On the perception of voicing in syllable-initial plosives in noise[a)]

Jintao Jiang,[b)] Marcia Chen, and Abeer Alwan[c)]

Department of Electrical Engineering, University of California, Los Angeles, CA 90095

Received.

Running title: Voicing perception in noise

[a)] Portions of this work were presented at the 139th Meeting of the Acoustical Society of America (2000) and EUROSPEECH 2001. This work is based in part on Marcia Chen's MS thesis, UCLA, 2001.

[b)] Now with the House Ear Institute, Los Angeles, CA 90057

[c)] Electronic mail: alwan@ee.ucla.edu

Abstract

Previous studies [Lisker, J. Acoust. Soc. Am. **57**, 1547-1551 (1975); Summerfield and Haggard, J. Acoust. Soc. Am. **62**, 435-448 (1977)] have shown that voice onset time (VOT) and the onset frequency of the first formant are important perceptual cues of voicing in syllable-initial plosives. Most prior work, however, has focused on speech perception in quiet environments. The present study seeks to determine which cues are important for the perception of voicing in syllable-initial plosives in the presence of noise. Perceptual experiments were conducted using stimuli consisting of naturally-spoken consonant-vowel syllables by four talkers in various levels of additive white Gaussian noise. Plosives sharing the same place of articulation and vowel context (e.g., /pa,ba/) were presented to subjects in 2AFC identification tasks, and a threshold SNR value (corresponding to the 79% correct classification score) was estimated for each voiced/voiceless pair. The threshold SNR values were then correlated with several acoustic measurements of the speech tokens. Results indicate that the onset frequency of the first formant is critical in perceiving voicing in syllable-initial plosives in additive white Gaussian noise, while the VOT duration is not.

PACS numbers: 43.71.Es, 43.71.Bp

I. INTRODUCTION

Research on speech perception and human auditory processes, particularly in the presence of background noise, helps to improve and calibrate such practical applications as noise-robust automatic speech recognition systems (e.g., Hermansky, 1990; Strope and Alwan, 1997) and aids for the hearing impaired. Allen (1994) has argued that the study of speech perception in noise is not only practically important for automatic speech recognition but also critical for a complete and correct understanding of the psychology of human speech perception. The present study examines the contributions of various acoustic parameters to the perception of the feature *voicing* in syllable-initial plosives in the presence of additive white Gaussian noise. Questions addressed in this paper include: How does the perception of voicing vary with signal-to-noise ratio (SNR)? Does the threshold SNR of voicing detection vary with place of articulation and/or vowel context? What acoustic properties account for the perception of voicing in noise?

The present study focuses on plosives that span three different places of articulation: labials (/b,p/), alveolars (/d,t/), and velars (/g,k/). These six consonants are further classified as voiced {/b/, /d/, /g/} or voiceless {/p/, /t/, /k/}. Plosive consonants are produced by first forming a complete closure in the vocal tract via a constriction at the place of articulation, during which there is either silence or a low-frequency hum (called *voicebar/prevoicing*). The vocal tract is then opened suddenly releasing the pressure built up behind the constriction; this is characterized acoustically by a transient and/or a short-duration noise burst (Stevens, 1998). The period between the release of the plosive and the beginning of voicing in the vowel is called the voice onset time (VOT). During this period there is silence and/or aspiration noise. The time interval between the onset of the following vowel and the instance when a formant frequency reaches its steady-state value is called the formant transition.

Previous studies on the voiced/voiceless distinction in plosives have focused primarily on the VOT duration. It has been shown that the VOT duration of voiced plosives is significantly shorter than voiceless plosives. Liberman *et al.* (1958) conducted perceptual experiments using synthetic syllable-initial plosives (/b/, /d/, and /g/), where the onset of F1 was cut back (delayed) by amounts varying between 10 and 50 ms relative to the burst; F2 and F3, however, began immediately after the release. The authors concluded that the VOT duration (as defined by the amount of F1 cutback) was a cue for voicing, and that replacing F2 and F3 with noise, instead of harmonics, increased the voiceless effect. Lisker and Abramson (1964) measured the VOT duration of several naturally-spoken utterances, comprised of {/b/, /d/, /g/, /p/, /t/, /k/} followed by the vowel /a/. The average VOT duration of voiced plosives was found to be 1, 5, and 21 ms for /b/, /d/, and /g/ in English, respectively. In contrast, the average VOT duration of voiceless plosives was found to be 58, 70, and 80 ms for /p/, /t/, and /k/ in English, respectively. There were no values in common between voiced/voiceless pairs of plosives with the same place of articulation (e.g., /b,p/). The VOT duration was therefore a good and easily measurable acoustic property to distinguish between voiced and voiceless plosives. Lisker and Abramson (1970) followed up their previous study with perceptual experiments using synthetic stimuli of varying VOT durations. Stimuli were made to sound like {/b,p/, /d,t/, /g,k/}, in English, followed by /a/. Tokens were identified as voiced if their VOT durations were under 15, 20, and 30 ms for /b/, /d/, and /g/, respectively; tokens were identified as voiceless if their VOT durations were over 30, 40, and 50 ms for /p/, /t/, and /k/, respectively. The study concluded that the voicing boundary for English tokens changed along the VOT duration continuum as a function of place of articulation.

Lisker and Abramson (1964; 1970) also demonstrated that VOT duration and the voicing boundary along the VOT duration continuum varied from language to language, although there were within-language regularity and cross-language predictability in the realization of voicing contrasts (Cho and Ladefoged, 1999).

More recent research has focused on how these different acoustic cues are integrated or traded during perception: Many studies have demonstrated the occurrence of cue interactions (e.g., Fitch *et al.*, 1980; Miller, 1977; Sawusch and Pisoni, 1974) and modeled the interactions (Massaro and Oden, 1980; Repp, 1983). According to Stevens and Klatt (1974; see also Lisker *et al.*, 1977), the VOT duration boundary between voiced and voiceless tokens was unstable and varied depending on the presence or absence of a rapidly changing F1 transition. In Stevens and Klatt's (1974) study, perceptual tests were conducted using synthetic /da,ta/ stimuli with varying F1 transition rates and VOT durations. The authors reported that, for one listener, classification seemed to be based solely on the VOT duration: The VOT duration boundary was always 30 ms regardless of transition slope. For another listener, results seemed to be based on the duration of the F1 transition. The other three listeners seemed to use a mixture of cues. The authors also conducted a separate experiment to show that a transition was perceptually detectable at about 13 ms and a rapid spectrum change was indicative of voicing. Lisker (1975) conducted further experiments on the F1 transition using synthetic stimuli of /ga/ and /ka/ and found that although F1 had a significant effect on the voiced/voiceless classification, it was neither necessary nor as sufficient as the VOT duration. In addition, it was not the dynamic quality (rapidly changing) of F1 but the low F1 onset frequency that indicated voicing. Summerfield and Haggard (1977) refined and extended Lisker's (1975) conclusion. The authors showed that when F1 onset frequency and F1 transition duration were independently controlled in the /g,k/ pair, only F1 onset frequency had a significant effect on voicing perception. When a factorial F1 (four steady-state frequencies) x F2 (four steady-state frequencies) design was used for the /d,t/ distinction, the authors further ruled out F2 or F1 x F2 interaction as possible voicing cues. These results suggest that there may be a perceptual trading relationship between F1 onset frequency and VOT duration (i.e., a low F1 onset frequency, usually from a high vowel, resulted in a longer threshold VOT) and that the trading relationship is vowel-dependent. Hall *et al.* (1995) found that VOT duration and aspiration amplitude were completely integral using speeded classification experiments. Benki (2001) experimented with synthetic vowel-consonant-vowel and consonant-vowel (CV) stimuli, where F1 transition, place of articulation (bilabial, alveolar, and velar), and VOT duration were manipulated. The author found that, in quiet, place of articulation and F1 transition characteristics were both important phonetic cues for voicing categorization and that the magnitude of the F1 transition effects was considerably larger than the place of articulation effect.

Other acoustic cues were also investigated with respect to voicing classification. Peterson and Lehiste (1960) found that the vowel duration was not a reliable cue for voicing in syllable-initial consonants. The loudness of the burst has been speculated to be a distinguishing characteristic for voicing (Lisker and Abramson, 1964): Voiceless plosives have a greater articulatory force, or a louder burst, than voiced ones. Klatt (1975) showed that aspiration during the period of closure was present in voiceless plosives, but absent in voiced plosives. In the same study, Klatt reported that the VOT duration changed as a function of the place of articulation and was longer before high vowels than before mid- and low vowels. Repp (1979) showed in perceptual tests with synthetic stimuli that the VOT duration boundary between /da/ and /ta/ was linearly dependent on the ratio between the amplitude of the aspiration and that of the vowel. A one-decibel increase of the ratio led to a shortening of the VOT duration boundary by, on average, 0.43 ms. In the same study, other experiments were conducted with varying burst and aspiration amplitudes. Both acoustic properties were seen to affect the voicing boundary (a louder burst or louder noise aspiration increased the voiceless effect), although the amplitude of aspiration had a larger effect. The fundamental frequency (F0) has also been suggested as a possible cue for voicing in stops (Haggard *et al.*, 1970; Ohde, 1984; Whalen *et al.*, 1993). Haggard *et al.* (1970) examined the perception of synthetic /bi,pi/ stimuli and showed that utterances with a low rising F0 across the transition region indicated voicing, while a high falling F0 indicated voiceless. Ohde (1984) examined several naturally-spoken CVC tokens by male talkers and found that the absolute value of F0 immediately following the onset of voicing was higher in voiceless stops (average 135 Hz) than voiced stops (average 103 Hz). Also, the drop in F0 between the first glottal pulse and the second pulse was larger for voiceless stops (average 16 Hz) than voiced stops (average 3 Hz).

The studies discussed above were all conducted in quiet environments; however, speech is often heard in the presence of background noise. Therefore, it is of great importance to investigate speech perception in noise. Extensive studies on the perceptual confusions between consonants in the presence of noise have been conducted, most notably by Miller and Nicely (1955). Their study used naturally-spoken utterances consisting of one of 16 consonants followed by the vowel /a/, in varying levels of noise and band pass filtering configurations. Their work showed that voicing was much less affected by noise than other features (such as the place of articulation). Voicing was still discriminable at SNRs as low as –12 dB, while place information, in contrast, was difficult to distinguish at SNRs less than +6 dB. The perceptual experiments for that study, however, allowed confusions between all consonants, and not just the plosives.

The perception of speech sounds in noise also depends on the noise characteristics. Hant and Alwan (2000) examined the perceptual confusion of synthetic plosive consonants in noise and found that there was a 5 to 10 dB drop in threshold SNRs between the perceptually-flat and speech-shaped noise, suggesting that adult native English listeners might be using high frequency cues to discriminate plosives in speech-shaped noise, but that those cues were unavailable in white noise. Nittrouer *et al.* (2003) showed clear differences in adults' perception of consonants in white versus speech-shaped noise, while there was no difference in children's perception. In a recent study of English phoneme confusions in multispeaker babble noise by native and non-native listeners, Cutler *et al.* (2004) showed that although both language background and noise were significant factors in perception, the interaction of these two factors was not significant.

Few studies have examined physical measures that could account for the changes in the perception of features or sounds in the presence of background noise (Hant and Alwan, 2000, 2003; Soli and Arabie, 1979). Soli and Arabie (1979) analyzed the consonant confusion data from Miller and Nicely (1955) and suggested (qualitatively) that consonant confusion data could be better explained by the acoustic properties of the consonants than by phonetic features. In (Hant and Alwan, 2000, 2003), the authors developed a general, time/frequency detection computational model to predict human speech perception in noise. The model predicted well the discrimination of synthetic voiced plosive CV syllables in perceptually-flat and speech-shaped noise. Their perceptual experiments and model showed that formant transitions are more perceptually salient than the plosive burst is in noise.

Previous literature has focused mostly on the /Ca/ context. Notable exceptions include Klatt (1975), Summerfield and Haggard (1977), and Hant and Alwan (2003). The present study examines how vowel context (/a/, /i/, or /u/) affects the relationship between the acoustic properties of the speech signal and results from perceptual experiments conducted in the presence of additive white Gaussian noise. First, measurements of several acoustic properties from a set of CV utterances were made (in quiet) and analyzed for possible voicing cues. Second, perceptual experiments were conducted using the speech tokens mixed with varying amounts of background noise. Finally, the acoustic measurements were examined in conjunction with the results from the perceptual experiments to determine which cues could possibly account for the perception of voicing in noise. The hypothesis, implicit in this paper, is that the perception of voicing in plosives in noise is affected by an interaction between SNR on the one hand, and factors that affect the acoustic characteristics of the plosive release on the other hand; such factors include talkers' gender, place of articulation, and vowel context. Specifically, post-hoc correlation analyses will demonstrate that tokens whose voicing is specified only by VOT are well recognized at high but not low SNRs, while tokens whose voicing is specified by both VOT and F1 onset frequency are well recognized at both high and low SNRs. That is, CVs possessing strong voicing cues in the vowels, such as the first formant frequency and VOT, yield better voicing perception in noise than those not having strong F1 onset frequency cues. In addition, we also analyzed the importance of other acoustic cues that have been hypothesized in the literature to be important for voicing classification (such as F0, the burst, and voicebar).

II. METHODS

A. Stimuli

The stimuli consisted of isolated, naturally-spoken consonant-vowel utterances (CVs), each comprising of a plosive from the set {/b/, /p/, /d/, /t/, /g/, /k/} followed by a vowel from the set {/a/, /i/, /u/}, for a total of 18 syllables. The speech signals were recorded in a sound attenuating room using a headset microphone, and were then sampled (sampling rate of 16 kHz with a 16 bits per sample representation). Four talkers, all native speakers of American English between the ages of 18 and 36 years, were recorded. Two talkers were male, and two were female. Each talker produced eight repetitions of each CV, but only four tokens were used for the current study (the first three tokens and the last one were discarded), resulting in a total of 16 tokens per CV. Syllables were sorted in voiced/voiceless pairs (such as /ba/ and /pa/), such that place of articulation and vowel context were identical, and the two syllables in each pair differed only in the voicing dimension. Thus, there were a total of nine CV pairs (see Table I).

The masking noise used in the perceptual experiments was a 1250-ms white Gaussian noise. At the beginning of each experimental session, 32 Gaussian noise sources were generated; each noise sample was generated randomly and modeled with a Gaussian distribution. During the presentation of each stimulus, a noise masker was randomly selected from the 32 Gaussian noise sources. The SNR was defined as the ratio of the peak root mean square (RMS) value in the CV to the RMS value of the noise token [$20\log10(peak\_RMS_{CV}) - 20\log10(RMS_{noise})$]. The first term occurred in the vowel part for most of the CV tokens. The peak RMS energy of a token was computed using a 30-ms rectangular window. The RMS energy of the noise was based on the entire noise segment. Hence, the SNR did not depend on the duration of the speech token.

B. Acoustic measurements

For acoustic measurements, tokens were first decimated to an 8 kHz sampling rate and then pre-emphasized with a coefficient of 0.97. All tokens were normalized such that the peak amplitude of the entire sampled waveform was set to the same level. Acoustic measurements were made for the speech tokens without noise.

The voicebar, burst, VOT, and F0 measurements were made by visually inspecting the time waveforms and wideband spectrograms of the tokens using the software CoolEdit Pro. Wideband spectrograms were calculated using a 6.4-ms Hamming window with frame shift of one sample (see Fig. 1). The voicebar/prevoicing was defined as the periodic low-frequency energy before the consonant release. The burst was defined as the short segment characterized by a sudden, sharp vertical line in the spectrogram. The segmentation of a burst was performed visually by examining both the waveform and spectrum in CoolEdit Pro. If multiple bursts were present, the burst duration (in ms) was measured from the beginning of the first burst to the end of the last. The maximum absolute amplitude of the sampled waveform in the burst segment was defined as burst peak amplitude (in dB). VOT duration was measured from the end of the burst to the beginning of the vowel, which was also the beginning of the first waveform period. VOT peak amplitude (in dB) was the maximum absolute value of the sampled waveform in the segment. F0 at the onset of the vowel was calculated from the inverse of the first pitch period, measured from peak to peak. The steady-state F0 frequency was similarly calculated from the length of a pulse measured at approximately 100 ms after the onset of the vowel. An F0 frequency change was then derived from the F0 onset and steady-state frequencies (as a drop in F0).

Formant frequency measurements were made from the time waveforms, spectrograms, LPC spectra [Fig. 2(a)], and short-time DFT spectra [Fig. 2(b)] using Matlab. A 20 ms (for tokens from male talkers) or 15 ms (for tokens from female talkers) Hamming window was applied to define an analysis segment. Different Hamming window lengths were used because of the differences in the fundamental frequency between male and female talkers. To obtain a spectrum, each segment was zero-padded for a 1024-point FFT analysis, and the frame shift was half the Hamming window length (i.e., 10 ms for male talkers and 7.5 ms for female talkers). For an LPC analysis, no zero padding was applied, the frame shift was 2.5 ms for all talkers, and the LPC order was between 8 and 12, depending on which gave better results. The vowel was defined to begin when the vocal folds began to vibrate after aspiration. Vowel measurements included the first three formants (F1, F2, and F3). The three formants were located by examining the LPC spectra [Fig. 2(a)], FFT spectra [Fig. 2(b)], and spectrograms. Three landmark points were defined for each formant: formant onset, offset, and steady-state [Fig. 2(c)]. The onset of the vowel, chosen manually, was defined as the center point of the frame that exhibited the following characteristics: an abrupt increase in the total energy of the frame and a sudden change in the spectral properties, particularly the introduction of a sharp F1 spectral peak. F1 always began in this frame, but not necessarily F2 or F3. Accordingly, F2 and F3 onsets were defined by examining the sudden spectral change in F2 and F3 frequency range, respectively. The end of a formant transition (offset), chosen automatically, was defined as the frame during which the rate of change of the formant frequency fell to less than 5 Hz per 2.5 ms, and the average rate of change for the next 12.5 ms was also less than 5 Hz per 2.5 ms [Kewley-Port, 1982; see Fig. 2(d)]. The steady-state point was centered at 95 ms after the onset, and the steady-state measurements were averaged over five frames. At the formant transition onset, offset, and steady-state points [Fig. 2(c)], the following parameters were recorded: time (relative to the beginning of the utterance and measured in ms), formant frequency (from the LPC spectrum and measured in Hz), and formant amplitude (from the LPC spectrum and measured in dB). From these measurements, formant transition duration, frequency and amplitude change, and slope were calculated for each of the first three formants. Formant transition duration was defined as the time difference between the formant transition offset and onset. Using this methodology, the duration and slope of formant transitions were not necessarily the same for the three formants. Formant frequency and amplitude changes were measured between the formant transition onset and steady-state (which was a more stable reference than the offset). Formant slope was calculated from formant transition onset and offset measures, since it required temporal information. These definitions imply that the formant slope does not necessarily relate to the ratio of formant frequency change to formant transition duration. Table II lists the acoustic parameters that were measured for each token.

C. Perceptual experiments

1. Participants

Listening experiments were conducted with four paid subjects (different from those who participated in the recording session): two males and two females, all native speakers of American English aged between 18 and 36 years, who passed a hearing test and participated in a training session.

2. Procedure

Experiments took place in a sound attenuating room. Digital speech stimuli were played via an Ariel Pro Port 656 board Digital-to-Analog converter (16 bits at a rate of 16 kHz). The resulting analog waveforms were amplified by a Sony 59ES DAT recorder. The amplified analog waveforms were then presented binaurally with identical waveforms to each ear via Telephonics TDH49P headphones. The system was calibrated within 0.5 dB (from 125 to 7500 Hz at third octave intervals) using a 6-cc coupler and a Larson Davis 800B sound level meter prior to each experiment. The sound level meter was set to the “A” weighting scale with a slow response. Preamp levels and digital internal level offsets were set to establish a relation between the digital RMS energy and the actual SPL level.

Each signal was played at 60 dB SPL, and the noise level was adjusted. The SPL of the speech signals were set based on its peak RMS energy on a 30-ms rectangular window. The SPL of the white Gaussian noise was adjusted based on its RMS energy to result in different SNRs. The speech signal was added to a 1250-ms noise (or silence) segment such that it was centered in the middle of the segment with equal duration of noise (or silence) before and after the speech signal. Sessions lasted for no longer than two hours, and subjects were instructed to take at least one break every hour. No feedback was given at any time.

The experiments were of the two alternate forced choice (2AFC) type. To counterbalance the effects of talker and token order, testing was administered in blocks of 64 pseudo-randomized items (32 tokens x 2 presentations) for each CV pair (for example, /ba,pa/). When an utterance was played, participants were asked to label the sound heard as either the voiced or voiceless consonant (e.g., /b/ or /p/). A computer program was developed to record participants' responses from their keyboard inputs. The test was then repeated at different SNR levels. The order of SNR conditions was –15 dB, -10 dB, -5 dB, 0 dB, 5 dB, 10 dB, and quiet (same order for all listeners). The CV pairs were presented in the order of /ba,pa/, /bi,pi/, /bu,pu/, /da,ta/, /di,ti/, /du,tu/, /ga,ka/, /gi,ki/, and /gu,ku/.

III. VOICING CLASSIFICATION BASED ON ACOUSTIC MEASUREMENTS

The acoustic measurements were analyzed using logistic regression (Benki, 2001; Menard, 1995; Nearey, 1997), where the quiet speech tokens were classified as either voiced or voiceless according to a single acoustic property measured without the addition of the white

Gaussian noise. A separate logistic regression model was applied to each acoustic variable for each CV pair:

$$\log[prob/(1\text{-}prob)] = \alpha + \beta Mea + e, \tag{1}$$

where *prob* is the probability of a token being voiceless, α is a constant, β is a weighting coefficient, *Mea* is one acoustic feature, and *e* is the error term. Table III lists the results in terms of percent correct classification based on logistic regression using the tokens from all talkers. If the gender effect was significant (Bonferroni-corrected $p < 0.050$), a dummy variable representing gender difference was added in the logistic regression for a better fitting:

$$\log[prob/(1\text{-}prob)] = \alpha + \beta Mea + \gamma G + e, \tag{2}$$

where γ is a weighting coefficient, and *G* represents gender (1 for male and 0 for female). The significance of gender effect was determined by

$$LR[2] = (-2LL_{\mathrm{MF}}) - (-2LL_{\mathrm{M}}) - (-2LL_{\mathrm{F}}), \tag{3}$$

where the likelihood ratio *LR* is a distributed chi-square with 2 degrees of freedom, and LL_{MF}, LL_{M}, and LL_{F} are the log likelihood functions derived from Equation (1) using data pooled from all talkers, male talkers, and female talkers, respectively. For each token, the plosive was either voiced or voiceless, and thus the *prob* in Equations (1) and (2) was either 0 or 1. After logistic regression, $\alpha + \beta Mea = 0$ for Equation (1) or $\alpha + \beta Mea + \gamma G = 0$ for Equation (2) was used for classification, and results were compared against ideal classification to obtain the percent correct scores. In Table III, the acoustic measures are listed according to their percent correct scores for each CV pair. Only acoustic measures with 90% or higher correct classification are listed.

Prior studies (Liberman *et al.*, 1958; Lisker and Abramson, 1964) have shown that the **VOT duration** (votD) is a significant acoustic cue for voicing in plosives. In the present study, the VOT duration proved to be the single best acoustic property for classification of voicing for the quiet speech tokens (Table III). All tokens could be correctly classified based solely on the VOT duration (see Fig. 3) with voiceless plosives having a significantly longer VOT duration than voiced plosives [$t(286) = 44.470$, $p = 0.000$]. The VOT duration was less than 20 ms for most of the voiced tokens, and over 45 ms for most of the voiceless tokens. Thus, a VOT duration of 30 ms could be considered as a voicing boundary, except for the /gi,ki/ pair, which had a VOT duration boundary of 40 ms. Another VOT property, **VOT peak amplitude** (votpA), which can be considered a measurement of the loudness of aspiration, produced relatively high percentage of correct identifications for the alveolar pairs /di,ti/ and /du,tu/.

Voicing classification based solely on the **duration of the voicebar** (vbD) did not result in a high percentage of correct classification results for all talkers, whereas it was an important factor in voicing classification for seven out of the nine CV pairs from female talkers. While the presence of a voicebar is almost certainly indicative of a voiced utterance, the converse does not hold. For all but one token, voiceless tokens showed no voicebar. Only two of the 72 voiced tokens spoken by the male talkers showed a voicebar, while most (65 out of 72) of the voiced tokens spoken by the female talkers showed a voicebar. This, however, does not necessarily indicate that gender plays an important role in whether or not there is a voicebar. Since there were only four talkers, the results can be attributed to individual differences. Previous studies have shown that some talkers tend to produce a voicebar while others do not (Lisker and Abramson, 1964).

The **loudness of the burst**, including intensity and duration, has been cited as a possible cue for voicing (Lisker and Abramson, 1964; Repp, 1979). However, the present study shows that classification based solely on the burst properties did not produce high percent correct classification. The **loudness of the burst** (peak amplitude; bstpA) was about 2.3 dB higher on average for voiceless tokens than for their voiced counterparts, but the range of values between voiced and voiceless tokens overlapped greatly. Burst measurements were short (about 5 ms) for both voiced and voiceless labial tokens. In general, **burst duration** (bstD) was more related to place of articulation and vowel context than to voicing distinctions: Velars appeared to have longer burst durations than labials and alveolars (about 10 ms longer in the /a/ context and 20 ms longer in the /i/ and /u/ contexts).

Summerfield and Haggard (1977) showed that F1 onset frequency is important in the context of a low vowel (e.g., /a/), but not in high vowel contexts (e.g., /i/ and /u/); similar results are shown in Table III. Table III shows that the F1 transition was an excellent classifier for voicing only in the /a/ context. All /Ca/ tokens could be correctly classified based solely on their **F1 onset frequency** (F1b, see Fig. 4). In the /a/ context, voiceless tokens usually showed an F1 onset frequency that is higher than 600 Hz, while voiced tokens were below that [mean absolute difference = 379 Hz, $t(94) = 15.281$, $p = 0.000$]. For tokens in the /i/ and /u/ contexts, however, the F1 onset frequency was in about the same range for both voiced and voiceless tokens with mean absolute differences of 43 and 33 Hz, respectively. As mentioned earlier, the F1 transition has been considered an important property for voicing classification (Stevens and Klatt, 1974). Accordingly, the **F1 frequency change** (F1df) was significantly different for voiced and voiceless tokens only in the /a/ context [mean absolute difference = 360 Hz, $t(94) = 15.915$, $p = 0.000$]. For tokens in the /i/ and /u/ contexts, the F1 frequency change was in about the same range for both voiced and voiceless tokens with mean absolute differences of 42 and 32 Hz, respectively. However, given the covariation between F1b and F1df and the results by Summerfield and Haggard (1977), F1 onset frequency (F1b), instead of F1 frequency change, was emphasized in this study. The percent correct classification from the **F1 slope** measurements (F1sl) was not as high as the onset frequency or frequency change of F1.

For voiceless plosives, a large portion of the formant transition occurs during the VOT and appears in the aspiration noise. Thus, voiceless tokens would exhibit short transition durations (in the vowel). Voiced plosives, on the other hand, are voiced throughout the transition and should therefore have longer formant transition durations. In this study, however, **F1 transition duration** (F1D) was not an accurate classifier for voicing, except for /da,ta/ and /di,ti/. **F1 amplitude at the end of transition** (F1eA) produced better voicing classification for the /a/ context than for the /i/ and /u/ contexts, but only the /ga,ka/ pair received above 90% correct classification based on F1eA. The **F1 amplitude change** measurements (F1dA) resulted in poor classification. **F2 and F3 measurements** were generally poor classifiers for voicing, while notable exceptions were F2 and F3 measurements for /da,ta/, /ga,ka/, /bi,pi/, and /di,ti/ (Table III).

On average, **F0 onset frequency** (F0b) was about 20 Hz higher for voiceless tokens than for voiced ones; however, the overlap between voiced and voiceless tokens was considerable. Similarly, **F0 frequency change** (F0df, i.e., an F0 drop) tended to be slightly greater (15.7 Hz) for voiceless tokens than for voiced ones but with a large overlap. Voicing classification based solely on F0b and F0df produced a high percentage of correct results only for /gi,ki/ (97% and 94%, respectively) with a significant gender effect (Bonferroni-corrected $p < 0.050$).

Note that in Table III, when male and female token differences were significant, separate thresholds were used. This was achieved by adding gender as a variable in logistic regression [see Equation (2)]. Therefore, high percent correct classification can still be achieved even when there is a significant male and female token difference.

In summary, the VOT duration was the single-best classifier in voicing for all nine CV pairs. The properties of the first formant ***frequency*** (F1 onset frequency, F1 frequency change, and F1 slope) were good classifiers only for the three /Ca/ pairs. Because the /Ca/ pairs exhibited stronger F1 cues than the /Ci/ and /Cu/ pairs, we expect them to have better voicing perception in noise than the /Ci/ and /Cu/ pairs. F1 transition duration, which is a ***temporal*** property rather than a ***frequency*** property, was a good classifier only for /da,ta/ and /di,ti/. Properties of F2 and F3 were good classifiers for four CV pairs. Voicebar duration was a reliable classifier for seven out of the nine CV pairs from the female talkers.

IV. PERCEPTUAL RESULTS

A. Percent correct classification

The percentage of correct voicing judgments was computed and listed as a function of SNR, place of articulation, and vowel context. The percent correct values shown in Table IV were calculated using all the data collected from the perceptual experiments, including all listeners and all talkers. Each data entry thus represents 256 responses (four talkers x four listeners x four recordings of each CV x two presentations x two consonants).

The listeners appeared to have had a particularly difficult time classifying the /bu,pu/ pair (with 89% correct voicing judgments) even when the SNR was 10 dB. However, for CV pairs other than /bu,pu/, the percent correct of voicing judgments was 93% or above when the SNR was 10 dB. Among the nine CV pairs, /da,ta/ yielded the best voicing judgment performance (96% correct) when the SNR was –5 dB. For SNRs of –10 dB and below, voicing judgments for all nine pairs were dramatically affected by noise (below 70% correct).

A four-way repeated measures analysis of variance (ANOVA) (gender x place of articulation x vowel x SNR) was used to analyze the perceptual results after an arcsine transformation was applied. The effects of vowel context [$F(2,6) = 35.254$, $p = 0.000$], SNR [$F(6,18) = 341.273$, $p = 0.000$] , place of articulation [$F(2,6) = 14.205$, $p = 0.005$], and gender of the talker [$F(1,3) = 13.066$, $p = 0.036$] were significant. The mean percent correct for vowel context, SNR, and place of articulation is plotted in Fig. 5. As expected, the vowel /a/ context yielded higher percent correct classification than the /i/ context [$F(1,3) = 35.573$, $p = 0.009$], but there was no significant difference between the /i/ and /u/ contexts [$F(1,3) = 1.473$, $p = 0.312$; see Fig. 5(a)]. Also as expected, the number of correct responses decreased as the SNR level decreased ($p < 0.050$), and decreased more quickly after the SNR was reduced below –5 dB. Figure 5(b) shows that most of the nine CV pairs had 100% or close to 100% correct voicing judgments in the absence of noise. Voicing discrimination degraded significantly between 0 and –10 dB SNRs. When the SNR was –15 dB, the percent correct of voicing judgments was about 50%, which is chance performance. This is consistent with Miller and Nicely's (1955) study that indicated voicing was still discriminable at a SNR of –12 dB.

As for place of articulation [Fig. 5(c)], percent correct scores for alveolars were significantly higher than those of bilabials [$F(1,3) = 14.706$, $p = 0.031$] but not significantly different from those of velars [$F(1,3) = 0.002$, $p = 0.964$].

The interactions between gender and SNR [$F(6,18) = 6.452$, $p = 0.001$], between vowel and SNR [$F(12,36) = 16.867$, $p = 0.000$], between place of articulation and SNR [$F(12,36) = 2.945$, $p = 0.006$], and between place of articulation and vowel [$F(4,12) = 4.067$, $p = 0.026$] were significant, while the interactions between gender and place of articulation [$F(2,6) = 0.044$, $p = 0.957$] and between gender and vowel [$F(2,6) = 0.127$, $p = 0.883$] were not significant. The gender and SNR interaction, vowel and SNR interaction, and place of articulation and vowel interaction effects are shown in Fig. 5. The interactions of more than two factors (gender, place of articulation, vowel, and SNR) were not significant ($p > 0.050$) except for place of articulation x vowel x SNR interaction [$F(24,72) = 2.027$, $p = 0.011$]. Figure 5(d) shows an interesting gender and SNR interaction effect. Voicing judgments for stimuli from male talkers degraded steadily with decreasing SNR, whereas for female talkers, the performance of voicing discrimination gradually degraded with decreasing SNR when the SNR was above –5 dB and then degraded dramatically for lower SNRs. A possible reason for the gender x SNR interaction may be due to the differences in F1 onset frequency between voiced and voiceless CVs for male and female talkers. Generally speaking, female talkers have shorter vocal tracts than male talkers, and thus produce higher formant frequencies including F1 onset frequencies for voiceless CVs, which were shown to influence voicing judgments (Summerfield and Haggard, 1977). For example, female talkers produced higher F1 onset frequencies in voiceless /Ca/ and /Ci/ tokens than male talkers.

Figure 5(e) shows that the /a/ context was the most robust one, most likely because the F1 differences between the voiced and voiceless plosives are prominent only in the /a/ context. These differences may have helped listeners perceive voicing more robustly. The figure also shows that voicing judgments for stimuli in the /i/ and /u/ contexts degraded steadily with decreasing SNR, whereas in the /a/ context, the performance of voicing discrimination gradually degraded with decreasing SNR when the SNR was above –5 dB and then dropped rapidly for lower SNRs.

Figure 5(f) shows the vowel and place of articulation interaction effect. For example, voicing classification was better for velars than for bilabials and alveolars in the /i/ and /u/ contexts.

In summary, the vowel /a/ context yielded higher percent correct voicing classification than the /i/ and /u/ contexts. Velars yielded higher percent correct classification than bilabials and alveolars in the /i/ and /u/ contexts. Voicing judgments for stimuli from male talkers (or the /i/ and /u/ contexts) degraded steadily with decreasing SNR, whereas for female talkers (or the /a/ context), voicing discrimination degraded dramatically when the SNR was below –5 dB.

B. Threshold SNRs for voicing classification in noise

A traditional approach for investigating voicing distinction has been to explore the perceptual boundaries between phonetic categories (e.g., VOT, F1, VOT x F1, etc.). In order to analyze how the acoustic properties account for the perceptual results, a single value for each CV pair was needed to represent the robustness of that CV pair in the presence of noise. That value, or threshold, was computed along the SNR continuum. The data for the nine CV pairs were arranged into plots as shown in Fig. 6 where percent correct is plotted versus SNR. A sigmoid was then fit to each plot (excluding data for the quiet condition) and described by the following equation:

$$y = c + \frac{d-c}{2} \cdot \left(1 - \frac{1 - e^{\frac{x-b}{a}}}{1 + e^{\frac{x-b}{a}}} \right), \tag{4}$$

where x represents SNR and y represents percent correct. The parameters a (rate of change of the sigmoid), b (halfway point), c (bottom of the sigmoid), and d (top of the sigmoid) were varied systematically to obtain the best fit sigmoid by minimizing the mean squared error. Note that in a regular logistic regression model, *c* and *d* are fixed to 0 and 1, respectively, and the model fitting is based on the maximum likelihood principle. Theoretically, the percent correct versus SNR curve should be flat at about 100% for very high SNR levels, flat at about 50% (chance performance) for very low SNR levels, and monotonically increasing in between. The characteristics of a sigmoid match these requirements, and it was therefore chosen as the curve that best represented the data. From the best fit sigmoid, the threshold SNR level corresponding to 79% correct responses was obtained (Chen, 2001; Hant, 2000; Levitt, 1971). Thus, a single threshold SNR value for each of the nine pairs of voiced/voiceless CVs was calculated to represent the perceptual robustness of that pair. A lower threshold SNR corresponded to better perceptual results (more robust to noise).

Threshold SNR levels for all CV pairs corresponding to 79% correct responses averaged over all talkers are shown in Fig. 6. These thresholds are separated by gender of the talker in Fig. 7. Figure 6 shows that for all talkers, threshold SNRs for CVs in the /i/ and /u/ contexts were lower (more robust) for velars (ranging from –5.2 to –3.8 dB) than for labials and alveolars (ranging from –3.0 to –1.6 dB). CVs in the /a/ context (Group 1 in Fig. 7) appeared to be significantly more robust in noise than those in the /i/ [$t(10) = 4.846$, $p = 0.001$] and /u/ [$t(10) = 5.556$, $p = 0.000$] contexts. This effect agrees with the results shown in Fig. 5(e). In most cases, the tokens from female talkers were more perceptually robust than those from male talkers [paired $t(8) = 3.475$, $p = 0.008$; much lower threshold SNRs (by over 3 dB) for /di,ti/, /bu,pu/, and /du,tu/ pairs; labeled as Group 2 in Fig. 7]. Such effect is consistent with that shown in Fig. 5(d).

Voicing classification for /gi,ki/ was more robust than those for /bi,pi/ and /di,ti/ (Fig. 6). Threshold SNRs for /gu,ku/ were lower than those for /bu,pu/ and /du,tu/ when the talker was a male, while the values were very close when the talker was a female (Fig. 7).

In summary, CVs in the /a/ context appeared to be more robust in noise than those in the /i/ and /u/ contexts. Threshold SNRs for CVs in the /a/ context were lower (more robust) for alveolars than for labials and velars. Threshold SNRs for CVs in the /i/ and /u/ contexts were lower for velars than for labials and alveolars. In most cases, the tokens from female talkers were more noise robust than those from male talkers. This can be attributed in part to differences in formant frequencies and F0 frequencies in addition to the existence of a voicebar for 65 out of the 72 voiced CV tokens from the female talkers.

V. CORRELATIONS BETWEEN THRESHOLD SNR VALUES AND ABSOLUTE ACOUSTIC DIFFERENCES OF THE MEANS

Correlations were computed between the nine threshold SNR values from the perceptual experiments and the absolute differences of the mean values of a measured acoustic property for the nine voiced/voiceless pairs (three places of articulation x three vowel contexts). The mean value of each acoustic measurement for every CV syllable was calculated from 16 tokens (four talkers x four tokens of the same syllable). The correlation is defined as

$$r = corr(\left|\overline{Mea}_{v} - \overline{Mea}_{vl}\right|, 10 - SNR_t) = -corr(\left|\overline{Mea}_{v} - \overline{Mea}_{vl}\right|, SNR_t), \tag{5}$$

where *corr* represents the Pearson correlation function, *v* represents voiced tokens, *vl* represents voiceless tokens, *Mea* represents one type of acoustic measurement, the bar over *Mea* represents the mean operation, SNR_t represents the threshold SNR values, and 10 - SNR_t indicates how much the threshold SNRs were below 10 dB. The absolute difference of the means was chosen under the assumption that the greater the distance between the means, the larger the separation between the associated distributions, and thus the more distinct the acoustic property in question is for that voiced/voiceless CV pair. Thus, if an acoustic property is an important cue for voicing, then a larger absolute difference between the means would correspond to better performance (a lower threshold SNR), while a smaller absolute difference between the means would correspond to poorer performance (a higher threshold SNR). Correlation using the absolute difference of the means has several shortcomings, but it provides a simple method for obtaining a numerical measure of how well the perceptual results correlate with acoustic properties. Pearson product correlation coefficients were obtained only for those acoustic properties that appear in Table III.

Table V lists the results of correlating threshold SNRs (using data from all talkers) with the absolute differences of the means of several acoustic properties. The first formant measurements yielded the highest and most significant correlations, **F1 onset frequency** (F1b) and **F1 transition frequency change** (F1df) having correlation coefficients of 0.86 (Bonferroni-corrected *p* = 0.048) and 0.87 (Bonferroni-corrected *p* = 0.044), respectively. A close examination of F1 measurements indicated that the absolute differences of means of F1b and F1df were perfectly correlated [*r*(7) = 0.999, *p* = 0.000]. Given such covariation between F1b and F1df, we may interpret the results of Summerfield and Haggard (1977) to mean that the F1 onset frequency, rather than the F1 frequency change, was the most important cue for voicing perception at low SNRs. **F1 amplitude at the end of transition** (F1eA) also yielded a relatively high correlation coefficient of 0.84 (Bonferroni-corrected *p* = 0.085). Summerfield and Haggard (1977) showed that the F1 amplitude was not an important cue for voicing in quiet conditions. In this study, in quiet, F1eA signaled voicing prominently only for /ga,ka/ (91% correct; Table III). In noise, however, F1eA might become important in voicing perception as demonstrated by a high correlation coefficient (Table V). **F1 slope** (F1sl) and **F1 transition duration** (F1D) also showed relatively high correlation coefficients of 0.79 and 0.78, respectively. These correlations occurred because the /a/ sound is the most robust in noise, and because simultaneously the /a/ context has the largest absolute F1 onset frequency difference for the voiced and voiceless syllable-initial plosives.

VOT duration (votD), on the other hand, showed a highly negative correlation with the perceptual results [*r*(7) = -0.85, Bonferroni-corrected *p* = 0.060]. A negative correlation coefficient indicates a larger distance between the means correlated with a higher (worse) threshold SNR. The VOT duration cue is easily corrupted by noise since the burst and aspiration parts are of low amplitude compared to the vowel onset.

F0 measurements (F0b and F0df) yielded negative but small-amplitude correlation coefficients. Thus, F0 cues might be easily disrupted in the voicing perception in noise.

Some of the correlations were further examined as shown in Fig. 8; the panels show threshold SNRs versus the absolute differences of the means for several acoustic properties. Numbers inside the panels are the correlation coefficients between the SNRs and the absolute differences of the means. For example, the absolute differences of the means of F1 onset frequency (F1b) and F1 amplitude at the end of transition (F1eA) were large for the /a/ context and small for the /i/ and /u/ contexts. Thus, the high correlation for F1 onset frequency indicated that the threshold SNRs were much lower for /Ca/s than for /Ci/s and /Cu/s, which is consistent with Fig. 6. As shown in Fig. 8, the highly negative correlation of threshold SNRs with VOT duration measures occurred because the CV pairs that were best separated by the VOT duration (all /d,t/ and /b,p/ pairs) also exhibited the highest (worst) threshold SNRs. In the speech production process under quiet recording conditions, there is a trading relation between VOT duration and F1 onset frequency (Summerfield and Haggard, 1977). That is, CV pairs with larger absolute VOT duration difference are produced with smaller F1 onset frequency differences, and vice versa (as is evident in Fig. 8). Because the VOT duration is masked at a higher SNR than the F1 onset frequency, the CV pairs with large VOT duration differences (and correspondingly small F1 onset frequency differences) are mistakenly perceived at a higher SNR than the CV pairs with small VOT duration differences (and correspondingly high F1 onset frequency differences).

In summary, the F1 onset frequency measurements yielded a significant positive correlation with voicing perception in noise. Because the /a/ context resulted in larger absolute differences in the F1 onset frequency measurements between the voiced and voiceless plosives than the /i/ and /u/ contexts, the voicing judgments in the /a/ context were more robust than those in the /i/ and /u/ contexts. VOT duration, on the other hand, showed a highly negative correlation.

VI. GENERAL DISCUSSION

The present study examines the acoustic correlates and perception in noise of voicing in naturally-spoken syllable-initial plosives. It is important to determine how human listeners trade off various acoustic cues (e.g., VOT and F1 onset frequency) at different SNR levels. The results demonstrate that the perception of voicing in plosives in noise is affected by an interaction between SNR on the one hand, and factors that affect the acoustic characteristics of the plosive release on the other hand; such factors include talkers' gender, place of articulation, and vowel context. Specifically, post-hoc correlation analyses suggest that VOT duration is the cue most predictive of voicing decisions in quiet conditions, but VOT duration appears to be masked at low SNRs, with a contrasting result that F1 onset frequency is a better voicing cue at low SNRs.

Specifically, in quiet conditions, all of the nine CV pairs were correctly classified at or near 100%. The VOT duration proved to be the single best acoustic property for voicing classification in syllable-initial plosives, and all tokens could be correctly classified based on their VOT durations. Furthermore, VOT duration boundaries differed across places of articulation, which agrees with results reported by Lisker (1975). VOT peak amplitude produced relatively high percent correct classification for two CV pairs. The F1 onset frequency was an excellent classifier for voicing only in the /a/ context. All /Ca/ tokens could be correctly classified based solely on their F1 onset frequencies or F1 frequency changes that covaried with F1 onset frequencies. Tokens without a prominent F1 onset frequency cue can be correctly identified, indicating that the F1 onset frequency is not a *necessary* cue for voicing, at least not in the /i/ and /u/ contexts. Obviously, there are multiple and redundant cues for voicing. In the presence of noise, the relative roles of different acoustic cues appear to change as a function of SNR.

For noisy speech, as expected, the voicing distinction was more difficult to perceive as the SNR level decreased. Listeners could still make correct voicing judgments even when the SNR level was –10 dB. However, for a SNR of –15 dB, listeners' responses were equivalent to random guesses (chance performance). These results are consistent with Miller and Nicely's (1955) study, which concluded that voicing is robust to noise, while place of articulation is not.

For voicing classification in noise, vowel effect was significant in the sense that voicing judgments were always more accurate in the /a/ context than in the /i/ and /u/ contexts (threshold SNRs were lower for /Ca/s than for /Ci/s or /Cu/s). In addition, voicing classification in /Ca/ syllables degraded gradually with decreasing SNR and then degraded rapidly for SNRs lower than –5 dB. In contrast, voicing classification in the /i/ and /u/ contexts degraded steadily across different SNR levels. Correlation analyses showed that the VOT duration contributed negatively to these threshold SNR differences. Instead, the F1 onset frequency differences between voiced and voiceless CVs for the nine pairs were highly correlated with threshold SNRs from the perceptual experiments. The highly negative correlation for the VOT duration and the significant positive correlation for the F1 onset frequency agree with the trading relationship between VOT duration and F1 onset frequency reported in (Summerfield and Haggard, 1977). The positive correlation for F1 onset frequency indicates that the F1 onset frequency is more important for the perception of voicing at low SNRs than the VOT duration. The high positive correlation occurred because the differences in F1 onset frequency between voiced and voiceless tokens were large and discriminative in the /a/ context, but not as large or discriminative in the /i/ or /u/ context (a perfect voiced/voiceless indicator only for CV syllables in the /a/ context). This is understandable because the information for the plosive consonants is not limited to a single time instant (Liberman *et al.*, 1967), but extends to the following vowel through coarticulation. The vowel /a/ has a high F1 onset frequency for a voiceless token and a low F1 onset frequency for a voiced token, and thus the difference in F1 onset frequency is prominent. The range of F1 onset frequency (or amplitude) change is much larger in the /a/ context than in the /i/ and /u/ contexts. For plosives, the release burst and aspiration consist of noise, which is weaker in amplitude than the vowel formants and is easily corrupted by noise, especially broadband noise. Therefore, in noisy speech, F1 onset frequency is a more dominant cue for voicing than VOT, and voicing perception is dependent on vowel context.

Similarly there was a significant gender and SNR interaction effect. Threshold SNR values for /di,ti/, /bu,pu/, and /du,tu/ with female talkers were significantly lower than those for their male counterparts' tokens (by over 3 dB). Such an interaction effect could be attributed to the fact that the properties of fundamental frequency and formant frequencies (or transitions) for female talkers are different from those for male talkers. For example, the absolute differences of means of F1 onset frequency in the /Ca/ voiced/voiceless pairs were larger for female talkers than for male talkers. In addition, a voicebar was present in the majority of voiced tokens by the female talkers, while only two out of 72 voiced tokens by the male talkers showed a voicebar. Further, for female talkers, distinct differences in F1 transition duration between the voiced and voiceless tokens were present in /di,ti/, /bu,pu/, and /du,tu/, but not for /gu,ku/. Some of the /Ci/s and /Cu/s from the male talkers were well classified using the F1 transition duration measurements, but they did not have low perceptual threshold SNRs. This may be due to their short F1 transitions. A longer F1 transition is more easily detectable, particularly in noise (Hant, 2000). This is, however, only useful if a voicing cue is present in the F1 transition. Therefore, for noisy speech, better voicing classification results were obtained only if the F1 transition contained distinct differences between the voiced and voiceless CVs, and if the F1 transition duration was relatively long (over 10 ms). Furthermore, for /di,ti/, /bu,pu/, /du,tu/, and /gu,ku/, the F1 transition duration was much longer for tokens generated by the female talkers (6 to 20 ms) than by the male talkers (1 to 10 ms), for both the voiced and voiceless CVs. This could be another explanation for the female tokens being more noise robust.

Results in this study also indicate that F0 differences between voiced and voiceless plosives are not important cues for voicing perception in additive white Gaussian noise.

Threshold SNRs for CVs in the /i/ and /u/ contexts were lower (more robust) for velars (/g,k/) than for labials (/b,p/) and alveolars (/d,t/). One possible explanation for the better performance of the velars (in the absence of a F1 transition cue) is that the differences of F0 frequency change, F0 onset frequency, and F1 amplitude at the end of transition between voiced and voiceless tokens, although not prominent, were larger for /gi,ki/ and /gu,ku/ than for /bi,pi/, /di,ti/, /bu,pu/, and /du,tu/.

In future studies, experiments will be conducted using a larger data set, with more talkers, in order to determine whether differences in gender that appeared in this study were in fact due to gender, or due to individual differences. Perceptual experiments could also be conducted using synthetic stimuli to construct acoustic continua and control interactions between the various acoustic properties (e.g., independently vary VOT, F1 onset frequency, F1 frequency change, and F1 amplitude). Given that the F1 onset frequency is an important cue for voicing perception at low SNRs, an interesting future research direction would be to investigate whether talkers will enhance the F1 onset frequency difference under noisy recording conditions to emphasize the voicing contrast.

ACKNOWLEDGMENTS

This work was supported in part by an NSF Post-Graduate Fellowship to Marcia Chen and by an NIH grant. We also thank Ann Bradlow and three anonymous reviewers for their constructive comments and suggestions. The views expressed here are those of the authors and do not necessarily represent those of the National Science Foundation and the National Institute of Health.

Allen, J. B. (**1994**). “How do humans process and recognize speech?” IEEE Trans. Speech Audio Process. **2**, 567-577.

Benki, J. (**2001**). “Place of articulation and first formant transition pattern both affect perception of voicing in English,” J. Phonetics **29**, 1-22.

Chen, M. (**2001**). “Perception of voicing for syllable-initial plosives in noise,” Master thesis, Electrical Engineering Department, University of California at Los Angeles.

Cho, T., and Ladefoged, P. (**1999**). “Variation and universals in VOT: evidence from 18 languages,” J. Phonetics **27**, 207-229.

Cutler, A., Weber, A., Smits, R., and Cooper, N. (**2004**). “Patterns of English phoneme confusions by native and non-native listeners,” J. Acoust. Soc. Am. **116**, 3668-3678.

Fitch, H. L., Halwes, T., Erickson, D. M., and Liberman, A. M. (**1980**). “Perceptual equivalence of two acoustic cues for stop-consonant manner,” Percept. Psychophys. **27**, 343-350.

Haggard, M., Ambler, S., and Callow, M. (**1970**). “Pitch as a voicing cue,” J. Acoust. Soc. Am. **47**, 613-617.

Hall, M. D., Davis, K., and Kuhl, P. K. (**1995**). “Interactions between acoustic dimensions contributing to the perception of voicing,” J. Acoust. Soc. Am. **97**, 3416.

Hant, J. (**2000**). “A computational model to predict human perception of speech in noise,” Ph.D. dissertation, Electrical Engineering Department, University of California at Los Angeles.

Hant, J., and Alwan, A. (**2000**). “Predicting the perceptual confusion of synthetic plosive consonants in noise,” in *Proceedings of Sixth Int. Conf. Spoken Lang. Proc.*, Beijing, China, pp. 941-944.

Hant, J., and Alwan, A. (**2003**). “A psychoacoustic-masking model to predict the perception of speech-like stimuli in noise,” Speech Commun. **40**, 291-313.

Hermansky, H. (**1990**). "Perceptual linear prediction (PLP) analysis for speech," J. Acoust. Soc. Am. **87**, 1738-1752.

Kewley-Port, D. (**1982**). "Measurement of formant transitions in naturally produced stop consonant-vowel syllables," J. Acoust. Soc. Am. **72**, 379-389.

Klatt, D. H. (**1975**). "Voice onset time, frication, and aspiration in word-initial consonant clusters," J. Speech Hear. Res. **18**, 686-706.

Levitt, H. (**1971**). "Transformed up-down methods in psychoacoustics," J. Acoust. Soc. Am. **49**, 467-477.

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

Liberman, A. M., Delattre, P. C., and Cooper, F. S. (**1958**). "Some cues for the distinction between voiced and voiceless stops in initial position," Lang. Speech **1**, 153-167.

Lisker, L. (**1975**). "Is it VOT or a first-formant transition detector?" J. Acoust. Soc. Am. **57**, 1547-1551.

Lisker, L., and Abramson, A. S. (**1964**). "A cross-language study of voicing in initial stops: Acoustical measurements," Word **20**, 384-422.

Lisker, L., and Abramson, A. S. (**1970**). "The voicing dimension: Some experiments in comparative phonetics," in *Proc. Sixth Int. Congress of Phonetic Sci.*, Prague, 1967 (Academia, Prague), pp. 563-567.

Lisker, L., Liberman, A. M., Erickson, D. M., Dechovitz, D., and Mandler, R. (**1977**). "On pushing the voice onset-time (VOT) boundary about," Lang. Speech **20**, 209-216.

Massaro, D. W., and Oden, G. C. (**1980**). "Evaluation and integration of acoustic features in speech perception," J. Acoust. Soc. Am. **67**, 996-1013

Menard, S. W. (**1995**). *Applied logistic regression analysis* (Sage Publications, Thousand Oaks, Calif.).

Miller, J. L. (**1977**). “Nonindependence of feature processing in initial consonants,” J. Speech Hear. Res. **20**, 519-528.

Miller, G. A., and Nicely, P. E. (**1955**). “An analysis of perceptual confusions among some English consonants,” J. Acoust. Soc. Am. **27**, 338-352.

Nearey, T. M. (**1997**). “Speech perception as pattern recognition,” J. Acoust. Soc. Am. **101**, 3241-3254.

Nittrouer, S., Wilhelmsen, M., Shapley, K., Bodily, K. and Creutz, T. (**2003**). “Two reasons not to bring your children to cocktail parties,” J. Acoust. Soc. Am. **113**, 2254.

Ohde, R. N. (**1984**). “Fundamental frequency as an acoustic correlate of stop consonant voicing,” J. Acoust. Soc. Am. **75**, 224-230.

Peterson, G. E., and Lehiste, I. (**1960**). “Duration of syllable nuclei in English,” J. Acoust. Soc. Am. **32**, 693-703.

Repp, B. (**1979**). “Relative amplitude of aspiration noise as a voicing cue for syllable-initial stop consonants,” Lang. Speech **22**, 173-189.

Repp, B. (**1983**). “Trading relations among acoustic cues in speech perception are largely a result of phonetic categorization,” Speech Commun. **2**, 341-361

Sawusch, J. R., and Pisoni, D. B. (**1974**). “On the identification of place and voicing features in synthetic stop consonants,” J. Phonetics **2**, 181-194.

Soli, S. D., and Arabie, P. (**1979**). “Auditory versus phonetic accounts of observed confusions between consonant phonemes,” J. Acoust. Soc. Am. **66**, 46-59.

Stevens, K. N. (**1998**). *Acoustic phonetics* (The MIT Press, Cambridge, Massachusetts).

Stevens, K. N., and Klatt, D. H. (**1974**). “Role of formant transitions in the voiced-voiceless distinction for stops,” J. Acoust. Soc. Am. **55**, 653-659.

Strope, B., and Alwan, A. (**1997**). “A model of dynamic auditory perception and its application to robust word recognition,” IEEE Trans. Speech Audio Process. **5**(5), 451-464.

Summerfield, Q., and Haggard, M. (**1977**). “On the dissociation of spectral and temporal cues to the voicing distinction in initial stop consonants,” J. Acoust. Soc. Am. **62**, 435-448.

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

TABLE I. CV pairs used in this study.

	Labials	Alveolars	Velars
/a/	/ba,pa/	/da,ta/	/ga,ka/
/i/	/bi,pi/	/di,ti/	/gi,ki/
/u/	/bu,pu/	/du,tu/	/gu,ku/

TABLE II. Acoustic measurements.

Name	Description	Name	Description
vbD	Voicebar duration	F3bA	F3 onset amplitude
bstD	Burst duration	F1eA	F1 amplitude at the end of transition
bstpA	Burst peak amplitude	F2eA	F2 amplitude at the end of transition
votD	VOT duration	F3eA	F3 amplitude at the end of transition
votpA	VOT peak amplitude	F1sl	F1 slope
F1D	F1 transition duration	F2sl	F2 slope
F2D	F2 transition duration	F3sl	F3 slope
F3D	F3 transition duration	F1dA	F1 amplitude change
F1b	F1 onset frequency	F2dA	F2 amplitude change
F2b	F2 onset frequency	F3dA	F3 amplitude change
F3b	F3 onset frequency	F1df	F1 frequency change
F1e	F1 frequency at the end of transition	F2df	F2 frequency change
F2e	F2 frequency at the end of transition	F3df	F3 frequency change
F3e	F3 frequency at the end of transition	F0b	F0 onset frequency
F1bA	F1 onset amplitude	F0df	F0 frequency change
F2bA	F2 onset amplitude		

TABLE III. Percent correct classification (shown as a superscript) of the quiet speech tokens (from all talkers) based on a single acoustic property measured without the addition of the white Gaussian noise.

/ba,pa/	/bi,pi/	/bu,pu/	/da,ta/	/di,ti/	/du,tu/	/ga,ka/	/gi,ki/	/gu,ku/
votD^{100}	votD^{100}	votD^{100}	votD^{100}	votD^{100}	votD^{100}	votD^{100}	votD^{100}	votD^{100}
F1b^{100}	F3b^{97a}		F1b^{100}	votpA^{100}	votpA^{94a}	F1b^{100}	F0b^{97a}	
F1df^{100}	F3eA^{97a}		F1df^{100}	F2b^{94a}		F2b^{100}	F0df^{94a}	
F1sl^{97}	F3sl^{94a}		F1D^{94a}	F1D^{91a}		F1df^{100}		
	F2dA^{94a}		F2b^{94a}	F3df^{91a}		F2df^{100}		
	F3df^{94a}		F2df^{94a}			F1sl^{94}		
	F2b^{91a}		F3b^{91a}			F1eA^{91a}		
	F2sl^{91}							
	F2df^{91a}							

[a] Male and female token difference was significant at a $p < 0.000179$ level (Bonferroni-corrected $p < 0.050$ for multiple comparisons; male and female tokens had separate thresholds on the acoustic property for voicing classification). When male and female token difference was not significant, one threshold was used.

TABLE IV. Percent correct voicing judgments as a function of SNR, place of articulation, and vowel context (data averaged across all talkers and all listeners).

	/b,p/			/d,t/			/g,k/		
SNR (dB)	a	i	u	a	i	u	a	i	u
Quiet	100	100	99	100	100	99	97	100	100
10	98	93	89	98	99	98	98	95	99
5	98	90	82	100	95	89	97	93	89
0	95	82	83	97	84	86	96	91	84
-5	88	70	74	96	75	69	91	80	80
-10	66	55	66	62	61	60	57	59	69
-15	47	49	53	44	51	50	49	50	59

TABLE V. Correlation coefficients (shown as superscripts) of threshold SNRs with absolute differences of the means of various acoustic properties across all talkers (with the highest correlation coefficient listed first).

F1df$^{0.87}$[b]	F2df$^{0.34}$	F0b$^{-0.29}$
F1b$^{0.86}$[b]	F3df$^{0.08}$	F2sl$^{-0.40}$
F1eA$^{0.84}$	F3b$^{0.01}$	F2dA$^{-0.45}$
F1sl$^{0.79}$	F3sl$^{-0.06}$	F3eA$^{-0.50}$
F1D$^{0.78}$	votpA$^{-0.08}$	votD$^{-0.85}$
F2b$^{0.46}$	F0df$^{-0.14}$	

[b] The correlation was significant at a $p < 0.00294$ level ($df = 7$, Bonferroni-corrected $p < 0.050$ for multiple comparisons).

FIG. 1. Spectrograms of /da/ and /ta/ tokens, sampled at 8 kHz, illustrating different acoustic properties (burst, VOT, F1 transition, and voicebar).

FIG. 2. (a) LPC spectrum of a /ta/ token during the vowel, (b) DFT spectrum of a /ta/ token during the burst, (c) formant transition measurements, and (d) determination of formant transition offset (with the F1 frequencies obtained from an 8-kHz waveform using LPC analyses).

FIG. 3. Histograms of VOT duration (votD) for the nine voiced/voiceless pairs with the voiced and voiceless tokens counted separately. The histogram bin centers range from 2.5 to 72.5 ms with a 5-ms step. VOT duration of more than 75 ms is counted into the 72.5-ms-center region.

FIG. 4. Histograms of F1 onset frequency (F1b) for the nine voiced/voiceless pairs with the voiced and voiceless tokens counted separately. The histogram bin centers range from 125 to 1025 Hz with a 50-Hz step. F1b of less than 100 Hz and of more than 1050 Hz is counted into the 125-Hz-center and 1025-Hz-center regions, respectively. Note that the bins are well separated by voicing categories only in the /a/ context.

FIG. 5. Percent correct voicing judgments as a function of (a) vowel context (/a/, /i/, or /u/), (b) SNR (dB), (c) place of articulation (/b,p/, /d,t/, or /g,k/), (d) gender (male or female) and SNR combination, (e) vowel and SNR combination, and (f) place of articulation and vowel combination. The error bars show standard errors of means. The 50% chance performance is indicated with “50*’ on the percent correct axis.

FIG. 6. A sigmoid fitting of percent correct scores as a function of SNR (dB) for the nine voiced/voiceless pairs. For each voiced/voiceless pair, the 79% threshold line is drawn, and the

threshold SNR value is labeled. The 50% chance performance is indicated with the dash lines and with a “50*” on the percent correct axis.

FIG. 7. Threshold SNR values (dB) for the nine voiced/voiceless pairs, separated by gender of the talkers. Group 1 (/Ca/ syllables) shows little difference in perceiving male versus female tokens, while Group 2 (/di,ti/, /du,tu/, and /bu,pu/) shows large gender related differences.

FIG. 8. Threshold SNRs and absolute differences of the means of several acoustic properties for the nine voiced/voiceless pairs. For each point (CV pair), the threshold SNR and the difference of the means resulted from the perception and measurement of 32 tokens (four talkers x four tokens of the same syllable x two syllables in one CV pair), respectively. Each point is indicated with its vowel context [/a/ (square), /i/ (diamond), or /u/ (circle)] and place of articulation [/b,p/ (solid), /d,t/ (bold), or /g,k/ (thin)]. Numbers inside panels are the correlation coefficients between the nine threshold SNRs and the nine absolute differences of the means. A fitted line is shown in only one panel for illustrative purposes.

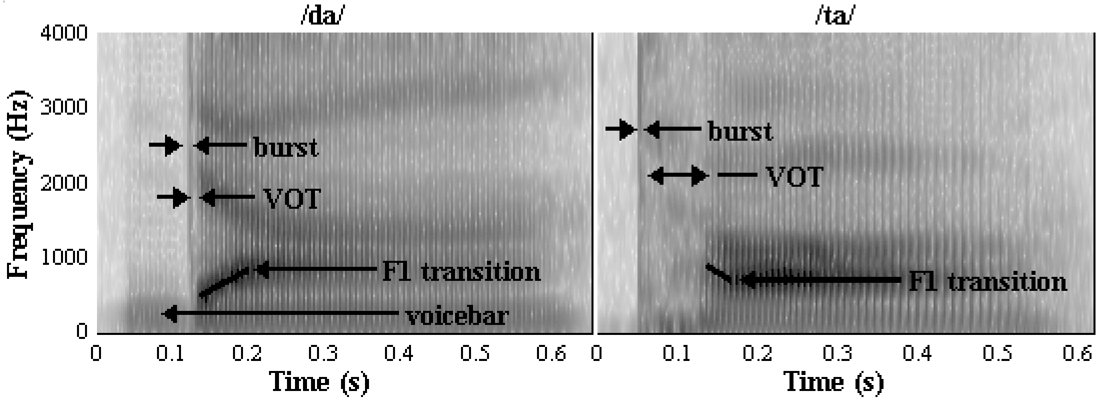

/da/
/ta/
4000
3000
2000
1000
0
Frequency (Hz)
burst
VOT
F1 transition
voicebar
burst
VOT
F1 transition
0
0.1
0.2
0.3
0.4
0.5
0.6
Time (s)
0
0.1
0.2
0.3
0.4
0.5
0.6
Time (s)

FIG. 1

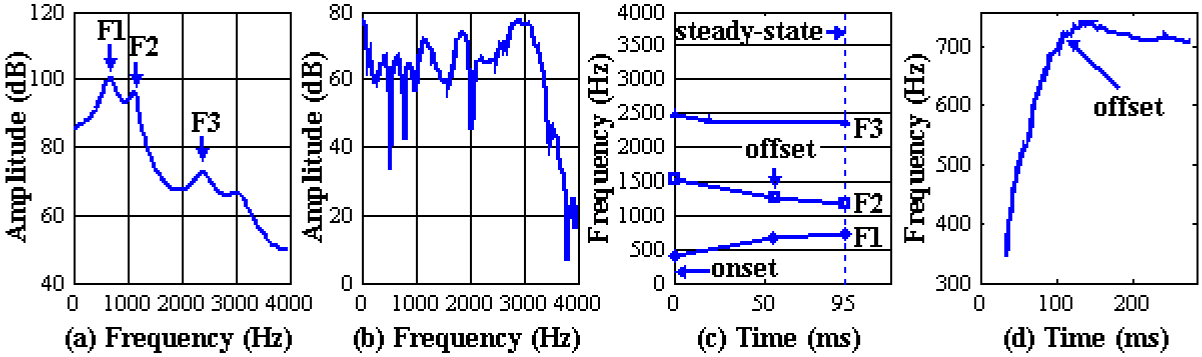

120
100
80
60
40
Amplitude (dB)
F1
F2
F3
0
1000 2000 3000 4000
(a) Frequency (Hz)
80
60
40
20
0
Amplitude (dB)
0
1000 2000 3000 4000
(b) Frequency (Hz)
4000
3500
3000
2500
2000
1500
1000
500
0
Frequency (Hz)
steady-state
offset
onset
F3
F2
F1
0
50
95
(c) Time (ms)
700
600
500
400
300
Frequency (Hz)
offset
0
100
200
(d) Time (ms)

FIG. 2

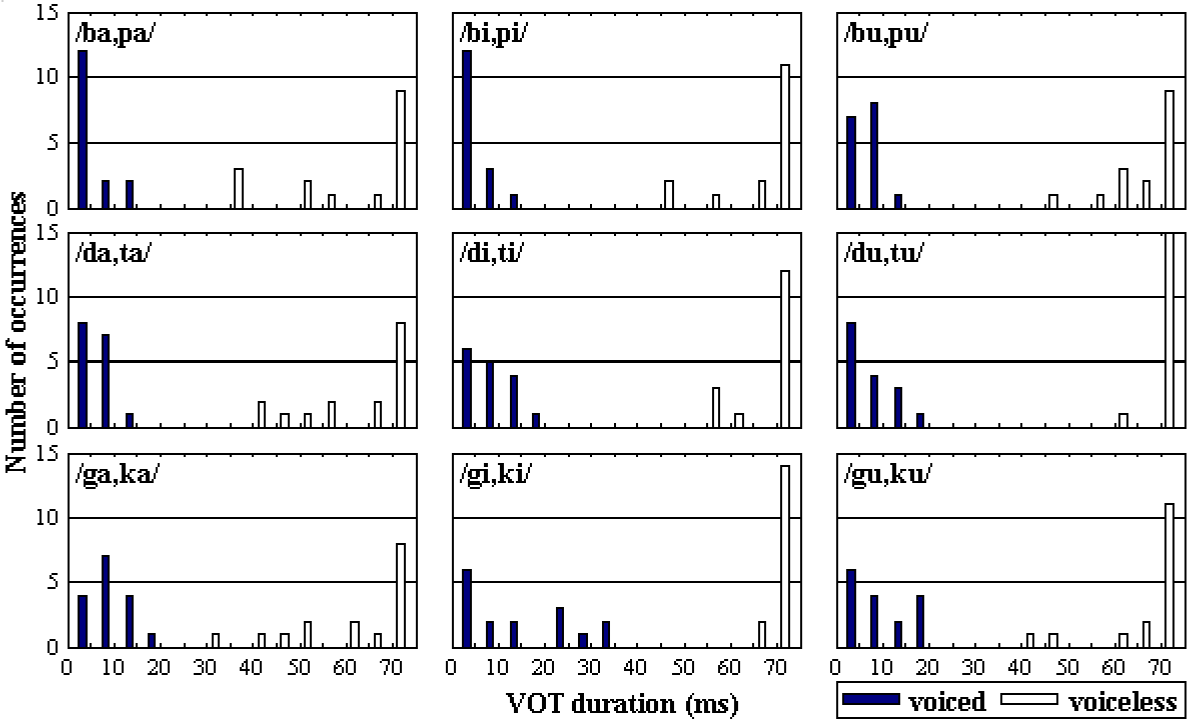

/ba,pa/
/bi,pi/
/bu,pu/
/da,ta/
/di,ti/
/du,tu/
/ga,ka/
/gi,ki/
/gu,ku/
Number of occurrences
0
5
10
15
0
10
20
30
40
50
60
70
VOT duration (ms)
voiced
voiceless

FIG. 3

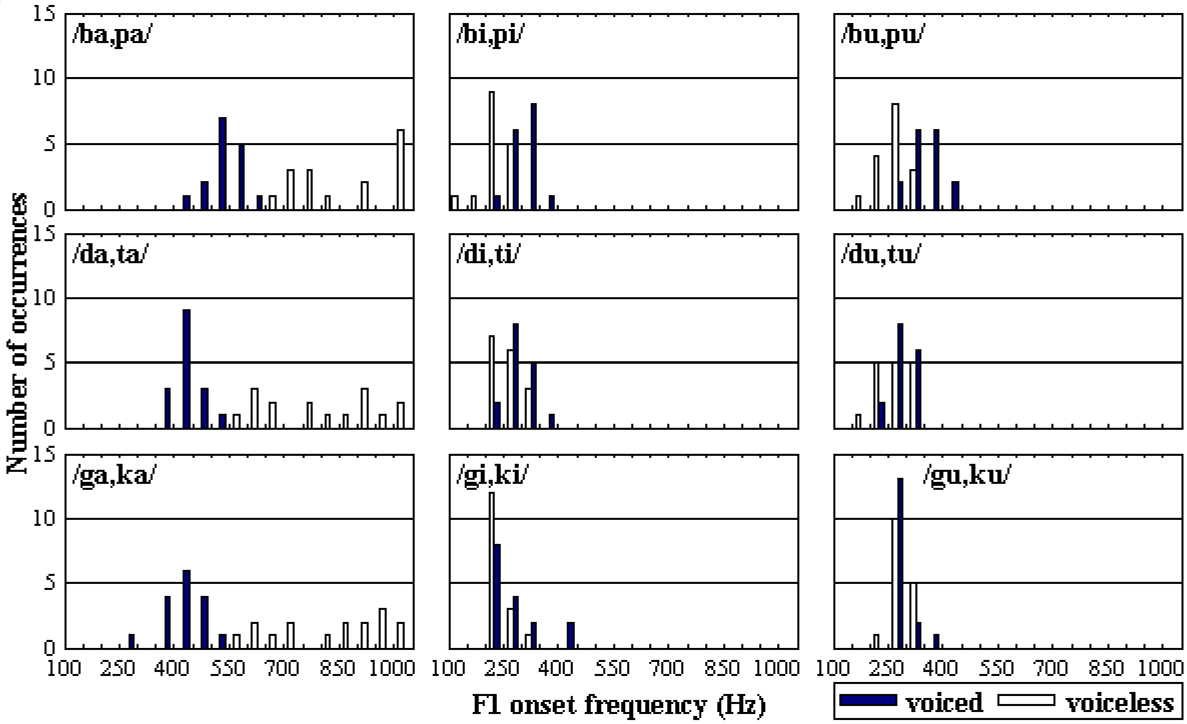

/ba,pa/
/bi,pi/
/bu,pu/
/da,ta/
/di,ti/
/du,tu/
/ga,ka/
/gi,ki/
/gu,ku/
Number of occurrences
0
5
10
15
100
250
400
550
700
850
1000
F1 onset frequency (Hz)
voiced
voiceless

FIG. 4

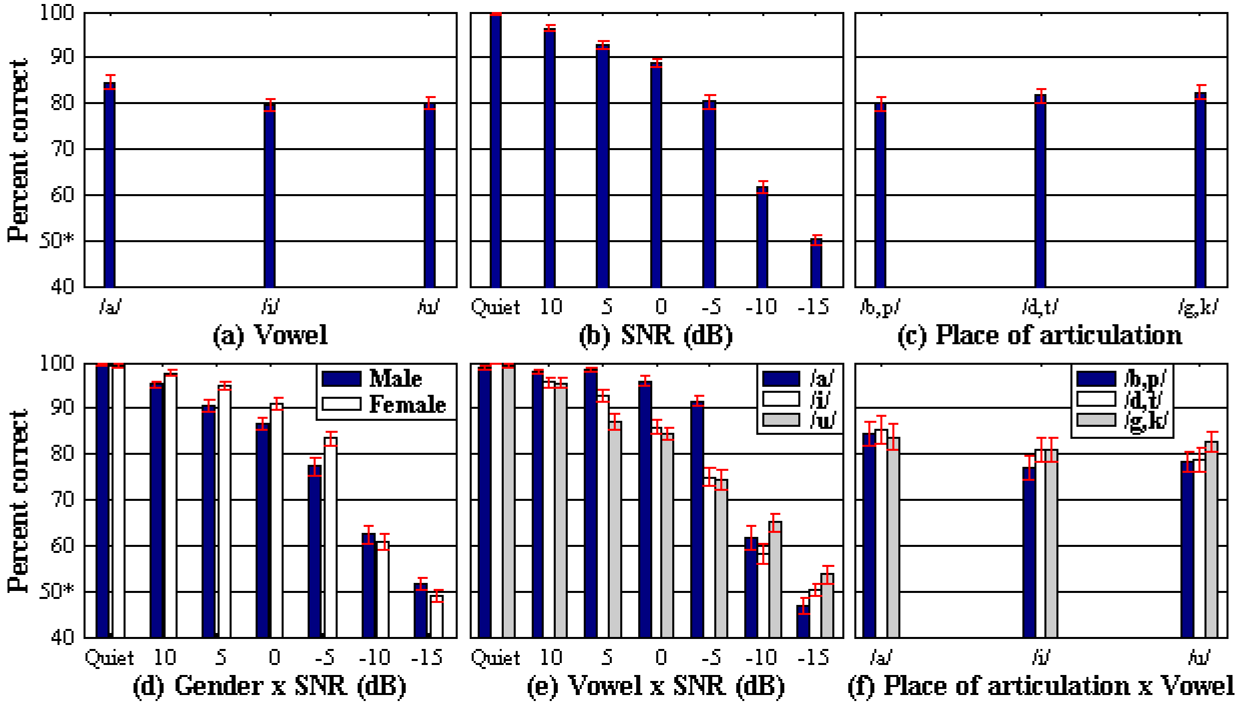

Percent correct
100
90
80
70
60
50*
40
/a/
/i/
/u/
(a) Vowel
Quiet
10
5
0
-5
-10
-15
(b) SNR (dB)
/b,p/
/d,t/
/g,k/
(c) Place of articulation
Male
Female
(d) Gender x SNR (dB)
/a/
/i/
/u/
(e) Vowel x SNR (dB)
/b,p/
/d,t/
/g,k/
(f) Place of articulation x Vowel

FIG. 5

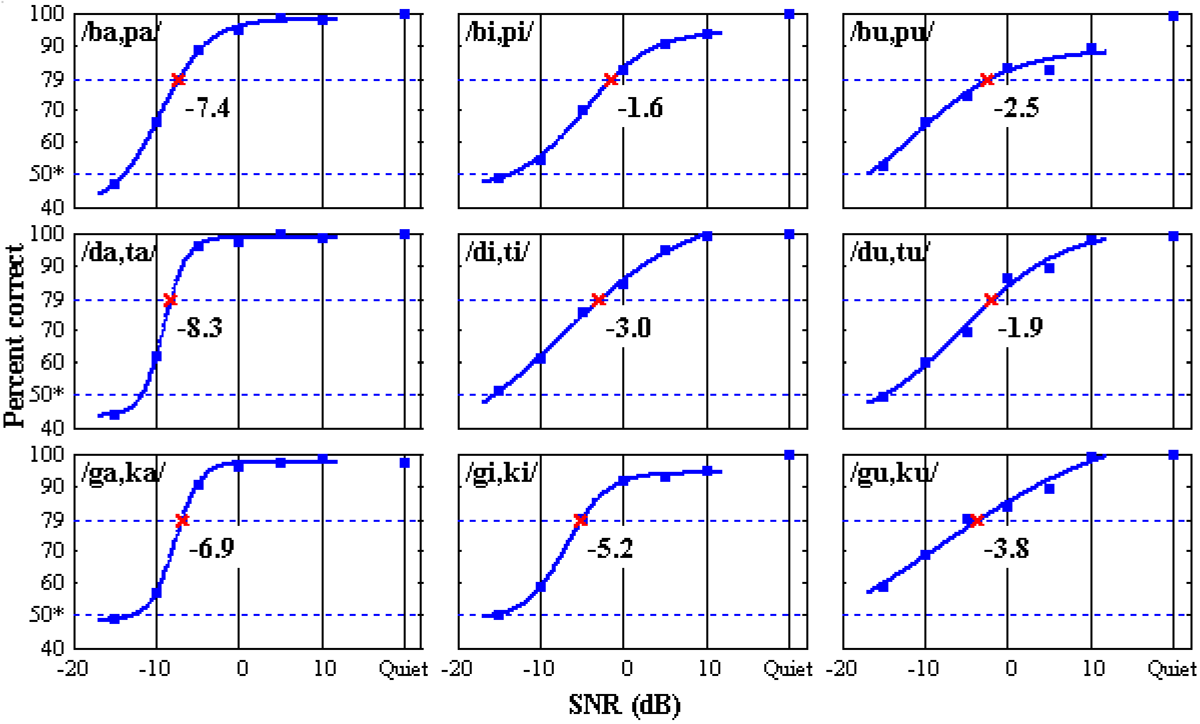

/ba,pa/
-7.4
/bi,pi/
-1.6
/bu,pu/
-2.5
/da,ta/
-8.3
/di,ti/
-3.0
/du,tu/
-1.9
/ga,ka/
-6.9
/gi,ki/
-5.2
/gu,ku/
-3.8
100
90
79
70
60
50*
40
Percent correct
-20
-10
0
10
Quiet
SNR (dB)

FIG. 6

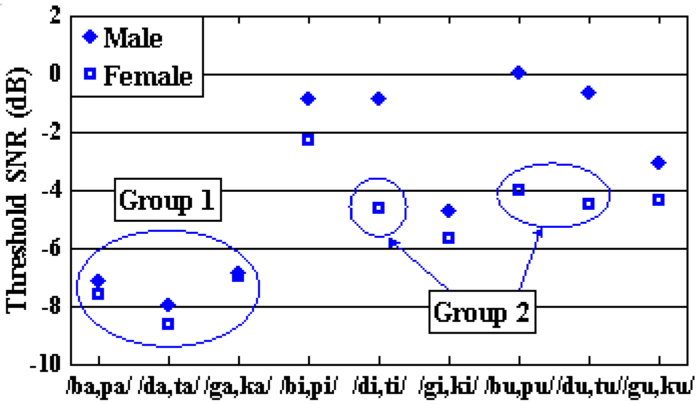

Threshold SNR (dB)
2
0
-2
-4
-6
-8
-10
◆ Male
□ Female
Group 1
Group 2
/ba,pa/
/da,ta/
/ga,ka/
/bi,pi/
/di,ti/
/gi,ki/
/bu,pu/
/du,tu/
/gu,ku/

FIG. 7

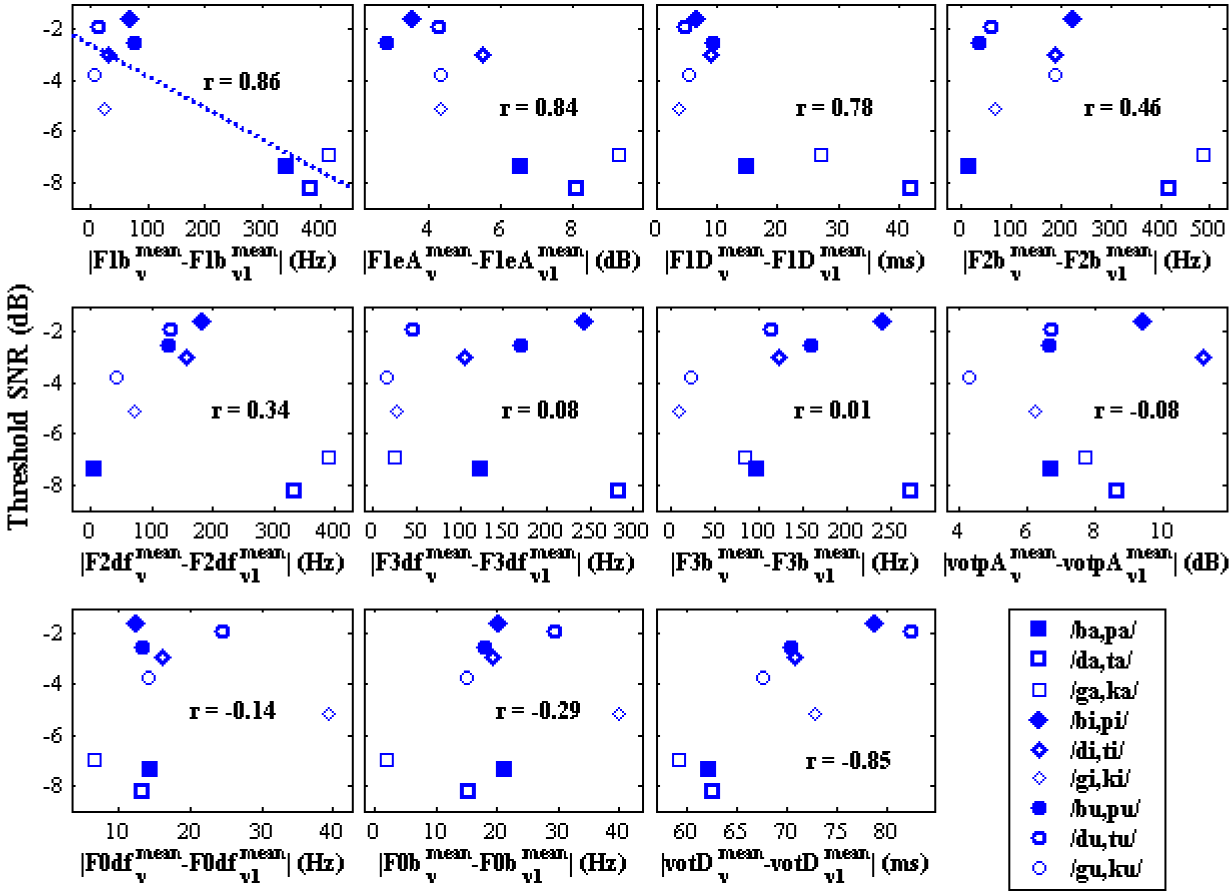
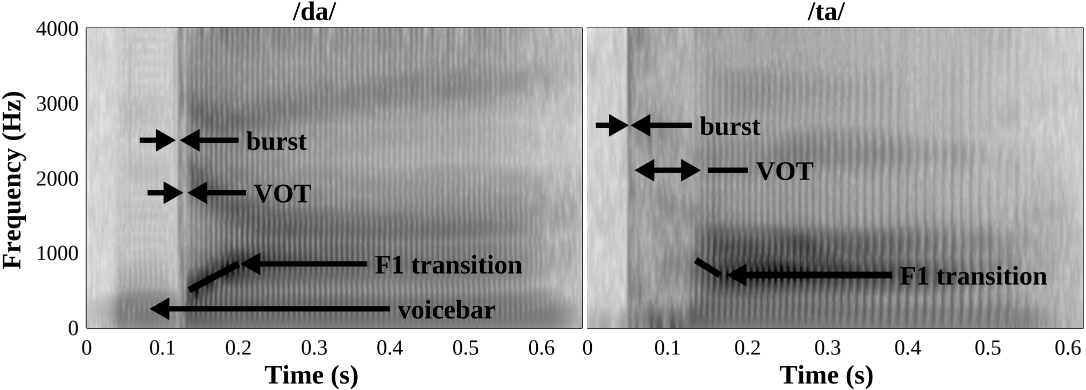
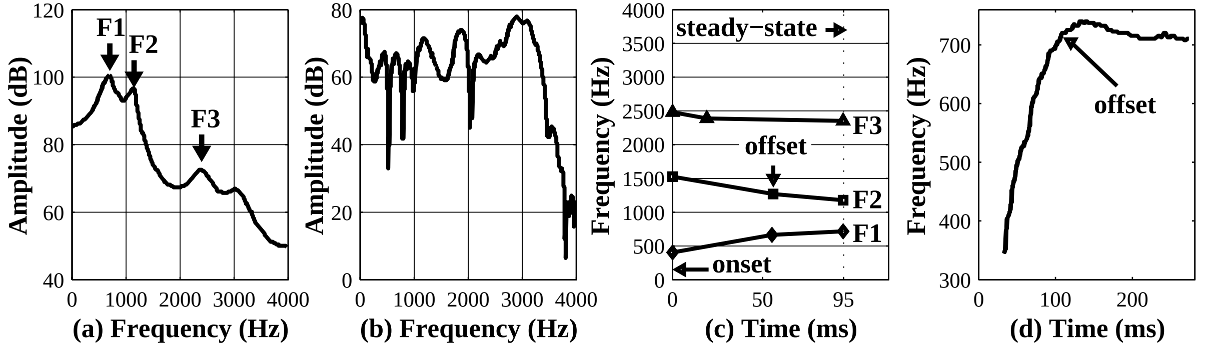
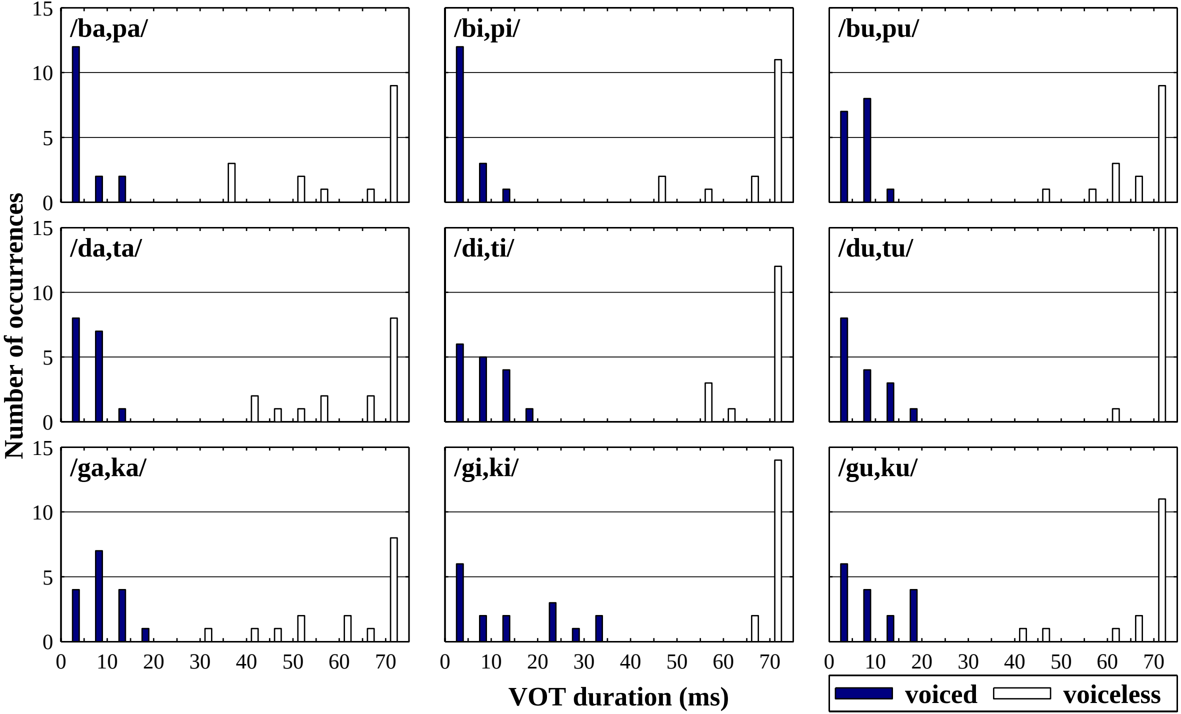
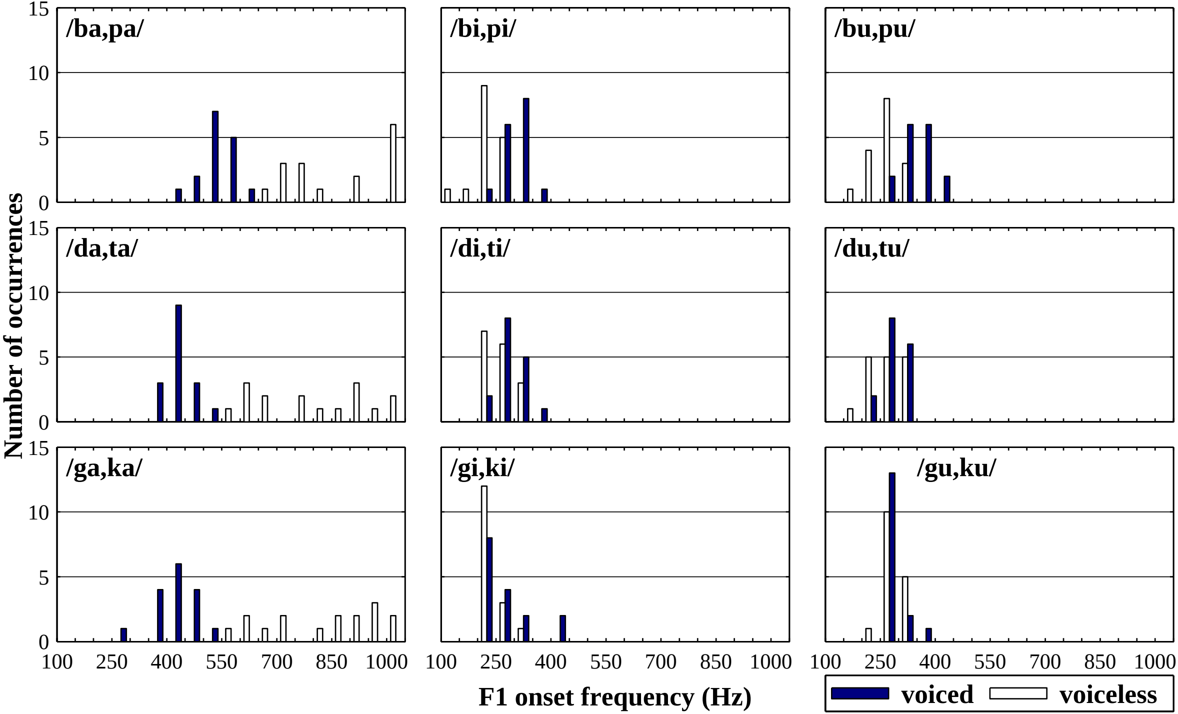

Threshold SNR (dB)
-2
-4
-6
-8
r = 0.86
0 100 200 300 400
|F1b v mean-F1b v1 mean| (Hz)
r = 0.84
4 6 8
|F1eA v mean-F1eA v1 mean| (dB)
r = 0.78
0 10 20 30 40
|F1D v mean-F1D v1 mean| (ms)
r = 0.46
0 100 200 300 400 500
|F2b v mean-F2b v1 mean| (Hz)
r = 0.34
0 100 200 300 400
|F2df v mean-F2df v1 mean| (Hz)
r = 0.08
0 50 100 150 200 250 300
|F3df v mean-F3df v1 mean| (Hz)
r = 0.01
0 50 100 150 200 250
|F3b v mean-F3b v1 mean| (Hz)
r = -0.08
4 6 8 10
|votpA v mean-votpA v1 mean| (dB)
r = -0.14
10 20 30 40
|F0df v mean-F0df v1 mean| (Hz)
r = -0.29
0 10 20 30 40
|F0b v mean-F0b v1 mean| (Hz)
r = -0.85
60 65 70 75 80
|votD v mean-votD v1 mean| (ms)
■ /ba,pa/
▣ /da,ta/
□ /ga,ka/
◆ /bi,pi/
◈ /di,ti/
◇ /gi,ki/
● /bu,pu/
◉ /du,tu/
○ /gu,ku/

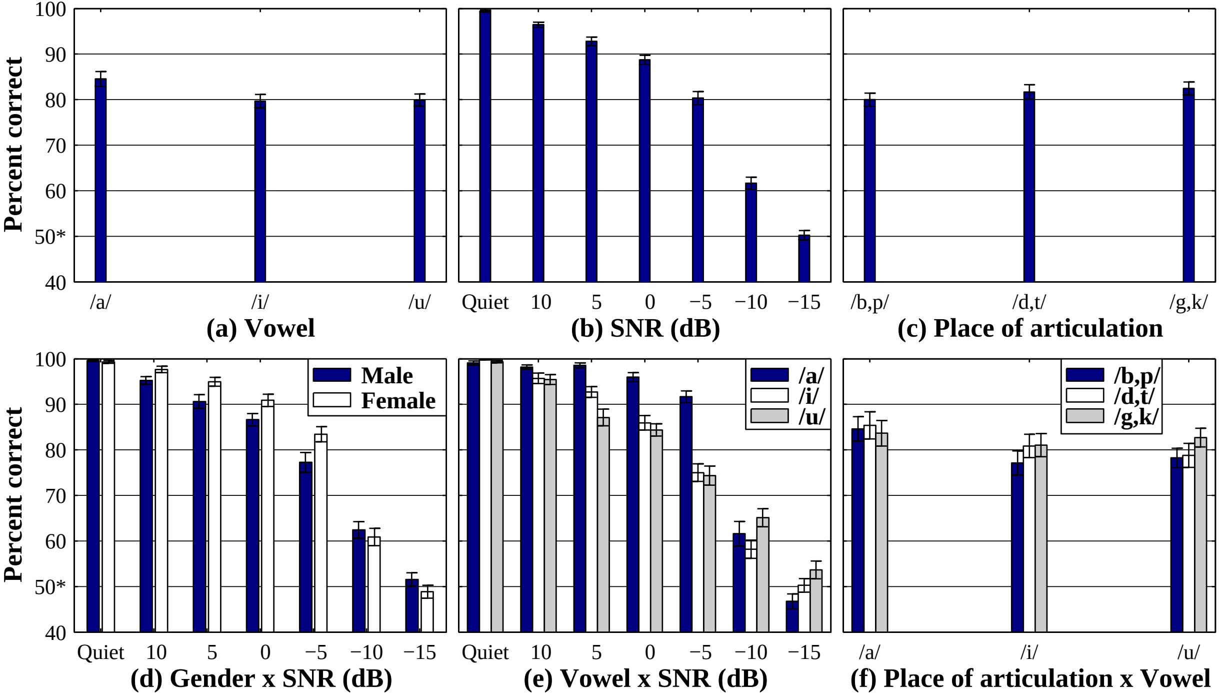
FIG. 8

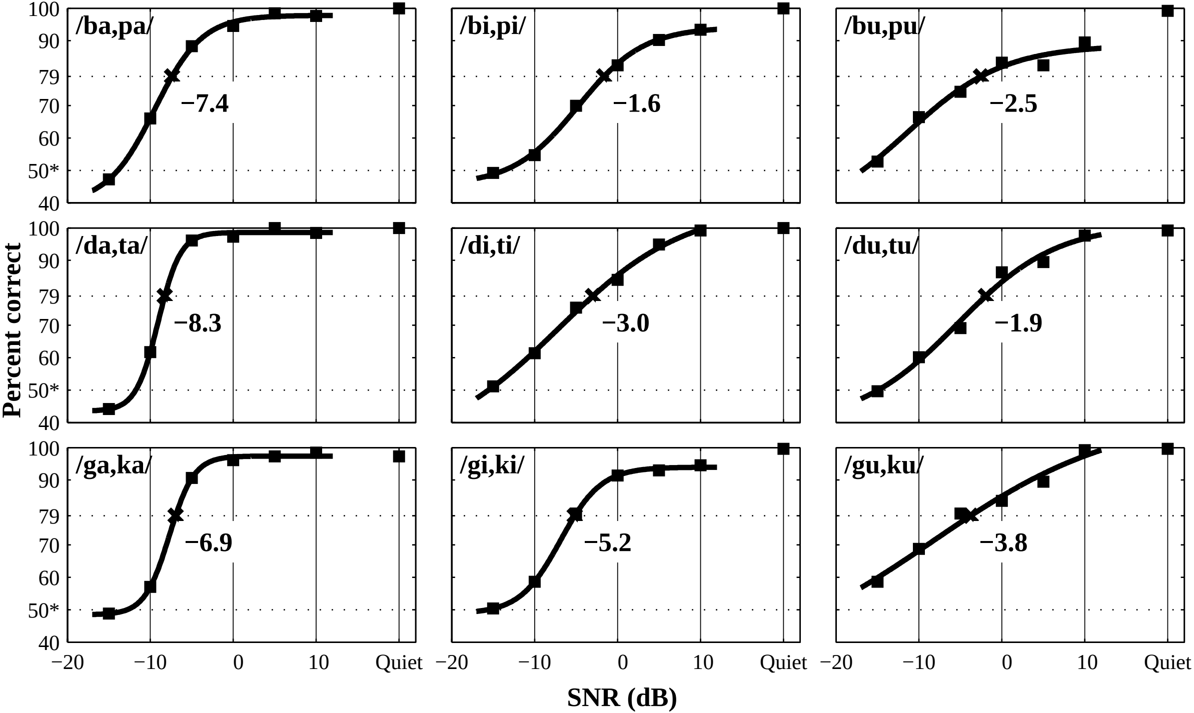

/da/
/ta/
Frequency (Hz)
4000
3000
2000
1000
0
burst
VOT
F1 transition
voicebar
burst
VOT
F1 transition
0
0.1
0.2
0.3
0.4
0.5
0.6
Time (s)
0
0.1
0.2
0.3
0.4
0.5
0.6
Time (s)

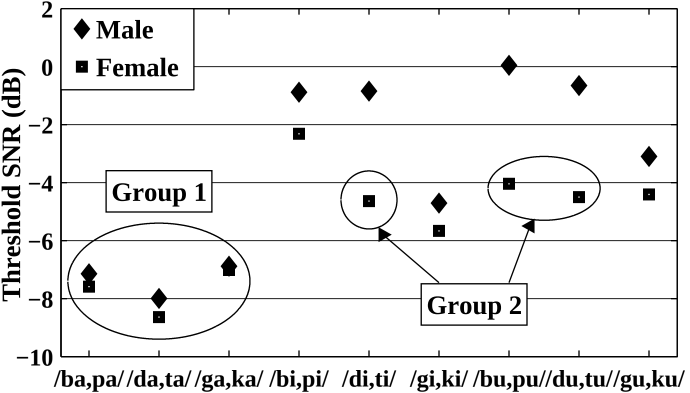

F1
F2
F3
Amplitude (dB)
120
100
80
60
40
0
1000
2000
3000
4000
(a) Frequency (Hz)
Amplitude (dB)
80
60
40
20
0
0
1000
2000
3000
4000
(b) Frequency (Hz)
steady−state
offset
onset
F3
F2
F1
Frequency (Hz)
4000
3500
3000
2500
2000
1500
1000
500
0
0
50
95
(c) Time (ms)
offset
Frequency (Hz)
700
600
500
400
300
0
100
200
(d) Time (ms)

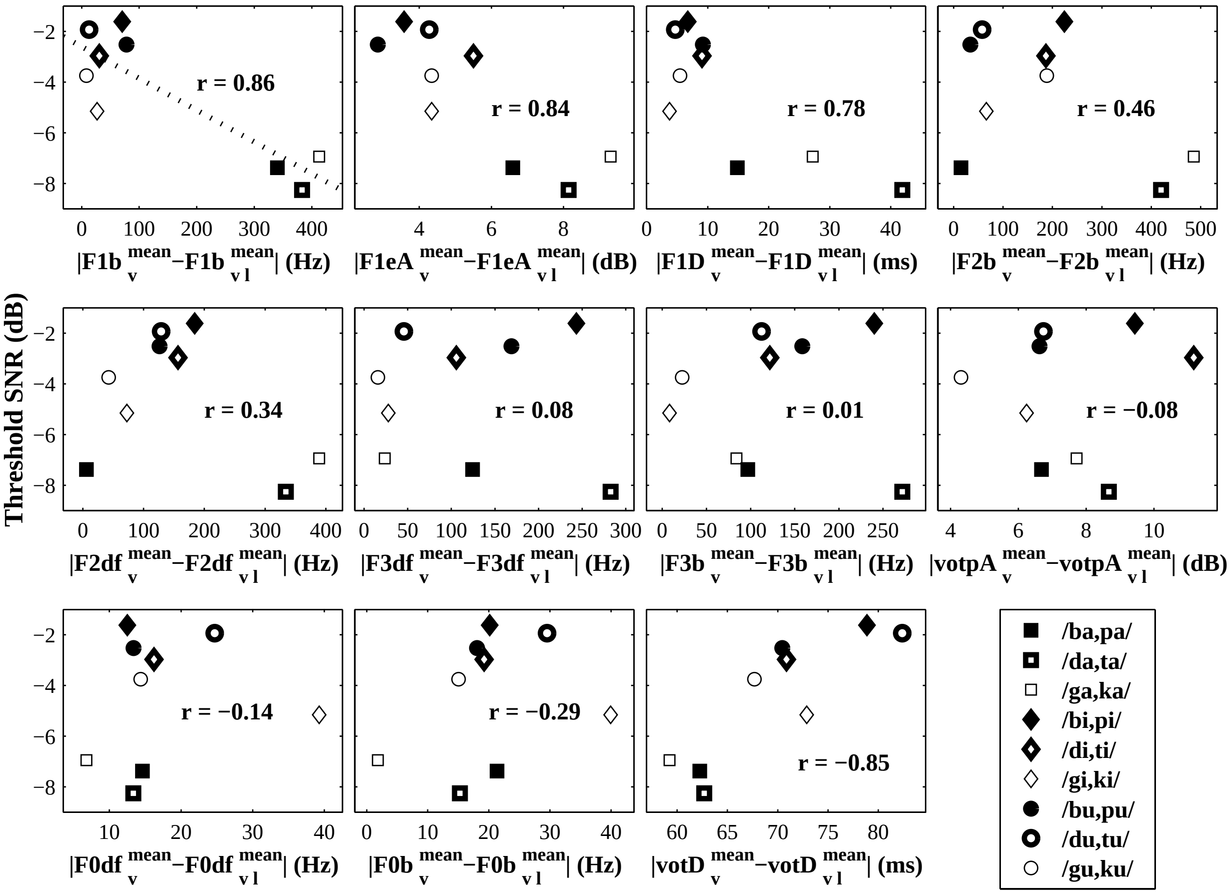

/ba,pa/
/bi,pi/
/bu,pu/
/da,ta/
/di,ti/
/du,tu/
/ga,ka/
/gi,ki/
/gu,ku/
Number of occurrences
0
5
10
15
0
10
20
30
40
50
60
70
VOT duration (ms)
voiced
voiceless


/ba,pa/
/bi,pi/
/bu,pu/
/da,ta/
/di,ti/
/du,tu/
/ga,ka/
/gi,ki/
/gu,ku/
Number of occurrences
0
5
10
15
100
250
400
550
700
850
1000
F1 onset frequency (Hz)
voiced
voiceless


Percent correct
100
90
80
70
60
50*
40
/a/
/i/
/u/
(a) Vowel
Quiet
10
5
0
−5
−10
−15
(b) SNR (dB)
/b,p/
/d,t/
/g,k/
(c) Place of articulation
Male
Female
(d) Gender x SNR (dB)
/a/
/i/
/u/
(e) Vowel x SNR (dB)
/b,p/
/d,t/
/g,k/
(f) Place of articulation x Vowel


/ba,pa/
−7.4
/bi,pi/
−1.6
/bu,pu/
−2.5
/da,ta/
−8.3
/di,ti/
−3.0
/du,tu/
−1.9
/ga,ka/
−6.9
/gi,ki/
−5.2
/gu,ku/
−3.8
Percent correct
100
90
79
70
60
50*
40
−20
−10
0
10
Quiet
SNR (dB)


Male
Female
Group 1
Group 2
Threshold SNR (dB)
2
0
−2
−4
−6
−8
−10
/ba,pa/
/da,ta/
/ga,ka/
/bi,pi/
/di,ti/
/gi,ki/
/bu,pu/
/du,tu/
/gu,ku/


Threshold SNR (dB)
−2
−4
−6
−8
r = 0.86
0 100 200 300 400
$|F1b_{v}^{mean} - F1b_{v\,l}^{mean}|$ (Hz)
r = 0.84
4 6 8
$|F1eA_{v}^{mean} - F1eA_{v\,l}^{mean}|$ (dB)
r = 0.78
0 10 20 30 40
$|F1D_{v}^{mean} - F1D_{v\,l}^{mean}|$ (ms)
r = 0.46
0 100 200 300 400 500
$|F2b_{v}^{mean} - F2b_{v\,l}^{mean}|$ (Hz)
r = 0.34
0 100 200 300 400
$|F2df_{v}^{mean} - F2df_{v\,l}^{mean}|$ (Hz)
r = 0.08
0 50 100 150 200 250 300
$|F3df_{v}^{mean} - F3df_{v\,l}^{mean}|$ (Hz)
r = 0.01
0 50 100 150 200 250
$|F3b_{v}^{mean} - F3b_{v\,l}^{mean}|$ (Hz)
r = −0.08
4 6 8 10
$|votpA_{v}^{mean} - votpA_{v\,l}^{mean}|$ (dB)
r = −0.14
10 20 30 40
$|F0df_{v}^{mean} - F0df_{v\,l}^{mean}|$ (Hz)
r = −0.29
0 10 20 30 40
$|F0b_{v}^{mean} - F0b_{v\,l}^{mean}|$ (Hz)
r = −0.85
60 65 70 75 80
$|votD_{v}^{mean} - votD_{v\,l}^{mean}|$ (ms)
■ /ba,pa/
▣ /da,ta/
□ /ga,ka/
◆ /bi,pi/
◈ /di,ti/
◇ /gi,ki/
● /bu,pu/
◉ /du,tu/
○ /gu,ku/