simply ceased to use mean pitch in the decision and did not replace that feature with a new and separate feature. Unlike the discussion above regarding LPC normalization, this would suggest that listeners did not adapt the decision criteria to the new stimuli in this case. These two observations could suggest that listeners adapt in some cases but not in others.

(7) The loss of mean-pitch information appears to spread the performance distribution more than other modifications. Unlike after the other modifications, the drop in accuracy after mean-pitch normalization is related to large drops in performance by a few of the participants. Though duration normalization had a minimal effect on overall performance, LPC and mean-pitch normalization had similar overall effects, so this distinction in spread of the accuracies is interesting. It would seem to suggest that only a few participants depended on mean pitch information in their decision (but those participants depended heavily on it), while all participants used LPCs in their decision process to some potentially lesser degree. Unfortunately, this experimental design does not allow for statistically analyzing this suggestion, but it is an intriguing possibility.

VII. CONCLUSION

In this report, we presented results from a baseline experiment with unprocessed speech followed by several experiments that each eliminated the variation in selected parameters from the stimuli. These stimuli were selected from a database of speakers relevant to the automatic speaker identification community, and our experiments demonstrate that these recordings can also be useful for perceptual research. It is the hope that other researchers will utilize these databases similarly to help bring the automatic and perceptual communities closer together and also to maintain a consistent set of stimuli for aggregating perceptual results.

From the results, we also drew several conclusions. First of all, this research shows once again that human listeners are highly robust to distortion and modification. While we did find changes in accuracy as a result of the changes in the stimuli, participants were still consistently able to perform to a high level, and always above chance.

Within the specific experiments, we found that manipulating phonetic duration has a minimal effect on participant performance. However, it is important to distinguish that this does not mean that longer syllabic-level timing like prosody would not have a greater importance.

Our results also showed that both mean pitch and LPCs are important cues for speaker discrimination, and that the loss of either does affect the ability of a listener to identify certain speakers or speakers uttering certain word-pairings, conclusions in agreement with past studies as well. However, after the loss of either, there was still always sufficient information remaining in the signal for participants to perform the task above chance in all conditions. This would suggest that unsurprisingly, a human listener uses features beyond pitch, spectral envelope, or duration in speaker discrimination, and indeed many past studies agree that listeners can perform speaker tasks without subsets of that information (Coleman, 1973; Lavner *et al*., 2000; Sheffert *et al*., 2002; Van Lancker *et al*., 1985a). On the basis of our acoustic analyses, raspiness may be one of these additional sources of information.

While the loss of either mean pitch or LPCs did have an effect on participant performance, the two did not affect it in the same way. Interestingly, the normalization of mean pitch appears to affect certain participants more than others, while normalization of LPCs has a more consistent effect. Furthermore, each normalization can have a drastic effect on specific phrase pairs or speakers, but comparatively very little effect on others.

The results of this experiment suggest several possible follow-up experiments to answer a few of the questions raised. First of all, testing raspiness and cortical features would be valuable, because they all showed importance in the baseline analysis, and several timbre features similar to raspiness were also suggested in previous research.

Normalizing multiple parameters simultaneously would also be useful to examine if the effects of these manipulations compound each other, behave independently, or diminish each other. Our analysis also raised the possibility that the type of database being used affects the decision strategies for the participants, and an experiment testing this effect would be useful, not only for understanding this study, but potentially for re-examining other past research as well. The relative effects of modification on "scam" trials as compared to "know" trials is also intriguing, and a follow-up experiment to determine the reasons for this discrepancy would be informative. Finally, testing participant performance on a database of mixed stimuli would allow an analysis of the relative effects of each type of modification on individuals.

ACKNOWLEDGMENTS

The authors would like to give special thanks to Dr. John Godfrey at Johns Hopkins University for his advice and help. We would also like to recognize the IC Postdoctoral Research Fellowship Program for funding these studies. This work was also supported in part by grants IIS-0846112 (NSF), 1R01AG036424-01 (NIH), N000141010278 and N00014-12-1-0740 (ONR).

Amino, K., and Arai, T. (**2009**). "Effects of linguistic contents on perceptual speaker identification: Comparison of familiar and unknown speaker identifications," Acoust. Sci. Technol. **30**(2), 89–99.

Amino, K., Sugawara, T., and Arai, T. (**2006**). "Idiosyncrasy of nasal sounds in human speaker identification and their acoustic properties," Acoust. Sci. Technol. **27**(4), 233–235.

Bachorowski, J.-A., and Owren, M. J. (**1999**). "Acoustic correlates of talker sex and individual talker identity are present in a short vowel segment produced in running speech," J. Acoust. Soc. Am. **106**(2), 1054–1063.

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

Bonada, J., Serra, X., Amatriain, X., and Loscos, A. (**2011**). "Spectral processing," in *DAFX: Digital Audio Effects*, 2nd ed., edited by U. Zölzer (John Wiley and Sons, West Sussex, UK), pp. 393–446.

Brandstein, L., Graff, D., and Walker, K. (**2013**). "Mixer-6 Speech LDC2013S03," Philadelphia: Linguistic Data Consortium.

Bricker, P. D., and Pruzansky, S. (**1966**). "Effects of stimulus content and duration on talker identification," J. Acoust. Soc. Am. **40**(6), 1441–1449.

Brown, B. L., Strong, W. J., and Rencher, A. C. (**1973**). "Perceptions of personality from speech: Effects of manipulations of acoustical parameters," J. Acoust. Soc. Am. **54**(1), 29–35.

Brown, B. L., Strong, W. J., and Rencher, A. C. (**1974**). "Fifty-four voices from two: The effects of simultaneous manipulations of rate, mean fundamental frequency, and variance of fundamental frequency on ratings of personality from speech," J. Acoust. Soc. Am. **55**(2), 313–318.

Chi, T., Gao, Y., Guyton, M. C., Ru, P., and Shamma, S. (**1999**). "Spectro-temporal modulation transfer functions and speech intelligibility," J. Acoust. Soc. Am. **106**(5), 2719–2732.

Childers, D. G., and Lee, C. K. (**1991**). "Vocal quality factors: Analysis, synthesis, and perception," J. Acoust. Soc. Am. **90**(5), 2394–2410.

Coleman, R. O. (**1973**). "Speaker identification in the absence of inter-subject differences in glottal source characteristics," J. Acoust. Soc. Am. **53**(6), 1741–1743.

Compton, A. J. (**1963**). "Effects of filtering and vocal duration upon the identification of speakers, aurally," J. Acoust. Soc. Am. **35**(11), 1748–1752.

Fitch, W. T. (**1997**). "Vocal tract length and formant frequency dispersion correlate with body size in rhesus macaques," J. Acoust. Soc. Am. **102**(2), 1213–1222.

Gaudrain, E., Li, S., Ban, V. S., and Patterson, R. D. (**2009**). "The role of glottal pulse rate and vocal tract identity in the perception of speaker identity," in *Interspeech 2009*, pp. 148–151.

Greenberg, C. S., Martin, A. F., Doddington, G. R., and Godfrey, J. J. (**2011**). "Including human expertise in speaker recognition systems: Report on pilot evaluation," in *Proceedings of ICASSP*, pp. 5896–5899.

Hartman, D. E., and Danhauer, J. L. (**1976**). "Perceptual features of speech for males in four perceived age decades," J. Acoust. Soc. Am. **59**(3), 713–715.

Kerstholt, J. H., Jansen, N. J. M., Van Amelsvoort, A. G., and Broeders, A. P. A. (**2004**). "Earwitnesses: Effects of speech duration, retention interval and acoustic environment," Appl. Cognit. Psychol. **18**, 327–336.

Köster, O., and Schiller, N. O. (**1997**). "Different influences of the native language of a listener on speaker recognition," Forensic Linguist. **4**(1), 18–28.

Kreiman, J., and Papcun, G. (**1991**). "Comparing discrimination and recognition of unfamiliar voices," Speech Commun. **10**, 265–275.

Kreiman, J., and Sidtis, D. (**2011**). *Foundations of Voice Studies: An Interdisciplinary Approach to Voice Production and Perception* (Wiley-Blackwell, Hoboken, NJ), pp. 156–301.

Kuwabara, H., and Takagi, T. (**1991**). "Acoustic parameters of voice individuality and voice-quality control by analysis-synthesis method," Speech Commun. **10**, 491–495.

Lass, N. J., Hughes, K. R., Bowyer, M. D., Waters, L. T., and Bourne, V. T. (**1976**). "Speaker sex identification from voiced, whispered, and filtered isolated vowels," J. Acoust. Soc. Am. **59**(3), 675–678.

Lavner, Y., Gath, I., and Rosenhouse, J. (**2000**). "The effects of acoustic modifications on the identification of familiar voices speaking isolated vowels," Speech Commun. **30**, 9–26.

Maheshwari, N., Philip, L., and Savithri, S. R. (**2008**). "Speaker identification by native and non-native speakers," in *Proceedings of FRSM-2008*, pp. 181–183.

Monsen, R. B., and Engebretson, A. M. (**1983**). "The accuracy of formant frequency measurements: A comparison of spectrographic analysis and linear prediction," J. Speech Hear. Res. **26**(1), 89–97.

Neural Systems Laboratory (**2003**). "NSL MATLAB Toolbox," University of Maryland http://www.isr.umd.edu/Labs/NSL/Software.htm (Last viewed July 23, 2013).

Perrachione, T. K., Del Tufo, S. N., and Gabrieli, J. D. E. (**2011**). "Human voice recognition depends on language ability," Science **333**, 595.

Pollack, I., Pickett, J. M., and Sumby, W. H. (**1954**). "On the identification of speakers by voice," J. Acoust. Soc. Am. **26**(3), 403–406.

Remez, R. E., Fellowes, J. M., and Nagel, D. S. (**2007**). "On the perception of similarity among talkers," J. Acoust. Soc. Am. **122**(6), 3688–3696.

Remez, R. E., Fellowes, J. M., and Rubin, P. E. (**1997**). "Talker identification based on phonetic information," J. Exp. Psychol. Human **23**(3), 651–666.

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. Human **28**(6), 1447–1469.

Skuk, V. G., and Schweinberger, S. R. (**2013**). "Influences of fundamental frequency, formant frequencies, aperiodicity, and spectrum level on the perception of voice gender," J. Speech Lang. Hear. Res. **57**, 1–12.

Slaney, M. (**1998**). "Auditory toolbox," Technical report 1998-010, Interval Research Corporation, Palo Alto, CA.

Smith, D. R. R., and Patterson, R. D. (**2005**). "The interaction of glottal-pulse rate and vocal-tract length in judgements of speaker, size, and age," J. Acoust. Soc. Am. **118**(5), 3177–3186.

Thompson, C. P. (**1987**). "A language effect in voice identification," Appl. Cognit. Psychol. **1**, 121–131.

Vallabha, G. K., and Tuller, B. (**2002**). "Systematic errors in the formant analysis of steady-state vowels," Speech Commun. **38**, 141–160.

Van Lancker, D., and Kreiman, J. (**1987**). "Voice discrimination and recognition are separate abilities," Neuropsychologia **25**(5), 829–834.

Van Lancker, D., Kreiman, J., and Emmorey, K. (**1985a**). "Familiar voice recognition: Patters and parameters, Part I: Recognition of backward voices," J. Phonetics **13**, 19–38.

Van Lancker, D., Kreiman, J., and Wickens, T. D. (**1985b**). "Familiar voice recognition: Patters and parameters, Part II: Recognition of rate-altered voices," J. Phonetics **13**, 39–52.

Voiers, W. D. (**1964**). "Perceptual bases of speaker identity," J. Acoust. Soc. Am. **36**(6), 1065–1073.

Voiers, W. D. (**1979**). "Toward the development of practical methods of evaluating speaker recognizability," in *Proceedings of ICASSP*, Washington, DC (IEEE), pp. 793–796.

Yarmey, A. D., Yarmey, A. L., Yarmey, M. J., and Parliament, L. (**2001**). "Commonsense beliefs and the identification of familiar voices," Appl. Cognit. Psychol. **15**, 283–299.