The influence of noise on vowel and consonant cues

Gaurang Parikh and Philipos C. Loizou[a)]
Department of Electrical Engineering, University of Texas at Dallas, Richardson, Texas 75083-0688

(Received 29 April 2005; revised 19 September 2005; accepted 20 September 2005)

This study assessed the acoustic and perceptual effect of noise on vowel and stop-consonant spectra. Multi-talker babble and speech-shaped noise were added to vowel and stop stimuli at −5 to +10 dB S/N, and the effect of noise was quantified in terms of (a) spectral envelope differences between the noisy and clean spectra in three frequency bands, (b) presence of reliable F1 and F2 information in noise, and (c) changes in burst frequency and slope. Acoustic analysis indicated that F1 was detected more reliably than F2 and the largest spectral envelope differences between the noisy and clean vowel spectra occurred in the mid-frequency band. This finding suggests that in extremely noisy conditions listeners must be relying on relatively accurate F1 frequency information along with partial F2 information to identify vowels. Stop consonant recognition remained high even at −5 dB despite the disruption of burst cues due to additive noise, suggesting that listeners must be relying on other cues, perhaps formant transitions, to identify stops. © *2005 Acoustical Society of America.* [DOI: 10.1121/1.2118407]

PACS number(s): 43.71.Es, 43.71.Ky [ALF] Pages: 3874–3888

I. INTRODUCTION

Since the classic study by Peterson and Barney (1952) on the distribution of the vowel formant frequencies on the F1-F2 plane, many studies were conducted in quiet on vowel perception (see reviews by Strange, 1989, 1999), estimation of difference limens for formant discrimination (Flanagan, 1955; Hawks, 1994; Liu and Kewley-Port, 2001), vowel modeling (e.g., Syrdal and Gopal, 1986; Molis, 2005), and other aspects of vowel discrimination. Many factors were found to be important for vowel identification including formant frequencies (Peterson and Barney, 1952), vowel duration (e.g., Ainsworth, 1972), F0 (e.g., Lehiste and Meltzer, 1973), spectral contrast (Leek *et al.*, 1987; Loizou and Poroy, 2001), formant contour (Hillenbrand and Gayvert, 1993; Hillenbrand and Nearey, 1999), spectral shape (e.g., Zahorian and Jaghargi, 1986; Ito *et al.*, 2001), and spectral change (e.g., Strange *et al.*, 1983). Of these factors, the formant frequencies and spectral shape, in particular, have been found to be major cues to vowel perception. The contribution of these cues to vowel perception has been the subject of a longstanding debate between a formant-based and a spectral-shape-based theory of vowel perception (see review in Ito *et al.*, 2001).

Evidence in support of a formant theory comes from studies in which speech synthesized from formant frequencies was found to be highly intelligible despite removal of detailed spectral shape information (e.g., Remez *et al.*, 1981). There is also ample evidence from animal neurophysiological studies that suggests that the formant frequencies are encoded in the temporal discharge patterns of the auditory nerve fibers (e.g., Young and Sachs, 1979). Discharge rates of large populations of auditory nerve fibers when plotted as a function of the characteristic frequency showed distinct peaks corresponding to the formants of the vowels, suggesting a place-code representation of vowel formants, at least for low sound intensities. Formants are also coded in the time pattern of the auditory fiber's discharges over a large range of sound intensities, suggesting a temporal-code representation (Young and Sachs, 1979). Scalp-recorded frequency-following responses (FFRs) from humans showed clearly discernible peaks at harmonics adjacent to the first two formant frequencies (Krishnan, 1999, 2002). The FFR data led Krishnan (2002) to conclude that the neural coding of formants based on phase locking is preserved even at high levels in the human brainstem.

Despite its simplicity, the formant theory does not account for the outcomes of several psychophysical and modeling studies, leading some researchers to a spectral-shape or whole-spectrum theory. Advocates of the whole-spectrum theory describe the vowels in terms of the properties of the spectrum as a whole rather than in terms of the individual formant frequency values. Several arguments were offered against a formant representation (e.g., Bladon, 1982). First, automatic formant frequency detection tends to be unreliable particularly in noise and in situations in which the harmonics are spaced widely apart (e.g., in children's speech). Errors made by formant tracking algorithms (e.g., formants merging) are not consistent with those made by human listeners. Second, formant-based models do not incorporate any characteristics of the peripheral auditory system such as critical band filtering. The whole-spectrum model proposed by Bladon and Lindblohm (1981) incorporated several aspects of auditory processing including critical-band filtering and loudness and yielded a high correlation with normal-hearing listener's judgments of vowel quality. Third, cues other than formant frequencies can affect vowel perception. Several researchers have demonstrated that the relative amplitude of adjacent formants can affect the perceived vowel quality (Chistovich and Lublinskaja, 1979; Ito *et al.*, 2001). Ito *et al.*, for instance, demonstrated that the formant amplitude

[a)]Author to whom correspondence should be addressed. Electronic-mail: loizou@utdallas.edu

ratio is equally or more effective than F2 as cue to place of articulation (front/back). Ito *et al.* (2001) concluded that the formant frequency cues are not the exclusive cues to vowel perception and that spectral shape could be crucial.

In most of the above studies the formant frequencies and spectral shape were varied independently of one another. In noise, however, both formant frequencies and spectral envelope (shape) will be affected to some degree, but it is not known which of the two will be affected the most. Also, it is not known how and to what degree the formant frequencies will be affected or to what degree the spectral envelope will be altered by the noise. These questions have not been addressed in previous studies on recognition of vowels in noise as those studies concentrated primarily on identification errors (e.g., Pickett, 1957; Nebalek and Dagenais, 1986) and on the relationship between vowel identification, hearing loss, and age (e.g., Nebalek, 1988), rather than on factors that might potentially be responsible for the identification errors.

The present study takes the first step in quantifying the effect of noise (multi-talker babble and continuous speech-shaped noise) on the spectrum of vowels. Given the importance of formant frequencies and spectral shape on vowel identification, we focus primarily on analyzing acoustic measurements of formant frequencies and spectral envelopes. Since we are ultimately interested in knowing whether listeners use formant frequency information to identify vowels in noise, we seek to establish first if there exists reliable and relatively accurate formant-frequency information in noise. More specifically, we measure how often there exists reliable F1 and/or F2 information in noise. This measure will tell us whether listeners have access to reliable and coherent formant frequency (F1 and F2) information in noise and, if so, to what degree. If there is no reliable formant information, then perhaps listeners make use of spectral shape information. To examine that, we evaluate the differences between the critical-band spectra of the clean and noisy vowels and correlate those differences with vowel identification scores. If the correlation analysis reveals that large critical-band spectral differences are associated with lower identification scores, then that would suggest that listeners are making use of spectral shape cues. The vowel representation based on critical-band spectra is chosen for two reasons. First, the critical-band spectra contain only gross peak information and no irrelevant harmonic details and can therefore be used to assess whether listeners make use of spectral-shape cues to identify vowels. Second, the critical-band vowel representation incorporates characteristics of the peripheral auditory system, such as critical-band filtering, and has been used by some as approximation to the "auditory excitation patterns" (Plomp, 1970; Klatt, 1982).

Parallel to examining the effect of noise on the vowel spectra, we also analyze the effect of noise on the consonant spectra, and particularly the stop consonants which have been studied extensively in the literature (see review in Kent and Read, 1992). The acoustic cues used for stop perception differ from those for vowel perception, hence different acoustic parameters are extracted and analyzed for the stop consonants. These parameters are based on theories and models proposed for understanding the cues to stop-consonant place of articulation. It is has long been recognized, since the Pattern Playback days at Haskins Labs (see Liberman *et al.*, 1952; Cooper *et al.*, 1952), that the spectrum of the stop burst is a major cue to place of articulation. Labial stops typically have low-frequency dominance, alveolars have high-frequency dominance, and velars are associated with a mid-frequency burst. Stevens and Blumstein (1978) explored the idea of constructing a spectral template that could be associated with each place of stop articulation. In these templates, the bilabials had a flat or a falling spectrum, the alveolars had a rising spectrum, and the velars had a compact (with a peak in mid-frequencies) spectrum. Using these templates, Blumstein and Stevens (1980) were able to classify stops with 85% accuracy. After several reevaluations of the spectral-template theory (e.g., Blumstein *et al.*, 1982; Walley and Carrell, 1983), research on place of articulation shifted toward the dynamic spectral change following the first tens of ms of the release. Other cues found in later studies to be important for place for articulation were the spectral change from the burst to voicing onset (e.g., Kewley-Port, 1983; Lahiri *et al.*, 1984) and formant transitions (e.g., Delattre *et al.*, 1955; Dorman and Loizou, 1996; Smits *et al.*, 1996).

In summary, the burst spectrum and formant transitions have been found to be major cues to stop-place of articulation. The effect of noise on these place cues, however, is not clear and has not been investigated. It is not known, for instance, how noise affects the tilt of the burst spectrum, and consequently whether a change in the spectral tilt would be accompanied by a shift in phonetic category. Similarly, it is also not known whether a change in burst frequency will be associated with low identification scores. While several studies (e.g., Dorman *et al.*, 1977; Dorman and Loizou, 1996; Smits *et al.*, 1996) used a conflicting-cue paradigm to probe the above questions, those studies were done in quiet. To answer the above questions, we will measure the slope and frequency of the burst spectrum in quiet and in noise and correlate these measurements with identification scores. We chose those two acoustic parameters because they are relatively simple to estimate in noise.

Previous studies examining stop consonant recognition in noise (e.g., Miller and Nicely, 1955; Wang and Bilger, 1973; Pickett, 1957) have not addressed the above questions. The studies by Pickett (1957) and Miller and Nicely (1955) on consonant identification in noise assessed the effect of noise on vowel/consonant perception, but only indirectly by analyzing the confusion errors. Such an analysis, however, leaves many questions unanswered and does not tell us specifically what caused the confusion errors. Only white noise was used in the studies of Miller and Nicely (1955) and Wang and Bilger (1973), making it difficult to generalize their findings to more realistic types of noise (e.g., multi-talker babble).

In summary, not many studies have assessed or quantified the perceptual effect of noise on vowel and consonant perception. In this study, we take the first step in quantifying the effect of multi-talker babble and continuous speech-shaped noise on the spectra of vowels and stop consonants. This paper attempts to answer several questions and has two

TABLE I. Mean F1 and F2 frequencies (in Hz) of the vowels used in this study.

		Had	Hod	Head	Hayed	Heard	Hid	Heed	Hoed	Hood	Hud	Who'd
F1	Male	627	786	555	438	466	384	331	500	424	629	319
	Female	666	883	693	492	518	486	428	538	494	809	435
F2	Male	1910	1341	1851	2196	1377	2039	2311	868	992	1146	938
	Female	2370	1682	1991	2437	1604	2332	2767	998	1102	1391	1384

interrelated aims. The first aim is to quantify the effect of noise by means of acoustic analysis of the vowel and consonant spectra. A subset of the aforementioned factors, known to be important for vowel and stop-consonant identification, will be assessed by comparing the acoustic parameters (e.g., spectral tilt, etc.) estimated in quiet with those estimated in noise. The acoustic parameter comparisons are meant to answer several questions, including the following. (1) How are the vowel spectral envelopes (critical-band spectra) affected? (2) How are the two formant frequencies (F1 and F2), known to be major cues to vowel recognition, affected? (3) How are the spectral tilt and frequency of the burst spectra affected? The above, and other, questions will be answered quantitatively by performing acoustic analysis of vowels and stop consonants embedded in −5 to 10 dB noise.

The second aim of this paper is to assess the perceptual effect of noise on vowel and stop-consonant identification. This will be done by performing correlation analysis between the acoustic parameter values and the vowel/stop-consonant identification scores. The second aim will be addressed in experiment 1. The results from the acoustic analysis and experiment 1 taken together will provide valuable insights on the cues used by listeners to understand speech in noise. Knowing how noise affects the spectrum of speech is important for several reasons. For one, such knowledge could help us design better noise reduction algorithms that could potentially improve hearing-impaired listeners' speech understanding in noise. Secondly, it could help us better understand which speech features are perceptually robust in additive noise, and consequently which features listeners attend to when identifying vowels or consonants in noise.

II. ACOUSTIC ANALYSIS

A. Method

1. Speech material

The vowel material consisted of the vowels in the words: "heed, hid, hayed, head, had, hod, hud, hood, hoed, who'd, heard." The stimuli were drawn from a large multi-talker vowel set used by Hillenbrand *et al.* (1995). A total of 66 vowel tokens were used for acoustic analysis: 33 vowels produced by male speakers and 33 vowels produced by female speakers. There were 6 tokens for each of the 11 vowels, 3 produced by male speakers and 3 by female speakers. A total of 20 different male speakers and 23 female speakers produced the 66 vowel tokens. Each speaker produced only a subset of the 11 vowels. The vowels were sampled at 16 kHz. Table I gives the steady-state F1 and F2 values of the vowel stimuli used in this study. The F1 and F2 values were sampled at the steady-state portion of the vowel and averaged across all speakers. The steady-state F1 and F2 values (Table I) of the vowel stimuli used in this study were provided by Hillenbrand *et al.* (1995).

Consonant material consisted of the stop consonants in VCV context, where V=/i a,u/ and C=/b d g p t k/. The stimuli were drawn from recordings made by Shannon *et al.* (1999). A total of 36 consonant tokens were used for acoustic analysis: 18 consonants (6 stops×3 vowel contexts) produced by a male speaker and 18 consonants produced by a female speaker. The consonants were sampled at 44.1 kHz.

2. Noise

Two types of noise were used, multi-talker babble (two male and two female talkers) and speech-shaped noise. The babble was taken from the AudiTEC CD (St. Louis) and was sampled at 16 kHz. The speech-shaped noise (sampled at 20 kHz) was constructed by filtering white noise through a 60-tap FIR filter with a frequency response that matched the long-term spectrum of the 11 male and 11 female vowels. Noise was first up-sampled to the sampling frequency of the vowel/consonant materials and then added to the vowels at −5, 0, 5, and 10 dB. Figure 1 shows the averaged long-term spectra of the multi-talker babble and speech-shaped noise.

3. Acoustic analysis of vowels

Prior to the acoustic analysis, the complete vowel data set was manually segmented to [h Vowel d]. The starting and ending times of the vocalic nuclei were measured by hand

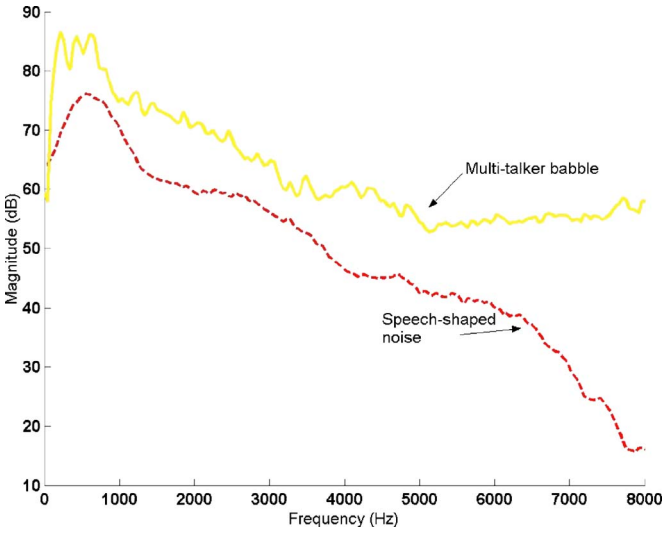

FIG. 1. (Color online) The long-term spectra of the multi-talker babble and continuous speech-shaped noise used in this study as maskers.

TABLE II. Lower and upper-edge edge (−3 dB) frequencies of the critical band filters.

Critical band	Lower edge frequencies (Hz)	Upper edge frequencies (Hz)	Center frequency (Hz)	Bandwidth (Hz)
1	0	100	50	100
2	100	200	150	100
3	200	300	250	100
4	300	400	350	100
5	400	510	450	110
6	510	630	570	120
7	630	770	700	140
8	770	920	840	150
9	920	1080	1000	160
10	1080	1270	1170	190
11	1270	1480	1370	210
12	1480	1720	1600	240
13	1720	2000	1850	280
14	2000	2320	2150	320
15	2320	2700	2500	380
16	2700	3150	2900	450
17	3150	3700	3400	550
18	3700	4400	4000	700
19	4400	53000	4800	900
20	5300	6400	5800	1100
21	6400	7700	7000	1300

from high-resolution digital spectrograms. To minimize the effect of formant movements due to /h/ and /d/, acoustic measurements were made starting from 20% of the vowel duration and ending at 80% of the vowel duration.

a. Critical-band spectral difference measurements. In order to quantify the effect of noise on different regions of the spectrum, we measured the critical-band spectral difference between the clean and noisy vowels in two different bands corresponding to the F1 and F2 regions. The measurements were based on a critical-band vowel representation. The critical-band spectra representation was chosen as it approximates the auditory "excitation patterns" (Plomp, 1970) of the vowels. The critical-band spectra were computed as follows.

The vocalic segment of the vowels (containing 20%–80% of the vowel duration) was first filtered through a 21-channel filterbank implemented using sixth-order Butterworth filters. The center frequencies of the filterbank were chosen according to critical-band spacing (Zwicker and Fastl, 1990) and are given in Table II. Estimates of the vowel spectra were then made by computing the root-mean-square (rms) energy of the 21-filterbank outputs within 10-ms windows. The 21-filterbank outputs provide approximations to the auditory excitation patterns (Plomp, 1970; Klatt, 1982).

The spectral difference between the clean and noisy vowels was then computed for two different frequency bands spanning the 0–8 kHz bandwidth, using a normalized Euclidean distance metric of the filterbank energies. This metric is similar to that used by Plomp (1970) based on third-octave bands and is used in our study for two reasons: (1) to quantify the effect of noise in individual frequency bands and (2) to assess the importance of spectral shape cues on vowel identification, since the critical-band spectra contain only gross peak information and no harmonic details. For the purposes of this study, the Euclidean distance metric was normalized by the energy of the clean spectrum within the specified frequency band. We found this normalization necessary due to the inherent differences in spectral magnitude levels at the low and high frequencies.

The two bands considered include a low-frequency (LF) band spanning the 0–1 kHz region and a middle-frequency (MF) band spanning the 1–2.7 kHz region. F1 typically resides in the LF band, and F2 resides in the MF band. Two spectral difference measurements were made, one for each band, every 10 ms:

$$\mathrm{LF} = \frac{\sqrt{\sum_{i=1}^{9} (F_i^c - F_i^n)^2}}{\sqrt{\sum_{i=1}^{9} (F_i^c)^2}}, \tag{1a}$$

$$\mathrm{MF} = \frac{\sqrt{\sum_{i=10}^{15} (F_i^c - F_i^n)^2}}{\sqrt{\sum_{i=10}^{15} (F_i^c)^2}}, \tag{1b}$$

where F_i^c denotes the ith filterbank energy of the clean vowel and F_i^n denotes the ith filterbank energy of the noisy vowel.

b. Measurements of formant frequency (F1 and F2) presence. Formant frequency measurements were made based on a 22-pole LPC spectrum computed over 20-ms Hamming-windowed segments. The LPC spectrum was obtained using a 2048-point FFT, yielding a 7.8-Hz frequency resolution. The frequencies of the first seven spectral peaks were extracted from the LPC spectrum every 20 ms. In order to get reliable F1 and F2 frequency estimates, we estimated the formant frequencies manually rather than using a peak-picking algorithm. An interactive MATLAB program was used that allowed the user to select among multiple spectral peaks the peaks corresponding to F1 and F2. This was done successively for each 20-ms segment of the vowel. The clean vowel spectrum was overlaid to the noisy vowel spectrum in order to get a rough estimate on the location of the F1/F2 frequencies of the noisy vowel spectra. Knowledge of acoustic phonetics played a role in the editing process, particularly knowledge about the close proximity of F1 and F2 for vowels such as /a/ and /u/.

Although it is relatively easy to identify (at least manually) F1 and F2 in quiet or relatively high S/N conditions, it is extremely difficult to identify F1 and F2 in extremely low S/N conditions. For that reason, F1 and F2 measurements were made only when the user felt confident that the selected peaks represented F1/F2 and not noise. For consistency purposes, several rules were adopted which classified each frame into four categories: (1) F1 not present, (2) F2 not present, (3) neither F1 nor F2 present, and (4) F1 and F2 reliably present. The percentage of frames that fell in each of the four categories was recorded for further analysis.

A frame was classified into category 1 (F1 not detected) whenever two or more peaks were present in the proximity[1] of the F1 region of the noisy vowel spectrum. A frame was classified into category 2 (F2 not detected) if either of the following two conditions were satisfied: (a) two or more

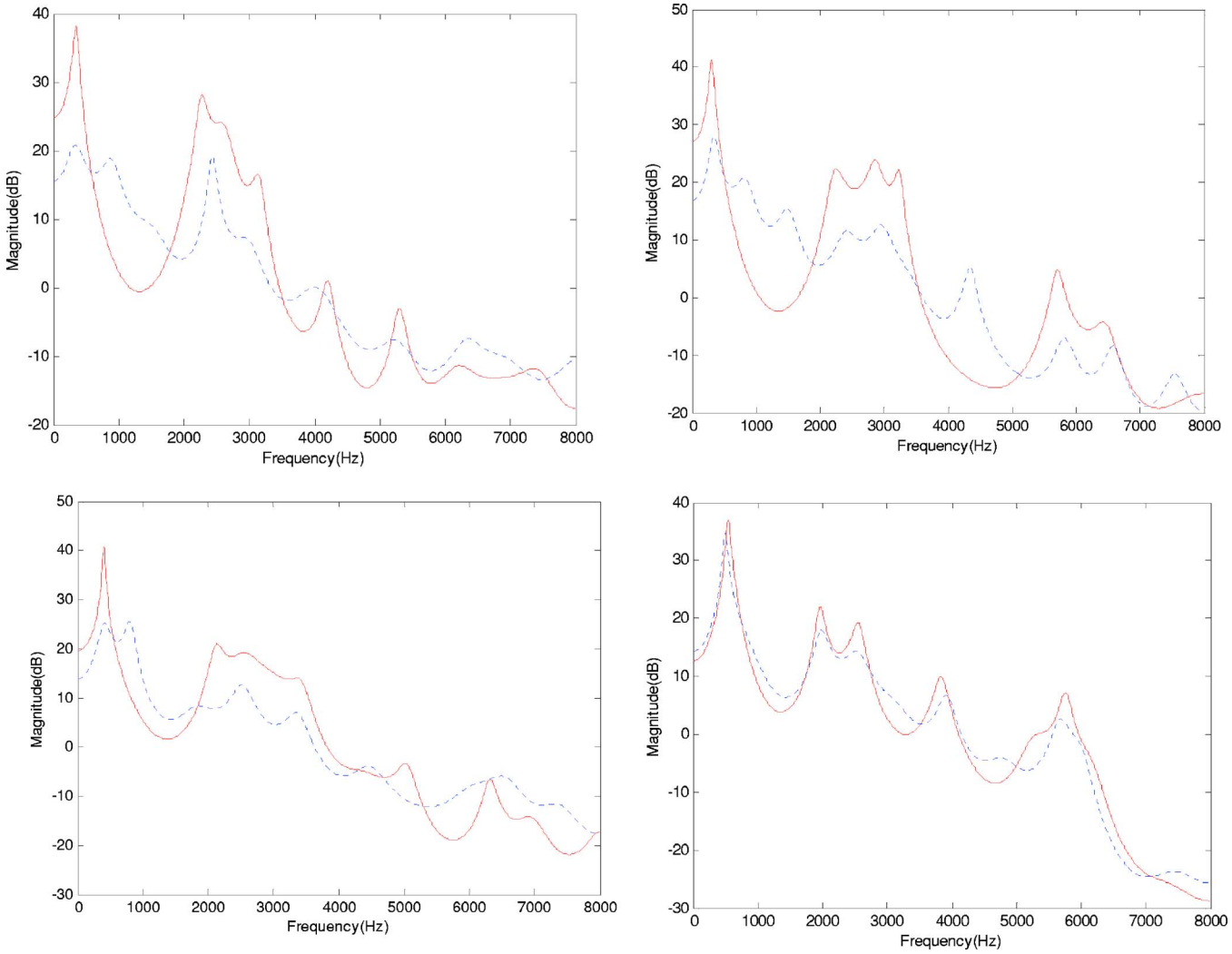

FIG. 2. (Color online) Example LPC spectra showing the four different categories used to label detection of F1 and F2. Solid lines indicate the clean spectra, and dashed lines indicate the corrupted vowel spectra. (Top left) Category 1: F1 not reliably detected because of multiple peaks in the proximity of F1 region. (Top right) Category 2: F2 not reliably detected because multiple peaks were present in the F1-F2 region. (Lower left) Category 3: Neither F1 nor F2 reliably detected. (lower right) Category 4: Both F1 and F2 reliably detected.

peaks were present in the proximity of the F2 region of the noisy vowel spectrum and (b) multiple peaks were present in the F1-F2 frequency range. A frame was classified into category 4 whenever a single peak was found near the F1 and F2 regions. Figure 2 shows example spectra from each category.

For the frames in which both F1 and F2 formants were reliably detected, we performed additional analysis to compare the formant frequencies estimated in noise with those estimated in quiet. We computed the absolute difference between the formant frequencies as follows:

$$\Delta F_1 = |F_1^q - F_1^n|,$$
$$\Delta F_2 = |F_2^q - F_2^n|, \tag{2}$$

where the superscript q indicates the formant frequency estimated in quiet and the superscript n indicates the corresponding formant frequency estimated in noise.

4. Acoustic analysis of consonants

Prior to analysis, the complete VCV consonant data set was manually segmented to [Vowel Consonant Vowel]. The starting and ending times of the burst were measured by hand from high-resolution digital spectrograms and displays of the time waveform. Only the first 10 ms of the release burst was considered, or the whole burst if smaller than 10 ms.

a. Burst frequency measurements. The burst frequency is defined as the frequency corresponding to the largest spectral magnitude of the burst spectrum (Liberman *et al.*, 1952). Burst frequency measurements were made using a 20-pole LPC spectrum. A Hamming window of 20 ms was used centered at the onset of the burst. In effect, only the latter half of the Hamming window was multiplied with the 10 ms of the release burst samples. The LPC spectrum was obtained using a 512-point FFT.

The frequency of the maximum amplitude of spectral peak was extracted from LPC spectra using a global peak-picking algorithm. Table III gives the consonant burst fre-

TABLE III. Burst frequencies (in Hz) of stop-consonants in three vowel contexts.

Vowel context	Stop consonant					
	/p/	/b/	/t/	/d/	/g/	/k/
/i/	3575	1938	7321	6675	3252	3208
/a/	1357	711	5103	5190	1852	1873
/U/	797	840	4522	3811	1701	1701

quencies estimated in quiet. The differences in burst frequencies between the clean and noisy burst spectra were computed and recorded.

b. Spectral tilt measurements. Spectral tilt measurements were made based on a four-pole LPC spectrum estimated using a 512-point FFT (fast Fourier transform). A low-order LPC was chosen to capture the trend of the burst spectrum while avoiding unnecessary details (e.g., peaks and valleys) in the spectrum. The spectral tilt was computed by evaluating the LPC magnitude spectrum at two frequencies, at 1000 and 5000 Hz. Spectral tilt was calculated as the difference between the spectral magnitudes (in dB) at 1000 and 5000 Hz. This frequency range was chosen to ensure that F2 was included in the spectral tilt computation[2] (Smits *et al.*, 1996). The dB difference values can be easily converted into slopes (in dB/Hz) by dividing the difference values by 4000 Hz (=5000−1000). A positive slope is indicated by a positive difference (rising spectrum), a negative slope is indicated by a negative difference (falling spectrum), and a slope close to zero (diffuse spectrum) is indicated by difference values close to zero. Figure 3 shows example LPC spectra of /p/ and /t/ along with their corresponding spectral tilt measurements.

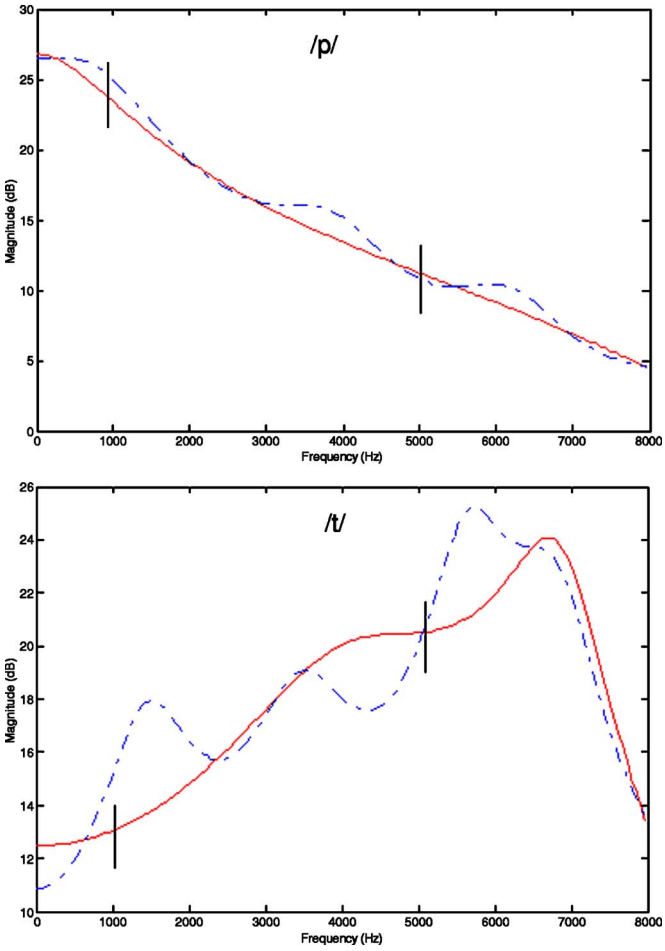

FIG. 3. (Color online) Example estimation of burst spectral slopes of /p/ and /t/. The dashed lines show the four-pole LPC spectra used for estimating the spectral slope, and the solid lines show the original burst spectra of /p/ and /t/. The vertical lines indicate the points (1000 and 5000 Hz) used to estimate the dB differences.

B. Results

1. Vowels

The acoustic measurements of the critical-band difference metric and number of formants reliably detected are shown in Figs. 4 and 5, respectively.

a. Critical-band spectral difference metric. Figure 4 shows the average critical-band spectral differences between clean and noisy spectra for each of the two frequency bands considered and for vowels embedded in speech-shaped noise (top panel) and multi-talker babble (bottom panel). Repeated measures ANOVA of the mean spectral difference (between noisy and clean vowels) of each vowel using noise type (multi-talker babble and speech-shaped noise), S/N level, and frequency band (low- and middle-frequency bands) as factors indicated a significant effect of noise type [$F(1,10)=7.18$, $p=0.023$], a significant effect of S/N level [$F(3,30)=11.7$, $p<0.005$], and a nonsignificant effect of frequency band [$F(1,10)=4.3$, $p=0.065$]. All the interactions between the within-subject factors were found to be significant ($p<0.05$). The interactions were partly caused by the fact that noise (babble and steady-speech-shaped) affected the two frequency bands of the vowel spectra differently depending on the S/N level. These interactions are probed in more detail below.

As shown in Fig. 4, the critical-band spectral difference between clean and noisy vowels decreased as the S/N increased. The largest spectral difference between the noisy and clean vowel spectra occurred in the mid-frequency band (1–2.7 kHz). This suggests that the F2 region was heavily masked by the noise. The F2 region was masked significantly ($p=0.049$) more than the F1 region (low frequency band) only for vowels corrupted by multi-talker babble at −5 dB S/N. For vowels corrupted by speech-shaped noise, the difference between the spectral distance measurements obtained for the low- and mid-frequency bands was not significant at any S/N level. Further *t* tests, with Bonferroni correction, indicated no statistically significant differences between the corresponding critical-band difference metrics for the two types of noise at any S/N level, suggesting that the two types of noise examined in this study affected the spectrum in the same way, at least for S/N levels higher than −5 dB.

b. Counts of formant frequencies. For reliability purposes, a second experimenter worked independently and remeasured 20% of the vowels embedded in the various S/N conditions. The second experimenter made use of the same software used by the first experimenter and the same rules for classifying the frames into one of the four categories.

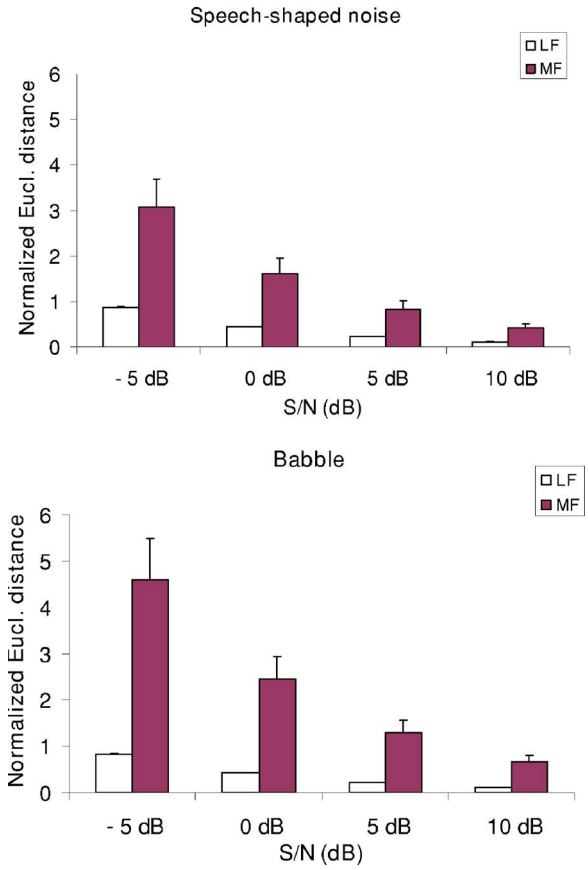

FIG. 4. (Color online) Mean spectral envelope differences (based on the Euclidean distance metric) between the clean vowel spectra and noisy vowel spectra in the low (LF) and middle (MF) frequency regions. Error bars indicate standard errors of the mean.

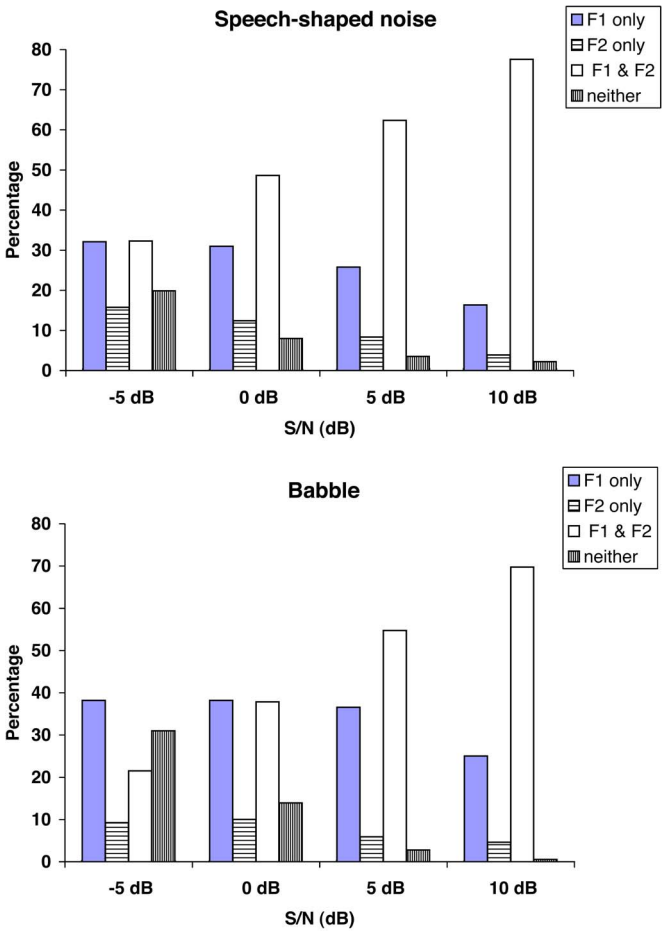

FIG. 5. (Color online) Percentage of vowel frames in which the F1 and/or F2 formants were reliably detected in various noise conditions.

Results from recounting the number of F1 and F2 peaks detected reliably in noise indicated an average difference of less than 1% (across all S/N conditions) between the measurements of the two experimenters.

Figure 5 shows the percentage of frames (out of a total of 539 frames) in which F1 and/or F2 were reliably identified for vowels embedded in speech-shaped noise (top panel) and for vowels embedded in multi-talker babble (bottom panel) at various S/N levels. For the low S/N conditions, the first formant (F1) was reliably identified more often compared to the second formant (F2). According to nonparametric tests (Mann-Whitney test), F1 was identified significantly ($p < 0.05$) more often than F2 in both types of noise.

F1 and F2 formants were identified significantly ($p = 0.036$) more often (according to Mann-Whitney tests) in −5 dB S/N speech-shaped noise than in −5 dB S/N multi-talker babble. No significant difference was found between the number of the two formants (F1 and F2) identified in babble and speech-shaped noise in other S/N conditions. F1 and F2 formants were reliably identified more than 50% of the time only for S/N values of 5 dB and higher. In multi-talker babble (5 dB S/N), the F1 and F2 formants were identified 54.6% of the time, while in speech-shaped noise the two formants were identified 62% of the time.

For the frames in which both formants were reliably identified, we computed the differences [ΔFs as per Eq. (2)] between the true (estimated in quiet) formant frequencies and the estimated formant frequencies in noise, averaged across all vowels. The results are tabulated in Table IV for +5 and +10 dB S/N. As can be seen from Table IV, the differences in formant frequencies (ΔF's) were extremely small. In fact, the ΔF values were close to the difference limens (DLs) of formant frequencies, known to be in the order of 1%–2% of the formant frequencies (Flanagan, 1955; Hawks, 1994).

2. Consonants

Acoustic measurements of the burst frequency and slope are shown in Figs. 6 and 7, respectively.

a. Burst frequency measurements. Figure 6 shows the burst frequency differences between measurements made in noise and in quite, averaged across all stop consonants for

TABLE IV. Mean absolute differences in F1 and F2 values between formant frequencies estimated in quite and those corrupted in +5 and +10 dB S/N. These differences were estimated only for frames which contained reliable F1 and F2 information.

Noise type	S/N level (dB)	ΔF1 (Hz)	ΔF2 (Hz)
Multi-talker	5	19.3	27.6
Speech shaped		15.8	22.8
Multi-talker	10	12.4	20.0
Speech shaped		11.1	16.9

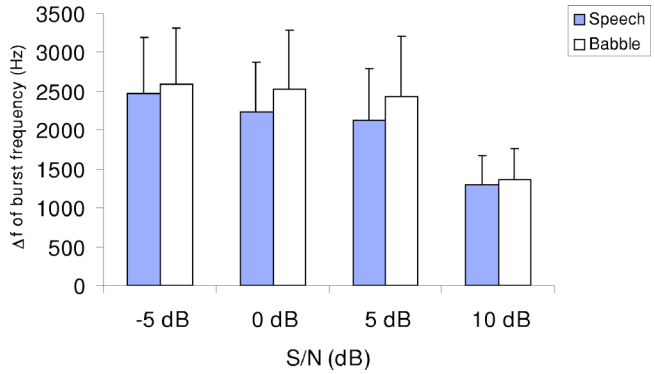

FIG. 6. (Color online) Mean shift in burst frequency of noisy stops relative to the burst frequency estimated in quiet. Error bars indicate standard deviations

different S/N levels and noise conditions. Repeated measures ANOVA of the burst frequency deviation (difference) of each stop consonant using noise type (multi-talker babble and speech-shaped noise) and S/N level as factors indicated a significant effect of noise type [$F(1,5)=10.48$, $p=0.023$], a significant effect of S/N level [$F(3,15)=5.26$, $p=0.011$], and a nonsignificant interaction between noise type and S/N [$F(3,15)=2.59$, $p=0.091$]. Posthoc comparisons indicated a significantly ($p=0.016$) larger deviation of the burst frequency for multi-talker babble at 5 dB S/N. No significant differences were found at the other S/N levels between the deviations of the burst frequencies for the two types of noise.

b. Spectral slope measurements. Figure 7 shows the mean spectral slopes of the labial and alveolar consonant bursts at various S/N levels for speech-shaped noise (top panel) and multi-talker babble (bottom panel). The slopes were given in terms of dB differences of the burst spectra sampled at 5000 and 1000 Hz. The dB difference values can be easily converted into slopes (in dB/Hz) by dividing the difference values by 4000 (=5000−1000) Hz. A rising spectrum is indicated by a positive difference (positive slope), a falling spectrum is indicated by a negative difference (negative slope), and a diffuse spectrum is indicated by difference values close to zero. Repeated measures ANOVA of the burst spectral slope of the stop consonants /b d p t/, using noise type (multi-talker babble and speech-shaped noise) and S/N level as factors, indicated a nonsignificant effect of noise type [$F(1,3)=0.0001$, $p=0.99$], a significant effect of S/N level [$F(3,9)=50.1$, $p<0.0005$], and a nonsignificant interaction between noise type and S/N [$F(3,9)=0.087$, $p=0.96$]. The effect of noise type was nonsignificant, suggesting that both types of noise affected the slope of the burst spectra the same way. This can be partly explained by the fact that both types of noise—babble and continuous speech-shaped noise—had low-frequency dominance (Fig. 1) and consequently affected the slope of the burst spectra the same way.

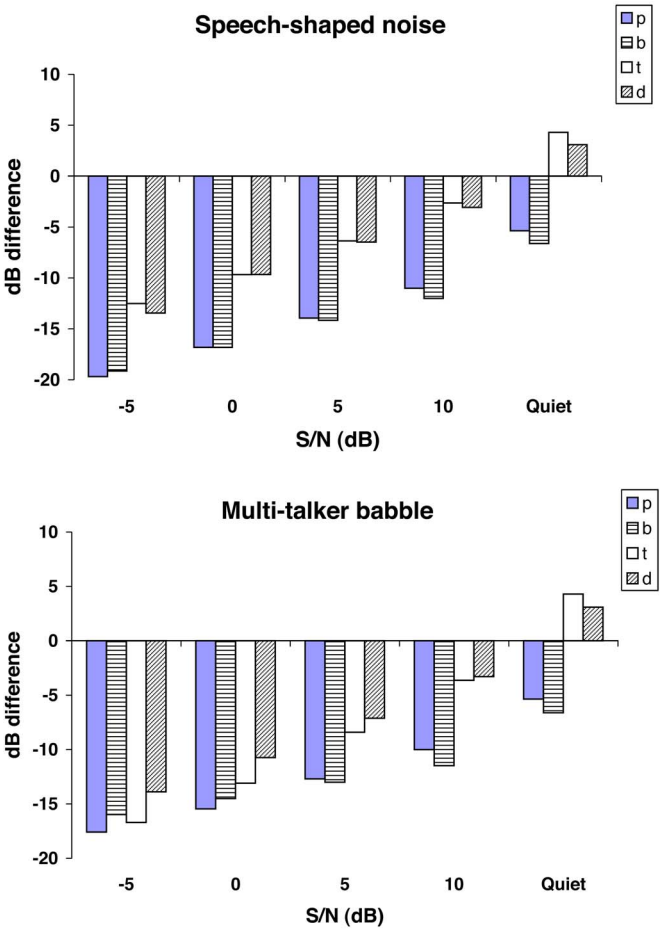

FIG. 7. (Color online) Spectral slope measurements of the burst spectra in terms of differences between the spectral magnitudes (in dB) at 1000 and 5000 Hz. A positive dB difference indicates a rising spectrum and a negative dB difference indicates a falling spectrum.

C. Discussion

Acoustic analysis indicated that the two types of noise (multi-talker babble and speech-shaped noise) examined in this study affected the vowel and stop-consonant spectra in a similar way. The differences between babble and speech-shaped noise were only prominent in extremely low S/N (−5 dB) conditions. This was found to be true for nearly all the acoustic parameters examined.

The vowel spectral difference measurements indicated a nonuniform effect of noise in the various frequency bands, as expected. The F2 region (1–2 kHz) was affected the most (at least in −5 dB S/N), suggesting that the second formant was heavily masked by the noise. This conclusion is also supported by the formant frequency data (Fig. 5) which indicated that F2 was not detected as reliably as F1. In contrast, F1 was detected more reliably and a smaller spectral difference between noisy and clean vowels was obtained for the F1 region. This outcome is consistent with the findings of the study by Diehl *et al.* (2003), i.e., the F1 peak is more resistant to noise than the F2 peak. These findings suggest that in noise listeners must be identifying vowels with a good F1 representation but a poor F2 representation. The question of whether this will impair vowel recognition will be investigated in experiment 1.

The data from the spectral difference measurements are not only important for understanding vowel perception in noise, but are also important for the development of noise-reduction algorithms. These findings point to a multi-band approach for noise reduction in which individual frequency

bands are treated differently by taking into account the non-uniform effect of noise. In spectral-subtractive-type algorithms (e.g., Boll, 1979), for instance, different rules could be applied to the low-, middle-, or high-frequency regions of the spectrum. Such an approach was proposed in Kamath and Loizou (2002) (see also Kamath, 2001) and could easily be extended to multi-channel hearing aids and multi-channel cochlear implants.

The formant frequency measurements indicated that F1 was identified reliably more often than F2. In multi-talker babble at −5 dB S/N, F1 was identified 60% of the time while F2 was identified only 30% of the time (these values were obtained by summing the percentages of "F1 only" and "F1&F2" in Fig. 5). In speech-shaped noise at −5 dB S/N, F1 was detected 64% of time, while F2 was detected 48% of the time. There were substantial differences in reliable detection of F1 and F2 across vowels. In multi-talker babble (−5 dB S/N), for instance, the F1 of the vowel /ɝ/ was identified most reliably (89%), while the F1 of the vowel /ɑ/ was identified least reliably (34%). F1 was generally identified more reliably in vowels with low F1 values (e.g., /i/, /e/). No clear pattern emerged for F2 identification. The F2 of the vowel /ɑ/ was identified most reliably (47%), while the F2 of the vowel /u/ was identified least reliably (11%). Significant differences between the two types of noise in percentage of formants identified were noted only for the extremely low S/N level, −5 dB. These findings are also supported by the critical-band spectral difference measurements—noise affected the F2 region more than the F1 region. Since the first two formants are known to be the primary cues to vowel identification, one would expect to find a strong correlation between presence of reliable F1 and F2 information and vowel identification, and this is investigated in experiment 1.

As shown in Fig. 7, the spectral tilt was severely affected. The spectral tilt of the alveolar stops (/t/, /d/), which is typically positive (see Fig. 3), became negative for S/N levels lower than 10 dB. Because of the low-pass nature of the noise (Fig. 1), the alveolar stops had now a falling spectrum. The spectral tilt of the labial stops (/b/, /p/) was not affected; the labial spectra remained falling, but had a more negative slope. It is questionable whether the change in spectral tilt alone will affect stop-consonant identification in noise, and this is investigated in experiment 1.

Consistent with the large changes in spectral tilt, the burst frequencies were also greatly affected. The average shift in burst frequency was 2500 Hz for S/N=−5, 0, and 5 dB, and 1500 Hz for S/N=10 dB. Particularly large shifts of about 5000 Hz were observed for the alveolar stops, consistent again with the reversal of the direction (sign) of the spectral slope. Even a small amount of noise (S/N=10 dB) produced a large shift in burst frequency, suggesting that the burst frequency might not be a robust cue to place of articulation, at least in noisy conditions.

III. EXPERIMENT 1: VOWEL AND STOP-CONSONANT IDENTIFICATION IN NOISE

The acoustic analysis quantified the effect of noise on the vowel and stop-consonant spectra. Several acoustic parameters such as the Euclidean distance metric between the noisy and clean critical-band spectra, the number of formants reliably detected, the spectral tilt, and so on, were extracted and analyzed. But, to what degree do these parameters correlate with perceptual data? This question is addressed in the present experiment.

A. Subjects

Nine normal-hearing listeners (20 to 30 years of age) participated in this experiment. All subjects were native speakers of American English. The subjects were paid for their participation. The subjects were undergraduate students (not trained in phonetics) from the University of Texas at Dallas.

B. Speech material

The same speech materials from acoustic analysis were used.

C. Procedure

The experiments were performed on a PC equipped with a Creative Labs SoundBlaster 16 soundcard. Stimuli were played to the listeners at a comfortable level through Sennheiser's HD 250 Linear II circumaural headphones. A graphical user interface was used that enabled the subjects to indicate their response by clicking a button corresponding to the word played.

The tests were conducted in two separate sessions, one for the vowels and one for the consonants. Prior to each test the subjects were presented with a practice session in which the identity of the test syllables (vowels or consonants) was displayed on the screen. In the practice session, the vowels and consonants were presented in quiet. In the test session, the vowels and consonants were completely randomized and presented in noise (−5, 0, 5, 10 dB S/N) and in quiet with no feedback. The order of the various S/N conditions was counterbalanced among subjects. Six repetitions per speaker group (female and male) were used for the vowel test, for a total of 12 repetitions per vowel. Three repetitions per vowel context and per speaker group were used for the consonant test, for a total of 18 repetitions (=3×2 speaker groups×3 vowel contexts) per consonant.

D. Results

The mean percent scores on vowel and stop-consonant identification are shown in Fig. 8 for different S/N levels and in quiet. Repeated measures ANOVA of the vowel scores, using noise type (multi-talker babble and speech-shaped noise) and S/N level as factors, indicated a significant effect of noise type [$F(1,8)=17.51$, $p=0.003$], a significant effect of S/N level [$F(2,16)=62.2$, $p<0.0005$], and a significant interaction between noise type and S/N level [$F(2,16)=11.04$, $p=0.001$]. Identification scores of vowels corrupted by multi-talker babble at −5 dB S/N were found to be significantly ($p=0.003$) lower than the scores of the vowels corrupted by speech-shaped noise. There was no statistically significant difference in vowel identification scores for the 0

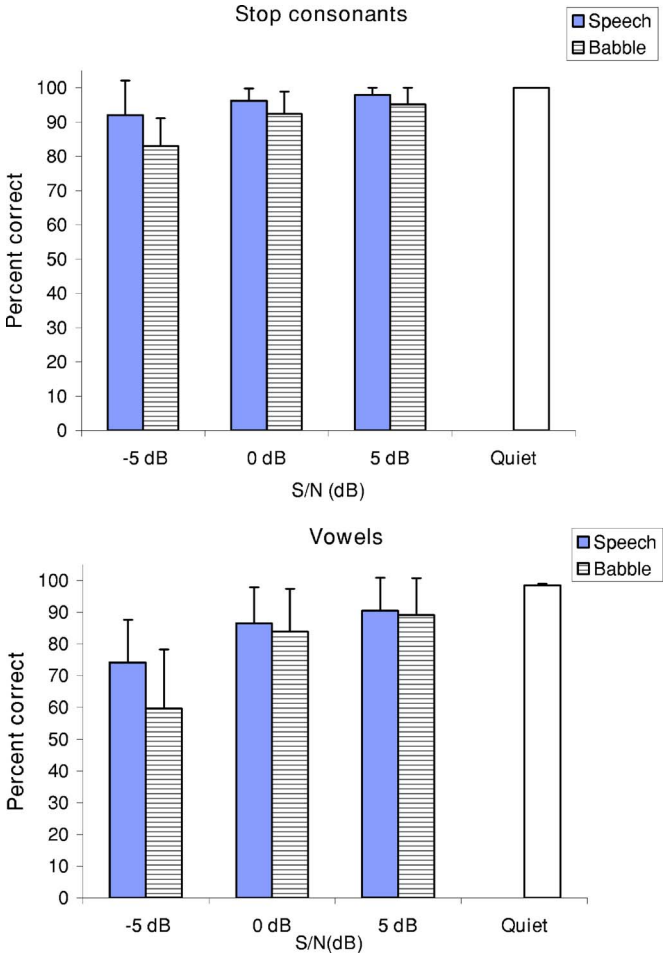

FIG. 8. (Color online) Mean stop-consonant and vowel recognition in noise and in quiet. Error bars indicate standard deviations.

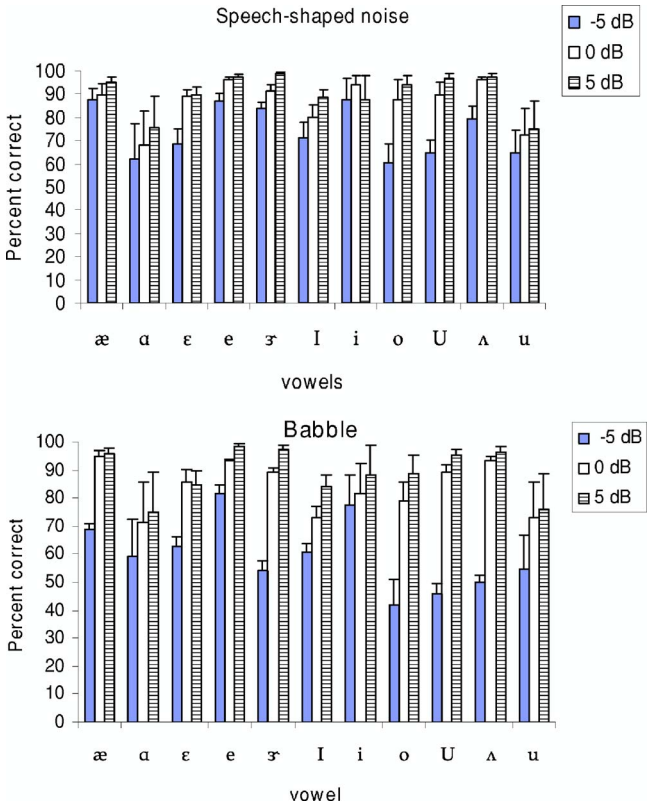

FIG. 9. (Color line) Individual vowel identification in noise. Error bars indicate standard deviations.

and 5 dB conditions between the two types of noise. In quiet, the vowels were identified with 98.5% accuracy,[3] while the stop-consonants were identified perfectly (100% correct) by all listeners.

The individual vowel identification scores are given in Fig. 9. Vowel identification remained high (90% correct) at 0 and 5 dB S/N levels, and dropped to 74% and 60% correct at −5 dB S/N for speech-shaped noise and babble respectively. At −5 dB S/N, the lowest score was obtained for the vowel /o/ and the highest score was obtained for the vowel /i/. The aggregate confusion matrix (compiled across all subjects) for the −5 dB condition is given in Table V. The vowels in the matrices are arranged in order of increasing F1. Adjacent vowels have therefore similar F1. The vowels in Table V are further arranged in four groups, enclosed in rectangles, according to whether they have generally low, medium, high, or very high F1 frequency. Most of the vowel confusions were along the main diagonal. The confusions which fell near the main diagonal and within the individual rectangles (groups) were caused by vowels with similar F1 but different F2. This suggests that listeners were using primarily F1 information to identify vowels, since the F2 region was heavily corrupted, but not necessarily obscured, by the noise (see Fig. 4; more on this in Sec. III E). The most dominant confusions between vowels with similar F1 included the pairs /o/-/ɝ/, /o/-/e/, /ɝ/-/U/, /I/-/U/, and /ɛ/-/ʌ/. Confusions that fell off the main diagonal and outside the rectangles indicated that the first formant was not perceived correctly. These included the confusions between the vowel pairs /ɑ/-/æ/, /U/-/u/, /I/-/ɛ/, /ʌ/-/ɑ/, and /ɛ/-/æ/. Note that the vowels in the pairs /U/-/u/, /I/-/ɛ/, and /ʌ/-/ɑ/ had similar F2.

ANOVA performed on the consonant scores indicated a significant effect of noise type [$F(1,8)=9.36$, $p=0.016$], a significant effect of S/N level [$F(2,16)=7.17$, $p=0.06$], and a significant interaction between noise type and S/N level [$F(2,16)=7.09$, $p=0.006$]. Identification scores of stop-consonants corrupted by multi-talker babble at −5 dB S/N were found to be significantly ($p=0.002$) lower than the scores of consonants corrupted by speech-shaped noise. There was no statistically significant difference in consonant identification between the two types of noise for the 0 or 5 dB conditions.

The individual stop-consonant identification scores are given in Fig. 10. Stop-consonant identification was impaired only at −5 dB S/N. At −5 dB, the lowest scores were obtained for the labial consonants /b/ and /p/. We suspect that this is because the babble noise (and speech-shaped noise) masks the low frequencies more than the high frequencies. The alveolar consonants have a spectral prominence in the mid to high frequencies and are therefore less susceptible to masking by the type of noise used in this study (Fig. 1). The aggregate confusion matrix (compiled across all subjects) for the −5 dB condition is given in Table VI. As can be seen, most of the confusions were place errors, consistent with the findings of Miller and Nicely (1955).

TABLE V. Aggregate confusion matrices obtained in identification of vowels corrupted with noise at −5 dB S/N.

Speech-shaped noise (S/N=-5 dB)

	/u/	/i/	/ɪ/	/ʊ/	/e/	/ɝ/	/o/	/ɛ/	/æ/	/ʌ/	/ɑ/
/u/	65	4	2	22	1	3	2			1	
/i/		88	11					1			
/ɪ/	2		88	2				8			
/ʊ/	2		6	65		7	5	4	2	8	1
/e/	1	1	6	2	87		1	1	1		
/ɝ/	2			11		84	2		1		
/o/	5		1	12	2	15	60	1	1	1	2
/ɛ/			6	2		5		69	10	8	
/æ/				1	1	1		6	87		4
/ʌ/				1		1	1	7	5	79	6
/ɑ/			10		5			2	19	2	62

Babble (S/N=-5 dB)

	/u/	/i/	/ɪ/	/ʊ/	/e/	/ɝ/	/o/	/ɛ/	/æ/	/ʌ/	/ɑ/
/u/	55	5	4	23	4	4	4			1	
/i/	2	77	14		3	1	1	1		1	
/ɪ/		1	60	16		3	1	13	3	2	1
/ʊ/	4		11	46	3	9	2	9	5	7	4
/e/	2	1	5		82	1	3	4	1	1	
/ɝ/	6	1	6	12	5	54	8	1	2	3	2
/o/	5	1	2	8	11	23	42	3		2	3
/ɛ/	1	2	4	1	2	2	1	63	6	17	1
/æ/		2	2	1	2	2	1	5	69	3	13
/ʌ/	1	1	5	4	3	2	2	10	6	50	16
/ɑ/		2	5	2	3	3	1		20	5	59

Correlation analysis (see Table VII) was performed between the vowel identification scores and the corresponding acoustic parameter values. For the critical-band difference metric, separate analysis was performed for the low-frequency (LF) band, the middle-frequency (MF) band, the low- and middle-frequency bands (LF+MF) combined and the whole spectrum (indicated as WF in Table VII). In the whole-spectrum analysis, the normalized Euclidean metric [Eq. (1)] was computed between the clean and noisy vowel spectra taking the whole bandwidth (0–8 kHz) into consideration, i.e., all 21 critical bands were included in the summation in Eq. (1). For the formant-frequency count data, separate correlation analysis was performed between the percentage of F1 identified and vowel identification scores, between the percentage of F2 identified and vowel scores, and between the percentage of both F1 and F2 (F1+F2) identified and vowel scores. Pearson's correlation analysis was used for the critical-band difference metric, and nonparametric Spearman's correlation analysis was used for the formant-frequency count data. Multiple-linear regression analysis was used for the LF+MF and F1+F2 data. The correlation coefficients (r) along with the corresponding p values are tabulated in Table VII for both speech-shaped noise and multi-talker babble at −5 and 0 dB S/N. No correlation was performed for the 5 dB condition because of ceiling effects.

A modestly high correlation (r=0.724) was found between the critical-band difference metric of the low frequency band [LF; Eq. (1a)] and vowel identification in multi-talker babble at −5 dB S/N (see Table VII). The correlation coefficient (r=0.558) for the LF band obtained for vowels in speech-shaped noise was not significant (p=0.07). A significant correlation was also found between the critical-band difference metric of the combined LF+MF bands and vowel

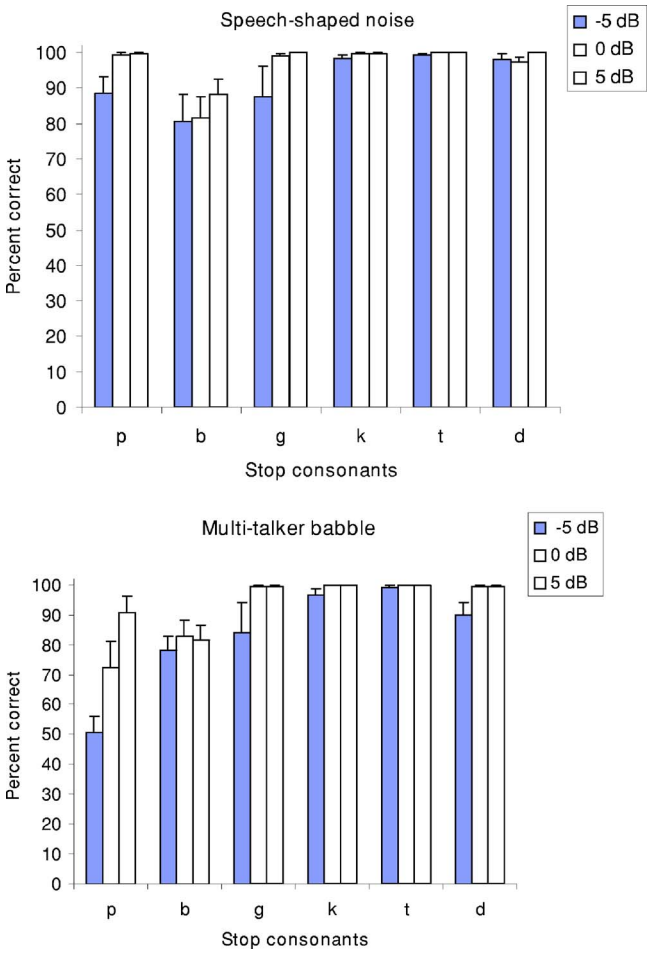

FIG. 10. (Color online) Individual stop-consonant identification in noise. Error bars indicate standard deviations.

TABLE VI. Aggregate confusion matrices obtained in identification of stop-consonants corrupted with noise at −5 dB S/N.

	Speech-shaped noise (S/N=−5 dB)					
	/b/	/d/	/g/	/k/	/p/	/t/
/b/	80	17	1		1	1
/d/	1	98				1
/g/		1	87	10		1
/k/			1	98	1	
/p/				11	88	
/t/						99
	Babble (S/N=−5 dB)					
/b/	78	6	2		2	12
/d/		90	9			
/g/		6	84	8		1
/k/				97	1	1
/p/	1	1	3	33	50	13
/t/		1				99

identification in −5 dB S/N for both multi-talker babble and speech-shaped noise conditions. No other significant correlations were found at −5 and 0 dB S/N in either babble or speech-shaped conditions. No significant correlations were found between the count of formant frequencies (F1, F2, F1+F2) and vowel identification at −5 and 0 dB S/N for either babble or speech-shaped noise conditions.

E. Discussion

One of the main goals of this study was to understand the cues used by listeners for vowel and stop-consonant recognition in noise. The good identification performance in vowel and stop recognition at −5 dB S/N challenged some of the traditional cues (e.g., formant frequencies, release burst, etc.) known to be important for vowel and stop perception, at least in quiet.

Vowel identification scores remained moderately high (75% correct in speech-shaped noise and 60% correct in babble) at −5 dB S/N despite the absence of clear and coherent F1 and F2 information. We exclude the possibility that listeners utilized exclusively formant frequency cues (F1 and F2 frequencies) for the main reason that the F2 region was heavily masked by the noise and F2 was not reliably identified (see Fig. 5). Only access to F1 information was reliable as evidenced by the formant frequency count data (Fig. 5). We also rule out the possibility that listeners utilized exclusively whole-spectrum shape cues to identify vowels because vowels with dissimilar spectral shapes were not identified correctly (Table V). This is illustrated in the example shown in Fig. 11 which compares the excitation patterns of the vowels /ɛ/ and /ʌ/ embedded in −5 dB babble. The excitation pattern of the vowel /ɛ/ is distinctly different from that of the vowel /ʌ/ as it is characterized by two peaks in the F2 region. In contrast, the excitation pattern of the vowel /ʌ/ has no peak in the F2 region, as it has a low F2 frequency. Yet, the vowel /ɛ/ was confused with the vowel /ʌ/ 17% of the time in babble noise (S/N=−5 dB). Similarly, the confusion of /o/ with /e/ cannot be explained by a pattern matching mechanism which relies on spectral shape information, since the excitation patterns of these vowels differ greatly in the F2 region (the difference between the F2 frequencies of these two vowels is larger than 1000 Hz). Further evidence is provided by the absence of correlation between the whole-spectrum difference metric (WF) and vowel identification scores (see Table VII).

TABLE VII. Correlation analysis between critical-band spectral differences of noisy and clean vowels for four different frequency regions (LF =0−1 kHz, MF=1−2.7 kHz, LF+MF=0−2.7 kHz, and WF=0−8 kHz) and vowel identification scores (top). Correlation analysis between presence of F1/F2 frequency information in noisy vowels and vowel identification scores (bottom). Correlation coefficients (r) with the corresponding p values are given in the two rightmost columns. Sample size for the correlation analysis was $n=11$.

Noise type	S/N (dB)	Independent variable	r	p
Speech shaped	−5 dB	LF	0.558	0.075
		MF	−0.491	0.125
		LF+MF	**0.733**	0.046
		WF	0.358	0.280
	0 dB	LF	0.183	0.589
		MF	−0.169	0.620
		LF+MF	0.251	0.770
		WF	0.135	0.692
Multi-talker babble	−5 dB	LF	**0.724**	0.012
		MF	−0.468	0.146
		LF+MF	**0.850**	0.006
		WF	0.474	0.141
	0 dB	LF	0.006	0.985
		MF	−0.355	0.285
		LF+MF	0.355	0.584
		WF	−0.090	0.791
Speech shaped	−5 dB	F1	−0.351	0.290
		F2	0.486	0.129
		F1+F2	0.369	0.264
	0 dB	F1	0.064	0.854
		F2	0.073	0.830
		F1+f2	0.021	0.950
Multi-talker babble	−5 dB	F1	−0.282	0.401
		F2	0.032	0.926
		F1+F2	0.100	0.770
	0 dB	F1	0.027	0.936
		F2	0.016	0.962
		F1+F2	0.279	0.406

The data from Figs. 4 and 5 taken together suggest that noise (babble and speech-shaped) produced vowels that have a poor (but not necessarily obscure) F2 representation, but a reasonably good F1 representation. Listeners must therefore be relying primarily on F1 frequency information to identify vowels in noise. Evidence of this is provided by the pattern of the vowel confusion errors (Table V). The most dominant confusions occurred along the main diagonal, i.e., between vowels that had similar F1 but different F2 (recall that the vowels in Table V are arranged in increasing F1, hence adjacent vowels have similar F1). Among the most dominant

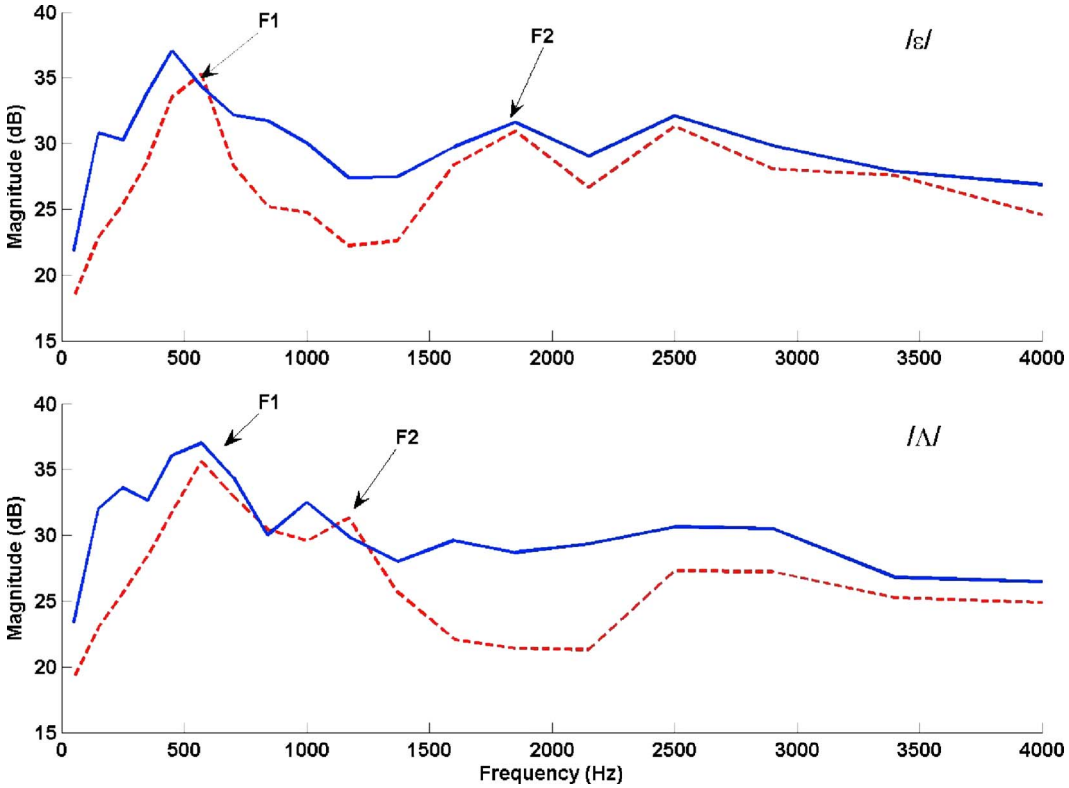

FIG. 11. (Color online) Excitation patterns of the vowel /ɛ/ (top panel) and vowel /ʌ/ (bottom panel) estimated in quiet (dashed lines) and in −5 dB S/N babble noise (solid lines). These patterns were computed based on a 20-ms segment of the vowels extracted from the steady-state portion of the words "head" and "hud" produced by a male speaker. Arrows show the locations of the F1 and F2 formants in quiet.

confusions included the pairs /o/-/ɜ˞/,/o/-/e/,/ɜ˞/-/U/,/I/-/U/, /ɛ/-/ʌ/. Hence, although listeners might not have a coherent idea on the location of *both* F1 and F2 frequencies, they had a good indication about the location of F1 and only a vague idea about the location of F2. Previous studies (e.g., Dubno and Dorman, 1987) have shown that that alone is sufficient for vowel identification. In the study by Dubno and Dorman (1987), for instance, listeners identified with greater than 95% accuracy six (synthetic) front vowels which had a normal F1 but higher frequency formants represented by a broad spectral plateau ranging from 1600 to 3500 Hz (i.e., F2 and F3 were flattened). We find it unlikely that a spectral-shape pattern matching mechanism based on the F1-region alone is responsible for the confusion patterns shown in Table V. We base this assertion on prior evidence from the study by Beddor and Hawkins (1990) where they showed that the perceived quality of vowels with prominent F1 peak is dominated primarily by the frequency of F1 rather than the spectral envelope. Our data from the critical-band difference metric (Fig. 4) suggest that the F1 peak remained prominent in noise.

The critical-band difference metric and the counts of F2 detection suggest a poor F2 representation due to the relatively heavy masking of the F2 region by noise. Despite that, the F2 region seems to contain some useful information. Evidence of this is given by the lack of confusion errors between the vowels /i/ and /u/. These vowels have nearly identical F1 but markedly different F2 frequencies (as much as 1300 Hz). We would therefore expect a large number of confusions between these two vowels if noise had completely obscured all information in the F2 region. But, the vowels /i/ and /u/ were rarely confused with each other by the listeners (Table V). Hence, listeners must be utilizing, in addition to F1 frequency information, some other form of information about F2. One possibility is that listeners are making use of partial spectral shape information contained in the F1-F2 region. This observation is collaborated with the data from the correlation analysis of the critical-band difference metric and the identification scores. A significant correlation was found between spectral envelope differences of noisy and clean vowels in the low-to-mid frequencies (based on the composite LF+MF difference metric) and vowel identification at −5 dB S/N. In the absence of enough formant frequency information listeners must be relying on other cues to identify vowels in extremely low S/N conditions. Other cues probably used by listeners (but not examined in this study) include vowel duration, spectral change, and formant contours. The salience of spectral change and formant contour cues in extremely low S/N conditions, however, is questionable and needs further investigation.

Stop-consonant identification remained high (80%–90% correct) even at −5 dB S/N. This was achieved despite the fact that both spectral tilt and burst frequency were significantly altered by the noise. The tilt of the alveolar burst spectra became negative after adding noise. Yet, the alveolar stops were identified perfectly (see Fig. 10) even at −5 dB. Previous studies (e.g., Smits *et al.*, 1996) have shown that the velar bursts are efficient cues to place of articulation, at least in quiet. In our study, the velar burst spectra were severely altered by noise, yet the identification of /k/ remained

at near 100%, and the identification of /g/ was only modestly affected at −5 dB. In summary, changes to the burst spectra did not impair stop-consonant identification. This suggests that in the presence of noise, the burst cues become unreliable and listeners must be relying on other cues, perhaps formant transitions and/or spectral change, to identify stops.

The present study assessed the perceptual and acoustic differences between babble and continuous speech-shaped noise. The data from acoustic analysis indicate no significant differences in the way the two types of noise affected the speech spectra, at least for S/N levels higher than and including 0 dB. Consistent with the data from acoustic analysis, listeners identified vowels and stop-consonants corrupted by babble and speech-shaped noise with the same accuracy at 0, 5, and 10 dB S/N. Identification scores were significantly lower only for multi-talker babble at −5 dB S/N. In brief, the data from this study indicate that the two types of noise examined are perceptually and acoustically equivalent, at least for low to moderate S/N levels (0–10 dB) and for single words presented in isolation.

IV. SUMMARY AND CONCLUSIONS

This study assessed the perceptual and acoustic effects of multi-talker babble and continuous speech-shaped noise on the vowel and stop-consonant spectra. Noise was added to vowel and stop-consonant stimuli at −5 to +10 dB S/N, and the spectra were examined by means of acoustic analysis. Acoustic analysis indicated the following.

(i) At −5 dB S/N, the largest spectral envelope difference between the noisy and clean vowel spectra occurred in the mid-frequency band (1–2.7 kHz) for both types of noise. This suggests that the second formant is heavily masked by noise at very low S/N levels.
(ii) The first formant (F1) was reliably detected in noise more often than F2, suggesting that listeners had access to correct F1 information but vague information about F2.
(iii) There was a large shift (of about 2500 Hz) of the apparent burst frequency of the stop consonants with the addition of noise.
(iv) The spectral slope of the consonant burst spectra was severely affected by noise. The alveolar stops (/d t/) which originally had a positive slope (rising spectrum) had now a negative slope (falling spectrum) for all S/N levels except for +10 dB S/N. This pattern was consistent with both types of noise.

The above acoustic parameters were subjected to correlation analysis with vowel and stop-consonant identification scores collected in experiment 1. The perception study and correlation analysis indicated the following.

(i) Vowel identification scores remained moderately high (75% correct in speech-shaped noise and 60% correct in babble) at −5 dB S/N despite the absence of clear and coherent F1 and F2 information. Based on acoustic analysis data (Figs. 4 and 5), we infer that at low S/N conditions listeners must be relying on relatively accurate F1 frequency information along with partial F2 information to identify vowels in noise.
(ii) Identification scores of vowels and consonants corrupted by multi-talker babble at −5 dB S/N were significantly lower than the corresponding scores obtained in speech-shaped noise. No significant differences were observed between identification scores of vowels (and consonants) corrupted by multi-talker babble and speech-shaped noise at S/N levels higher than −5 dB.
(iii) Stop-consonant identification remained high (80%–90% correct) even at −5 dB S/N, despite the fact that both the spectral tilt and burst frequency were significantly altered by the noise. This suggests that listeners must be relying on other cues, perhaps formant transitions, to identify stops.

ACKNOWLEDGMENTS

This research was supported by Grant No. R01 DC03421 from the National Institute of Deafness and other Communication Disorders, NIH. This project was the basis for the Master's thesis of the first author (G.P.) in the Department of Electrical Engineering at the University of Texas—Dallas. We would like to thank Kalyan Kasturi for all his help with the listening tests. Many thanks to Dr. Alexander Francis and the anonymous reviewers for their comments.

[1]We define F1 proximity as the frequency region that extends from F1 to the mid-point between F1 and F2, i.e., from F1 to F1+(F1+F2)/2. Similarly F2 proximity is defined as the frequency region encompassing F2±(F1+F2)/2.
[2]A similar procedure was used in Lahiri *et al.* (1984) to measure the spectral tilt. In their study, spectral tilt was obtained by drawing a straight line by hand through the F2 and F4 peaks in the LPC spectrum.
[3]The normal-hearing listeners in the Hillenbrand *et al.* (1995) study identified a large set of vowels with 95.4% accuracy (our study included a subset of those vowels). In their study, however, listeners were also presented with vowels produced by children and the test vowel set included the vowel /ɔ/, which was excluded from the present study. The vowel /ɔ/ was excluded because many speakers of American English do not maintain the /ɑ/-/ɔ/ distinction.

Ainsworth, W. A. (**1972**). "Duration as a cue in the recognition of synthetic vowels," J. Acoust. Soc. Am. **51**, 648–651.
Beddor, P., and Hawkins, S. (**1990**). "The influence of spectral prominence on perceived vowel quality," J. Acoust. Soc. Am. **87**, 2684–2704.
Bladon, R. (**1982**). "Arguments against formants in the auditory representation of speech," in *The Representation of Speech in the Peripheral Auditory System*, edited by R. Carlson and B. Grandstrom (Elsevier Biomedical, Amsterdam), pp. 95–102.
Bladon, R., and Lindbhlom, B. (**1981**). "Modeling the judgment of vowel quality differences," J. Acoust. Soc. Am. **69**, 1414–1422.
Blumstein, S., and Stevens, K. (**1980**). "Perceptual invariance and onset spectra for stop consonants in different vowel environments," J. Acoust. Soc. Am. **67**, 648–662.
Blumstein, S., Issacs, E., and Mertus, J. (**1982**). "The role of gross spectral shape as a perceptual cue to place of articulation in initial stop consonants," J. Acoust. Soc. Am. **72**, 43–50.
Boll, S. (**1979**). "Suppression of acoustic noise in speech using spectral subtraction," IEEE Trans. Acoust., Speech, Signal Process. **ASSP-27**, 113–120.
Chistovich, L., and Lublinskaja, V. (**1979**). "The center of gravity effect in vowel spectra and critical distance between the formants," Hear. Res. **1**, 185–195.
Cooper, F., Delattre, P., Liberman, A., Borst, J., and Gerstman, L. (**1952**).

"Some experiments on perception of synthetic speech sounds," J. Acoust. Soc. Am. **24**, 597–606.
Delattre, P., Liberman, A., and Cooper, F. (**1955**). "Acoustic loci and transitional cues for consonants," J. Acoust. Soc. Am. **27**, 769–774.
Diehl, R., Lindblom, B., and Creeger, C. (**2003**). "Increasing realism of auditory representations yields further insights into vowel phonetics," *Proc. 5th Intl Congress Phonetic Sciences*.
Dorman, M., and Loizou, P. (**1996**). "Relative spectral change and formant transitions as cues to labial and alveolar place of articulation," J. Acoust. Soc. Am. **100**, 3825–3830.
Dorman, M., Studdert-Kennedy, M., and Ralphael, L. (**1977**). "Stop consonant recognition: Release bursts and formant transitions as functionally equivalent context-dependent cues," Percept. Psychophys. **22**, 109–122.
Dubno, J., and Dorman, M. (**1987**). "Effects of spectral flattening on vowel identification," J. Acoust. Soc. Am. **82**, 1503–1511.
Flanagan, J. (**1955**). "A difference limens for vowel formant frequency," J. Acoust. Soc. Am. **27**, 288–291.
Hawks, J. (**1994**). "Difference limens for formant patterns of vowel sounds," J. Acoust. Soc. Am. **95**, 1074–1084.
Hillenbrand, J., and Gayvert, R. (**1993**). "Identification of steady-state vowels synthesized from the Peterson and Barney measurements," J. Acoust. Soc. Am. **94**, 668–674.
Hillenbrand, J., and Neary, T. (**1999**). "Identification of resynthesized /hVd/ utterances: Effects of formant contour," J. Acoust. Soc. Am. **105**, 3509–3523.
Hillenbrand, J., Getty, L., Clark, M., and Wheeler, K. (**1995**). "Acoustic characteristics of American English vowels," J. Acoust. Soc. Am. **97**, 3099–3111.
Ito, M., Tsuchida, J., and Yano, M. (**2001**). "On the effectiveness of whole spectral shape for vowel perception," J. Acoust. Soc. Am. **110**, 1141–1149.
Kamath, S. (**2001**). "A multi-band spectral subtraction method for speech enhancement," Masters thesis, Dept. of Electrical Engineering, University of Texas—Dallas.
Kamath, S., and Loizou, P. (**2002**). "A multi-band spectral subtraction method for enhancing speech corrupted by colored noise," *Proc. ICASSP*, Orlando, FL.
Kent, R., and Read, C. (**1992**). *The Acoustic Analysis of Speech* (Singular, San Diego, CA), Chap. 6.
Kewley-Port, D. (**1983**). "Time varying features as correlates of place of articulation in stop consonants," J. Acoust. Soc. Am. **73**, 322–335.
Klatt, D. (**1982**). "Prediction of perceived phonetic distance from critical band spectra: A first step," *IEEE ICASSP*, pp. 1278–1281.
Krishnan, A. (**1999**). "Human frequency-following responses to two-tone approximations of steady-state vowels," Audiol. Neuro-Otol. **4**, 95–103.
Krishnan, A. (**2002**). "Human frequency-following responses: representation of steady-state synthetic vowels," Hear. Res. **166**, 192–201.
Lahiri, A., Gerwirth, L., and Blumstein, S. (**1984**). "A reconsideration of acoustic invariance for place of articulation in diffuse stop consonants. Evidence from a cross-language study," J. Acoust. Soc. Am. **76**, 391–404.
Leek, M., Dorman, M., and Summerfield, Q. (**1987**). "Minimum spectral contrast for vowel identification by normal-hearing and hearing-impaired listeners," J. Acoust. Soc. Am. **81**, 148–154.
Lehiste, I., and Meltzer, D. (**1973**). "Vowel and speaker identification in natural and synthetic speech," Lang Speech **16**, 356–364.
Liberman, A., Delattre, P., and Cooper, F. (**1952**). "The role of selected stimulus variables in the perception of unvoiced stop consonants," Am. J. Psychol. **65**, 497–516.
Liu, C., and Kewley-Port, D. (**2001**). "Vowel formant discrimination for high-fidelity speech," J. Acoust. Soc. Am. **116**, 1224–1233.
Loizou, P., and Poroy, O. (**2001**). "Minimum spectral contrast needed for vowel identification by normal hearing and cochlear implant listeners," J. Acoust. Soc. Am. **110**, 1619–1627.
Miller, G. A., and Nicely, P. E. (**1955**). "An analysis of perceptual confusions among some English consonants," J. Acoust. Soc. Am. **27**, 338–352.
Molis, M. (**2005**). "Evaluating models of vowel perception," J. Acoust. Soc. Am. **118**, 1062–1071.
Nebalek, A. (**1988**). "Identification of vowels in quiet, noise and reverberation: Relationships with age and hearing loss," J. Acoust. Soc. Am. **84**, 476–484.
Nebalek, A., and Dagenais, P. (**1986**). "Vowel errors in noise and in reverberation by hearing-impaired listeners," J. Acoust. Soc. Am. **80**, 741–748.
Peterson, G., and Barney, H. (**1952**). "Control methods used in a study of the vowels," J. Acoust. Soc. Am. **24**, 175–184.
Pickett, J. (**1957**). "Perception of vowels heard in noises of various spectra," J. Acoust. Soc. Am. **29**, 613–620.
Plomp, R. (**1970**). "Timbre as a multidimensional attribute of complex tones," in *Frequency Analysis and Periodicity Detection in Hearing*, edited by R. Plomp and G. Smoorenbburg (Sijthoff, Leiden), pp. 397–414.
Remez, R., Rubin, P., Pisoni, D., and Carrell, T. (**1981**). "Speech perception without traditional speech cues," Science **212**, 947–950.
Shannon, R., Jensvold, A., Padilla, M., Robert, M., and Wang, X. (**1999**). "Consonant recording for speech testing," J. Acoust. Soc. Am. **106**, L71–L74.
Smits, R., Bosch, L., and Collier, R. (**1996**). "Evaluation of various sets of acoustic cues for the perception of prevocalic stop consonants: I. Perception experiment," J. Acoust. Soc. Am. **100**, 3852–3864.
Stevens, K., and Blumstein, S. (**1978**). "Invariant cues for the place of articulation in stop consonants," J. Acoust. Soc. Am. **64**, 1358–1368.
Strange, W. (**1989**). "Evolving theories of vowel perception," J. Acoust. Soc. Am. **85**, 2081–2081.
Strange, W. (**1999**). "Perception of vowels: Dynamic constancy," in *The Acoustics of Speech Communication*, edited by J. Pickett, (Allyn and Bacon, Needham Heights) 153–166.
Strange, W., Jenkins, J. J., and Johnson, T. L. (**1983**). "Dynamic specification of coarticulated vowels," J. Acoust. Soc. Am. **74**, 695–705.
Syrdal, A., and Gopal, H. (**1986**). "A perceptual model of vowel recognition based on the auditory representation of American English vowels," J. Acoust. Soc. Am. **79**, 1086–1100.
Walley, A., and Carrell, T. (**1983**). "Onset spectra and formant transitions in the adult's and children's perception of place of articulation in stop consonants," J. Acoust. Soc. Am. **73**, 1011–1022.
Wang, M., and Bilger, R. (**1973**). "Consonant confusions in noise: A study of perceptual features," J. Acoust. Soc. Am. **54**, 1248–1266.
Young, E., and Sachs, M. (**1979**). "Representation of steady-state vowels in the temporal aspects of the discharge patterns of populations of auditory nerve fibers," J. Acoust. Soc. Am. **66**, 1381–1403.
Zahorian, S. A., and Jagharghi, A. J. (**1993**). "Spectral-shape features versus formants as acoustic correlates for vowels," J. Acoust. Soc. Am. **94**, 1966–1982.
Zwicker, E., and Fastl, H. (**1990**). *Psychoacoustics, Facts and Models*, (Springer Verlag, Berlin).