# The contribution of obstruent consonants and acoustic landmarks to speech recognition in noise

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

(Received 14 April 2008; revised 8 September 2008; accepted 10 September 2008)

The obstruent consonants (e.g., stops) are more susceptible to noise than vowels, raising the question whether the degradation of speech intelligibility in noise can be attributed, at least partially, to the loss of information carried by obstruent consonants. Experiment 1 assesses the contribution of obstruent consonants to speech recognition in noise by presenting sentences containing clean obstruent consonants but noise-corrupted voiced sounds (e.g., vowels). Results indicated substantial (threefold) improvement in speech recognition, particularly at low signal-to-noise ratio levels (−5 dB). Experiment 2 assessed the importance of providing partial information, within a frequency region, of the obstruent-consonant spectra while leaving the remaining spectral region unaltered (i.e., noise corrupted). Access to the low-frequency (0–1000 Hz) region of the clean obstruent-consonant spectra was found to be sufficient to realize significant improvements in performance and that was attributed to improvement in transmission of voicing information. The outcomes from the two experiments suggest that much of the improvement in performance must be due to the enhanced access to acoustic landmarks, evident in spectral discontinuities signaling the onsets of obstruent consonants. These landmarks, often blurred in noisy conditions, are critically important for understanding speech in noise for better determination of the syllable structure and word boundaries. © *2008 Acoustical Society of America.* [DOI: 10.1121/1.2997435]

PACS number(s): 43.71.Es, 43.71.Gv [MSS] Pages: 3947–3958

## I. INTRODUCTION

Speech sounds can generally be classified into two broad categories, vowels and consonants. A number of studies assessed the contribution of information provided by vowels versus consonants to speech intelligibility in quiet (Cole *et al.*, 1996; Owren and Cardillo, 2006; Kewley-Port *et al.*, 2007). In noise, however, listeners may employ a different set of acoustic cues, and therefore, the contribution of vowels versus consonants to speech intelligibility might change. This is so because noise masks vowels and consonants differently and to a different extent. For one, the low-energy obstruent consonants (e.g., stops) are masked more easily by noise (e.g., Phatak and Allen, 2007; Parikh and Loizou, 2005) than the high-energy vowels and semivowels. Phatak and Allen (2007), for instance, showed that aside from a small subset of consonants, the vowel-to-consonant recognition ratio is well above unity for a large range of signal-to-noise ratio (SNR) levels (−20 to 0 dB), suggesting that vowels are easier to recognize than consonants in speech-weighted noise. The study by Parikh and Loizou (2005) showed that the information contained in the first two vowel formants is preserved to some degree even at low SNR levels. In contrast, both the spectral tilt and burst frequency of stop consonants, which are known to convey place of articulation information (e.g., Blumstein and Stevens, 1979), were significantly altered by noise. This raises then the question as to whether the degradation of speech intelligibility in noise can be attributed, at least partially, to the loss of information carried by obstruent consonants. Secondly, noise corrupts acoustic landmarks that are produced by (abrupt) spectral discontinuities such as those created by the closing and release of stop consonants. These landmarks were posited to be crucial in lexical-access models (Stevens, 2002). Some argued (Assmann and Summerfield, 2004) that these landmarks serve as robust speech features that listeners exploit to understand speech in noise. Furthermore, evidence from animal neurophysiology studies (Smith, 1979; Delgutte and Kiang, 1984) suggests that the auditory system is particularly responsive (manifesting adaptation effects) to abrupt changes to the signal, such as those occurring at the onsets of landmarks. These adaptation effects are beneficial in speech recognition as they enhance contrast between successive speech segments separated by an abrupt spectral change. In brief, while the acoustic cues to voiced speech segments (e.g., vowels) may resist, to some extent, the presence of noise (e.g., Parikh and Loizou, 2005), the acoustic cues to the unvoiced (low-energy) segments are severely corrupted and perhaps rendered useless.

In the present study, we assess the contribution of information provided by obstruent consonants to speech intelligibility in noise. We do this, in experiment 1, by using noise-corrupted sentences in which we replace the noise-corrupted obstruent consonants with clean obstruent consonants, while leaving the sonorant sounds (vowels, semivowels, and nasals) corrupted. In experiment 2, we control the amount of spectral and temporal information that is present in the clean obstruent sound spectra by presenting to listeners partial spectral information spanning a specific frequency range ($0-F_C$ Hz), while leaving the remaining spectrum ($\geq F_C$ Hz) corrupted. By restricting, for instance, the $F_C$ value to low

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

frequencies (e.g., $F_C \leq 1000$ Hz), we present to listeners only voicing and F1 information, and subsequently only information about low-frequency acoustic landmarks (Stevens, 2002). Experiment 2 thus assesses the contribution of low- and high-frequency acoustic landmarks to speech recognition in noise.

We choose to study the obstruent consonants (Shriberg and Kent, 2003), which includes stops, fricatives, and affricates, for several reasons. First, the obstruent consonants are more susceptible to noise masking and reverberation compared to the more intense vowels and semivowels (e.g., Phatak and Allen, 2007). Secondly, the spectra of the majority of the obstruent consonants have high-frequency prominence that may or may not be accessible or audible to people with sensorineural hearing loss (HL) (e.g., Ching *et al.*, 1998; Hogan and Turner, 1998). Hence, the present experiments will assess the maximum gain in intelligibility that could be obtained if listeners had access to (clean) obstruent-consonant information. If the gain in intelligibility is substantial, that would suggest that special algorithms ought to be designed for hearing aid applications to enhance access or detection of the obstruent consonants. Furthermore, the restoration of acoustic landmarks might provide some insight regarding the difficulty hearing-impaired (HI) listeners experience when communicating in noise.

## II. EXPERIMENT 1: CONTRIBUTION OF OBSTRUENT CONSONANTS TO SPEECH INTELLIGIBILITY IN NOISE

### A. Methods

#### 1. Subjects

Thirteen normal-hearing listeners participated in this experiment. All subjects were native speakers of American English and were paid for their participation. The subjects' age ranged from 18 to 40 years, with the majority being undergraduate students from the University of Texas at Dallas.

#### 2. Stimuli

The speech material consisted of isolated /VCV/ syllables taken from Shannon *et al.* (1999) and sentences taken from the IEEE database (IEEE, 1969). Since this study was concerned with the contribution of obstruent consonants, only the following consonants were used from the recordings of Shannon *et al.* (1999): /p, t, k, b, d, g, f, ð, s, ʃ, v, z, tʃ, dʒ/. The semivowels and nasals were excluded. The context vowel in the VCV syllables was /ɑ/. The selected consonants were produced by three male and three female speakers. The talkers had no noticeable regional accent (standard American Midwest dialect). We chose to use both sentences and VCV syllables as testing material in order to assess the contribution of obstruent consonants in situations where listeners have access to contextual information (i.e., sentence recognition task) and compare that outcome with the situation where listeners do not use high-level linguistic knowledge (e.g., VCV task).

All IEEE sentences were produced by one male and one female speaker. The male speaker was born in Texas and the female speaker was born in North Carolina. Neither talker had a noticeable Southern regional accent. The sentences were recorded in a sound-proof booth (Acoustic Systems) in our laboratory at a 25 kHz sampling rate. Details about the recording setup and copies of the recordings are available in Loizou, 2007. The speech materials were corrupted by a 20-talker babble (Auditec CD, St. Louis, MO) at −5 and 0 dB SNRs. The average long-term spectrum of the multitalker babble is shown in Parikh and Loizou (2005). These SNR levels were chosen to avoid floor effects. The babble interferer started at 100 ms before the beginning of each speech token and stopped at least 100 ms after the end of the sentence (and the VCV syllable). This masker has temporal characteristics similar to those found in continuous speech-shaped noise maskers in that it does not have any temporal envelope dips that would allow the listener to "glimpse" the target. Hence, for the most part, the masker used in the present study can be considered to be continuous.

The IEEE sentences were manually segmented into two broad phonetic classes: (a) the obstruent sounds, which included the stops, fricatives, and affricates, and (b) the sonorant sounds, which included the vowels, semivowels, and nasals. The segmentation was done in two steps. In the first step, a highly accurate F0 detector, taken from the STRAIGHT algorithm (Kawahara *et al.*, 1999), was used to provide the initial classification of voiced and unvoiced speech segments. The stop closures were classified as belonging to the unvoiced segments. The F0 detection algorithm was applied every 1 ms to the stimuli using a high-resolution fast Fourier transform (FFT) to provide for accurate temporal resolution of voiced/unvoiced boundaries. Segments with nonzero F0 values were initially classified as voiced and segments with zero F0 value (as determined by the STRAIGHT algorithm) were classified as unvoiced. In the second step, the voiced and unvoiced decisions were inspected for errors and the detected errors were manually corrected. Voiced stops produced with prevoicing (e.g., /b/) as well as voiced fricatives (e.g. /z/) were classified as obstruent consonants. Waveform and time-aligned spectrograms were used to refine the voiced/unvoiced boundaries. The criteria for identifying a segment belonging to voiced sounds (sonorant sounds) included the presence of voicing, a clear formant pattern, and absence of signs of a vocal-tract constriction. For the special boundary that separates a prevocalic stop from a following semivowel (as in *truck*), we adopted the rule used in the phonetic segmentation of the Texas Instruments-Massachusetts Institute of Technology (TIMIT) corpus (Seneff and Zue, 1988). More precisely, the unvoiced portion of the following semivowel or vowel was absorbed in the stop release, and was thus classified as an obstruent consonant. The two-class segmentation of all IEEE sentences was saved in text files (same format as the TIMIT .phn files) and is available from a CD ROM in Loizou, 2007. The VCV syllables were similarly segmented to consonants and vowels (/ɑ/). An example segmentation of the syllable /ɑ t ɑ/ is shown in Fig. 1.

To examine whether there exists a balanced distribution and coverage of obstruent consonants across the various IEEE lists, we computed the relative duration of the obstruent consonants in the IEEE corpus. The relative duration

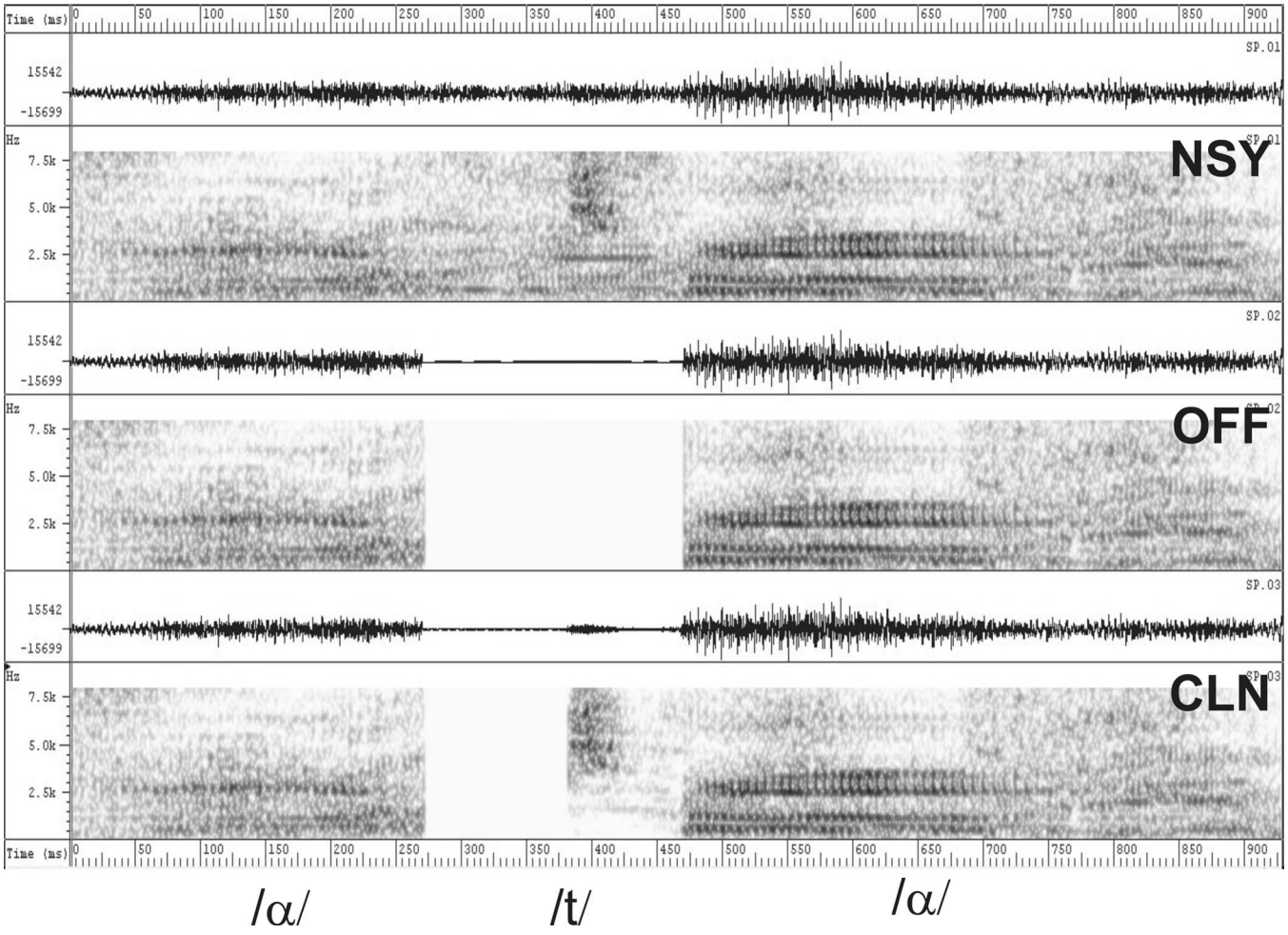


FIG. 1. Time waveforms and spectrograms of the syllable /ɑ t ɑ/, which was corrupted by the multitalker babble at 0 dB SNR (top panel, NSY condition), processed in the OFF condition (middle panel) and processed in the CLN condition (bottom panel).

was expressed in proportion of the total duration of each sentence. The mean total duration of the obstruent consonants in the IEEE sentences was found to be 36.2% (s.d. =3.1%) of the total duration of each sentence, consistent with that reported by Mines *et al.* (1978). This analysis indicates that there is a very small variability in the total duration of obstruent consonants across the IEEE sentence lists. The study by Mines *et al.* (1978) analyzed the distribution of phonemes taken from a large database (comprising nearly 104 000 phonemes) and found that the phonemes belonging in the obstruent class occur 34% of the time.

### 3. Conditions

The 13 subjects were quasirandomly assigned to the conditions involving processed IEEE sentences produced by the male and female speakers (seven listened to the IEEE sentences produced by a male speaker and six listened to the IEEE sentences produced by a female speaker). Similarly, the subjects were quasirandomly assigned to the processed VCV syllables produced by three male and three female speakers (seven listened to VCV syllables produced by two speakers, and six listened to VCV syllables produced by six speakers, with equal distribution between male and female speakers). Subjects were tested in three conditions. In the first control condition, the noisy speech stimuli (sentences and VCV syllables) were left unaltered. In this condition, the obstruent consonants remained corrupted by the multitalker babble, and will refer to it as the NSY (noisy) condition. In the second condition, the obstruent consonants were removed from the noise-corrupted sentences, and silence was inserted in their place. We will refer to this condition as the OFF condition. In the third condition, the noisy obstruent-consonant segments were replaced with the corresponding clean obstruent-consonant segments. The modified sentences contained noise-corrupted (at −5 and 0 dB SNRs) sonorant sounds, but clean obstruent consonants. We will refer to this condition as the CLN (clean) condition. Figure 1 shows example time waveforms and spectrograms for the syllable /ɑ t ɑ/ (produced by a male speaker) processed in the three conditions.

### 4. Procedure

The experiments were performed in a sound-proof room (Acoustic Systems, Inc.) using a PC connected to a Tucker-Davis system 3. Stimuli were played to the listeners monaurally through Sennheiser HD 250 Linear II circumaural headphones at a comfortable listening level. Prior to the sentence test, each subject listened to a set of noise-corrupted sentences to familiarize them with the testing procedure. During the test, subjects were asked to write down the words they

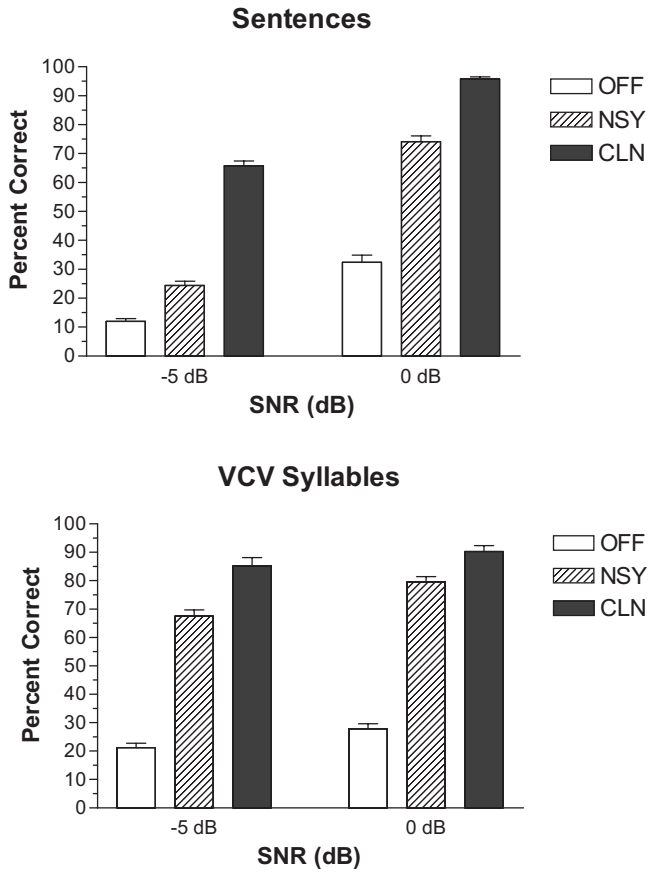


FIG. 2. Mean percent correct scores obtained in the various conditions for the sentence recognition task (top panel) and VCV recognition task (bottom panel). The error bars indicate standard errors of the mean.

heard. In the consonant test, subjects were presented with a total of 12 repetitions of each consonant. The stimuli were presented in blocks of six repetitions each. The consonant stimuli were completely randomized. All test sessions were preceded by one practice session in which the identity of the consonants was indicated to the listeners. The whole listening test lasted for about 4–5 h, which was split into two sessions lasting 2–2.5 h each. Five minute breaks were given to the subjects every 30 min. To collect responses, a graphical interface was used that allowed the subjects to identify the consonants they heard by clicking on the corresponding button in the graphical interface. Subjects participated in a total of 12 conditions (=2 SNR levels × 3 processing conditions × 2 types of speech material). Two lists of sentences (i.e., 20 sentences) were used per condition, and none of the lists were repeated across conditions. The sentence lists were counterbalanced across subjects. Sentences were presented to the listeners in blocks, and 20 sentences were presented in each block for each condition. The order of the test conditions was randomized across subjects.

### B. Results and discussion

The mean scores for all conditions are shown in Fig. 2. For the sentence recognition task, performance was measured in terms of percent of words identified correctly (all words were scored). Two-way ANOVA (with repeated measures) indicated a significant effect of SNR level ($F[1,12]=1330.1$), $p<0.0005$), a significant effect of stimulus processing ($F[2,24]=1189.2$, $p<0.0005$) and a significant interaction ($F[2,24]=67.1$, $p<0.0005$). The interaction was caused by the fact that the improvement in performance obtained when the listeners had access to the clean obstruent consonants was larger at −5 dB SNR than 0 dB.

Similar ANOVA analysis was applied to the scores obtained on the isolated consonant recognition task. Two-way analysis of variance (ANOVA) (with repeated measures) indicated a significant effect of SNR level ($F[1,12]=72.1$, $p<0.0005$), a significant effect of stimulus processing ($F[2,24]=538.1$, $p<0.0005$), and a significant interaction ($F[2,24]=7.9$, $p=0.002$).

*Post hoc* tests (Scheffe) applied to the sentence scores at −5 dB SNR revealed highly significant differences ($p<0.0005$) between the scores obtained in the OFF and NSY (control) conditions, and between the NSY and CLN conditions. This was also found to be true with the sentence scores at 0 dB SNR. *Post hoc* tests (Scheffe) applied to the VCV scores revealed significant differences ($p<0.005$) between the OFF and NSY conditions and between the NSY and CLN conditions at both SNR levels.

It is clear from Fig. 2 that listeners benefited a great deal from having access to the clean obstruent consonants, despite the fact that the sonorant sounds (vowels, semivowels, and nasals) were corrupted by the multitalker babble. This benefit was found to be most prominent in the low SNR conditions (−5 dB SNR). Sentence recognition improved nearly by a factor of 3 when the clean obstruent consonants were introduced amid the noise-corrupted sonorant sounds in sentences. In the isolated VCV task, performance improved by 20% points at −5 dB SNR.

Subject performance dropped significantly when the obstruent consonants were removed, thus providing additional evidence about the contribution and importance of the information provided by obstruent consonants to speech intelligibility in noise. Note that we cannot directly compare our findings with those of Cole *et al.* (1996) and Kewley-Port *et al.* (2007) for the following reasons. First, silence was inserted in the present study (in the OFF condition) rather than white noise or tones. In doing so, we ruled out any phonemic restoration effects that could have contributed (positively) to performance (Bashford and Warren, 1987). Second, our study used sentences embedded in noisy backgrounds and thus the task was more difficult. Furthermore, noise does not mask vowels and consonants the same way or equally, thereby making it difficult to truly assess the contribution of vowels versus consonants in noisy environments.

It is interesting to note that performance on the VCV task was significantly ($p<0.005$, based on t-tests) above chance (7.1% correct) despite the fact that the consonants were removed from the VCV syllables. This can be attributed to the subjects making use of the vowel formant transitions into and out of the medial consonant. As demonstrated by Cole *et al.* (1996) and Owren and Cardillo (2006), the formant transitions at the edges of vowels contain information about the neighboring consonants that may enable subjects to identify words in sentences.

Performance was expected in this experiment to improve

as new acoustic information was added, but the magnitude of the improvement was difficult to predict, particularly for the sentence recognition task. Sentence recognition scores improved significantly when the clean obstruent consonants were introduced, and the improvement was largest at −5 dB SNR. In the −5 dB SNR condition, for instance, performance improved from 25% correct to 65% correct. The underlying mechanisms responsible for such large improvement are not clear from the present Experiment. Listeners had access to multiple spectral/temporal cues when the clean obstruent consonants were introduced. These included F1/F2 transitions to/from the vowel and sonorant sounds, accurate voicing information, and consequently better access to acoustic landmarks that perhaps aided the listeners in identifying word boundaries in the noisy speech stream. In order to delineate and identify the cues and/or factors that contributed to the large improvements in performance, we present to listeners in experiment 2 partially clean spectral information of the obstruent consonants.

## III. EXPERIMENT 2: CONTRIBUTION OF PARTIAL SPECTRAL INFORMATION OF OBSTRUENT CONSONANTS TO SPEECH INTELLIGIBILITY

The previous experiment demonstrated the importance and contribution of obstruent consonants to speech intelligibility in noisy (continuous) backgrounds. The listeners in experiment 1 had access to the full spectrum of the obstruent consonants and thus it was not clear which temporal/spectral characteristic or spectral feature of the consonants contributed the most or generally which temporal/spectral cues did the listeners use. In the present experiment, we introduce systematically access to the clean spectrum in the region of $0-F_C$ Hz ($F_C$=500–3000 Hz), while leaving the remaining spectrum ($>F_C$ Hz) of the obstruent consonants unaltered (i.e., noise corrupted). The issue is determining the region(s) of the clean obstruent-consonant spectrum that contributes the most to intelligibility. By setting $F_C$, for instance, to a small frequency value (e.g., $F_C$=500 Hz), we provide to listeners reliable F1 and voicing information. This information in turn provides access to reliable low-frequency acoustic landmarks that might assist listeners better identify syllable or word boundaries (Stevens, 2002). Hence, the use of low-frequency values for $F_C$ can help us assess the contribution of low-frequency acoustic landmarks to speech intelligibility in noise. Similarly, by setting $F_C$ to higher values (e.g., $F_C$ >2000 Hz), we introduce more spectral cues (e.g., F2 formant transitions) including high-frequency landmarks (e.g., onset of frication noise) that can potentially aid the listeners identify more words in the noisy speech stream. In brief, by controlling the value of $F_C$, we can investigate in the present experiment the contribution of information carried by low- and high-frequency acoustic landmarks to speech recognition in noise.

### A. Methods

#### 1. Subjects and material

The same 13 subjects who participated in experiment 1 also participated in the present experiment. The same speech materials (IEEE sentences and VCV syllables) were used as in experiment 1.

#### 2. Signal Processing

The noisy obstruent-consonant stimuli (at −5 and 0 dB SNRs) were processed so as to create conditions in which listeners had access to the clean obstruent-consonant spectrum in the region of $0-F_C$ Hz, while leaving the remaining spectral region ($>F_C$ Hz) unaltered (noise corrupted). The obstruent consonants were processed as follows. An FFT was applied to the clean and corrupted obstruent-consonant segments every 10 ms with no overlap between adjacent segments (rectangular windowing was used[1]). A new (complex) spectral vector was formed consisting of the clean FFT spectral components for frequencies spanning the range $0-F_C$ Hz, and the noise-corrupted FFT spectral components for frequencies higher than $F_C$ Hz. The inverse FFT was applied to this new spectral vector to reconstruct the time-domain signal. This series of operations was repeated until the end of the obstruent-consonant segment was reached. The remaining sonorant segments (vowels, semivowels, and nasals) in the noise-corrupted sentence stimuli were left unaltered.

Four different values of the upper-cutoff frequency $F_C$ were considered: 500, 1000, 2000, and 3000 Hz. These values were chosen to encompass the F1 and F2 regions. Note that in experiment 1 the value of $F_C$ was set to half the sampling frequency (Nyquist frequency), i.e., $F_C$=12.5 kHz, for the CLN condition. Figure 3 (bottom panel) shows an example spectrogram of a sentence processed using $F_C$ =1000 Hz at −5 dB SNR.

#### 3. Procedure

The procedure was identical to that used in experiment 1. Subjects were tested in a total of 16 (=4 values of $F_C$ ×2 SNR levels×2 types of material) conditions. Two lists of sentences (i.e., 20 sentences) were used per condition, and none of the lists from experiment 1 was used or repeated across conditions. The sentence lists were counterbalanced across subjects. Sentences were presented to the listeners in blocks, with 20 sentences presented in each block for each condition. The order of the test conditions was randomized across subjects.

### B. Results and discussion

The mean scores for all conditions are shown in Fig. 4. The baseline NSY score ($F_C$=0 Hz) and the score obtained with clean obstruent consonants ($F_C$=12.5 kHz) are also plotted for comparison. Two-way ANOVA (with repeated measures) applied to the sentence scores indicated a significant effect of SNR level ($F[1,12]=1094.9$, $p<0.0005$), a significant effect of frequency $F_C$ ($F[5,60]=80.1$, $p<0.0005$), and a significant interaction ($F[5,60]=13.5$, $p$

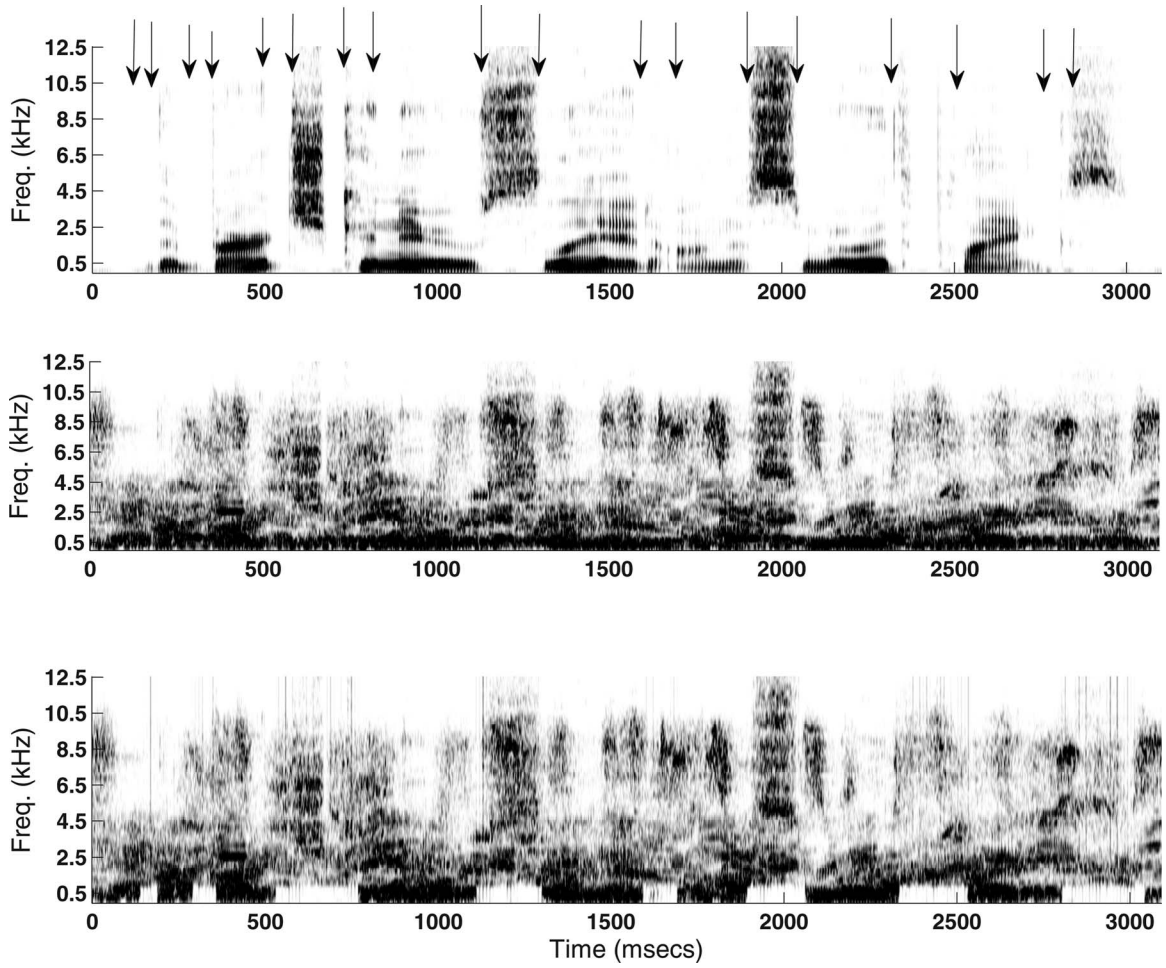


FIG. 3. Wide-band spectrogram (bottom panel) of the sentence: "The birch canoe slid on the smooth planks" processed using $F_C$=1000 Hz. The top and middle panels show the corresponding spectrograms of the same sentence in quiet and in multitalker babble (−5 dB SNR) conditions, respectively. The arrows in the top panel indicate the acoustic landmarks present in the clean stimulus.

$<0.0005$). The control NSY condition ($F_C$=0 Hz) was also included in the ANOVA. Two-way ANOVA (with repeated measures) applied to the VCV scores indicated a significant effect of SNR level ($F[1,12]=72.8$, $p<0.0005$), a significant effect of frequency $F_C$ ($F[5,60]=33.4$, $p<0.0005$), and a significant interaction ($F[5,60]=5.8$, $p<0.0005$). The performance of the two groups of listeners on the VCV identification task was comparable and the difference in scores for the processed VCV stimuli produced by two versus six speakers was not significant.[2] The interaction (in both sentence and VCV analysis) was caused by the fact that the improvement in performance noted with increasing values of $F_C$ was larger at −5 dB SNR than at 0 dB SNR.

*Post hoc* tests (Scheffe) were run to determine the smallest value of $F_C$ at which significant improvements in performance are observed relative to the control condition (NSY). The sentence scores obtained with $F_C$=1000 Hz at 0 dB SNR were significantly ($p$=0.005) higher than the control (baseline) scores ($F_C$=0 Hz). The sentence scores obtained with $F_C$=500 Hz at −5 dB SNR were significantly ($p$=0.008) higher than the control scores. Similar *post hoc* tests were run on the VCV scores. The scores obtained with $F_C$=3000 Hz at −5 dB SNR were significantly ($p$=0.026) higher than the control scores. The scores obtained in the CLN condition ($F_C$=12.5 kHz) at 0 dB SNR were significantly ($p$=0.013) higher than the control scores.

The sentence and VCV data clearly demonstrated that it is not necessary to have access to the full spectrum of the (clean) obstruent consonants to observe significant improvements in intelligibility. Access to the low-frequency (0–1000 Hz) region of the clean obstruent-consonant spectra was sufficient to realize significant improvements in performance. Sentence intelligibility scores (at −5 dB SNR) improved significantly by 2% points, when listeners had access to the low-frequency region (0–1000 Hz) of the obstruent consonants. Similar, albeit smaller, improvements were also noted with the VCV syllables. Considerably larger improvements in intelligibility were noted when listeners had access to the 0–3000 Hz region of the obstruent spectrum. Sentence scores improved from 24% to 51% correct, while VCV scores improved from 67% to 80% correct at −5 dB SNR.

The outcome that having access to the low-frequency (0–1 kHz) region of the spectrum yielded significant gains in intelligibility led us to wonder whether this was due to better transmission of voicing information. To further investigate this, we subjected the consonant confusion matrices to information transmission analysis (Miller and Nicely, 1955). The results were analyzed in terms of voicing, place, and

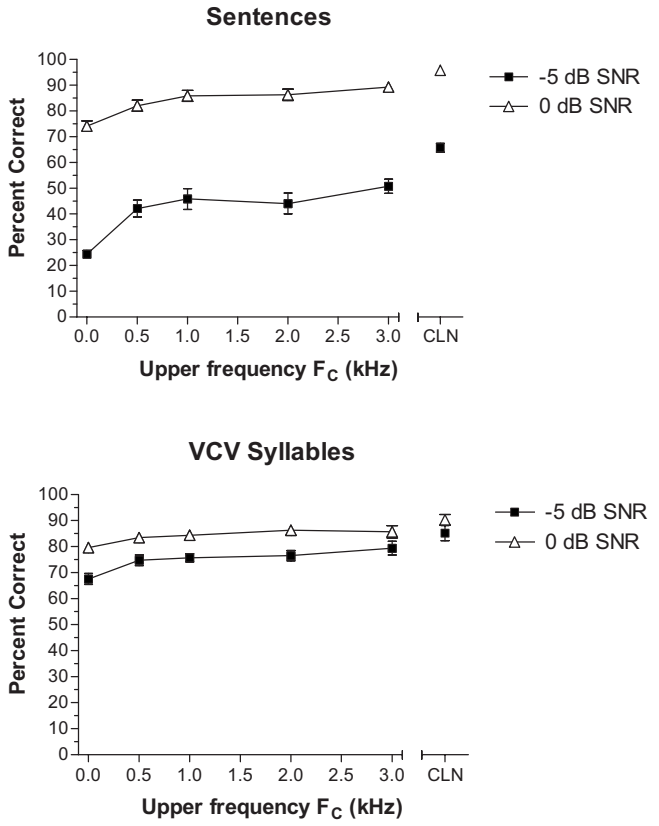


FIG. 4. Mean percent correct scores as a function of the upper-cutoff frequency $F_C$(Hz) and SNR level. The spectrum of the obstruent consonants was clean in the frequency region of $0-F_C$ Hz and was left unaltered (i.e., noise corrupted) in the region thereafter (i.e., $>F_C$ Hz). In the CLN condition, the listeners had access to the whole spectrum of the clean obstruent consonants (i.e., $F_C$=12.5 kHz). The $F_C$=0 Hz condition corresponds to the control (NSY) condition in which the whole spectrum of the obstruent consonants was left noise corrupted. The error bars indicate standard errors of the mean.

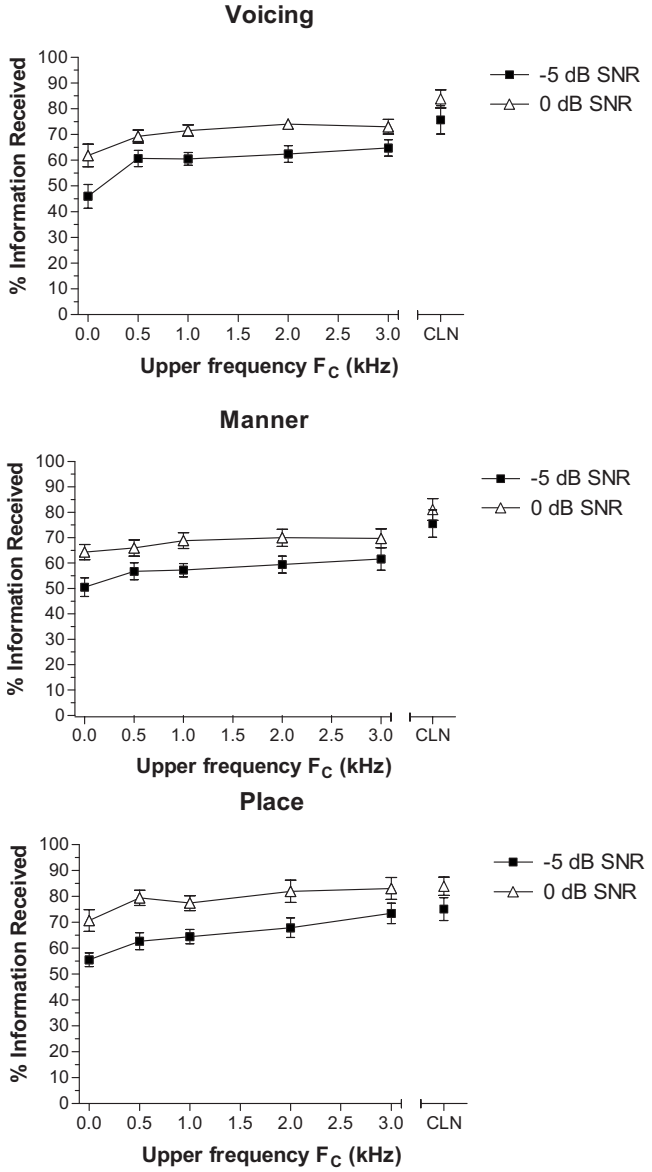


FIG. 5. Information transmission analysis (plotted as a function of the upper-cutoff frequency $F_C$) for the articulatory features of voicing, manner, and place of articulation. The error bars indicate standard errors of the mean.

manner features and are plotted in Fig. 5. Two-way ANOVA (with repeated measures) applied to the voicing scores indicated a significant effect of SNR level ($F[1,12]=62.4$, $p<0.0005$), a significant effect of frequency $F_C$ ($F[5,60]=11.8$, $p<0.0005$), and a nonsignificant interaction ($F[5,60]=1.1$, $p=0.36$). Similar ANOVA applied to the manner and place feature scores revealed significant effects for SNR level and frequency $F_C$ and nonsignificant interactions.

As shown in Fig. 5 (top panel), large improvements were noted in the transmission of voicing information when $F_C \geqslant 500$ Hz. More precisely, transmission of voicing information improved from 45% to 60% at −5 dB SNR when listeners had access to the low-frequency region (0–500 Hz) of the obstruent-consonant spectrum. *Post hoc* tests (Fisher's LSD) confirmed that the improvement in voicing relative to the baseline (control) condition was significant ($p<0.05$) for all values of $F_C \geqslant 500$ Hz at −5 dB SNR level, and for all values of $F_C \geqslant 1000$ Hz at 0 dB SNR.

Improvements to manner and place feature transmission were also noted relative to the baseline scores, but higher values of $F_C$ were required. Significant ($p<0.05$) improvements to transmission of manner feature were noted for $F_C \geqslant 3000$ Hz at −5 dB SNR and for $F_C$=12.5 kHz (CLN) at 0 dB SNR. Significant ($p<0.05$) improvements to transmission of place feature information were noted for $F_C \geqslant 2000$ Hz for both SNR levels. Place information at −5 dB SNR improved substantially from 55% (baseline score) to approximately 75% when $F_C$=3000 Hz. This was not surprising, given that the 0–3000 Hz region encompasses the F1/F2 region, known to carry place of articulation information (e.g., Liberman *et al.*, 1952).

In summary, the above feature analysis (Fig. 5) indicates that much of the improvement in performance is due to better transmission of voicing and place feature information. The improvement in the transmission of manner of articulation information was found to be relatively smaller. No improvement was noted at 0 dB SNR in the transmission of manner information for any value of $F_C$ except that used in the CLN condition. The improvement from better transmission of voicing information was quite substantial amounting to roughly 50% of the overall improvement on sentence recognition (Fig. 4). To further explore the contribution of voic-

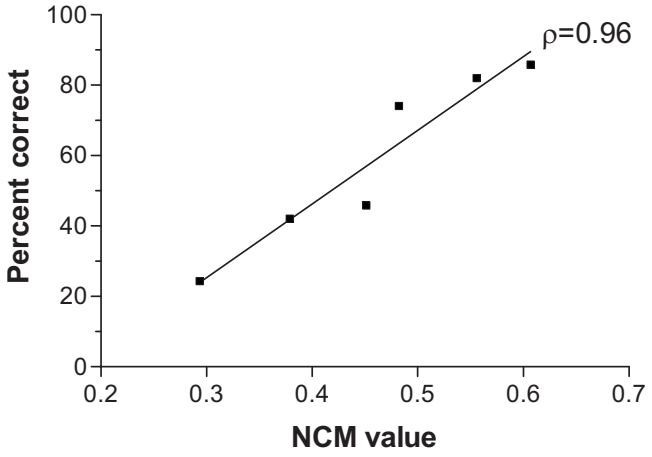


FIG. 6. Scatter plot of the sentence recognition scores obtained for sentences processed with $F_C \leq 1000$ Hz (at −5 and 0 dB SNRs) against the predicted scores obtained using the NCM measure—see text in the Appendix for more details.

ing cues to speech intelligibility in noise, we ran the normalized covariance metric (NCM) (a measure similar to the speech-based speech-transmission index (STI)—see the Appendix) to all the sentence stimuli processed with $F_C \leq 1000$ Hz. This measure assessed the contribution of low-frequency (<1 kHz) envelopes to speech intelligibility and was tested using a total of 120 sentences processed with $F_C \leq 1000$ Hz at −5 and 0 dB SNR levels. The NCM values obtained for each condition (based on 20 sentences/condition) were averaged and correlated with the corresponding speech recognition scores (the control stimuli were also included in the correlation analysis). The resulting correlation coefficient (see Fig. 6) was quite high ($\rho=0.96$) demonstrating the importance of voicing information in speech recognition in noise.

The data obtained in this experiment (Figs. 4 and 5) with $F_C \leq 1000$ Hz are particularly interesting. As mentioned above, and as shown in Fig. 5, listeners had better access to F1 and voicing information when $F_C \leq 1000$ Hz (some F2 information might have been present also in transitions to/from back vowels). Better and more reliable access to voicing leads in turn to better access to low-frequency acoustic landmarks. Figure 3 shows an example spectrogram of a sentence (middle panel) embedded in −5 dB SNR. It is clear from this spectrogram that most of the acoustic landmarks present in the clean signal (shown in the top panel and indicated with arrows) are missing due to the masking of the signal by noise (see middle panel in Fig. 3). The low-frequency acoustic landmarks, in particular, such as those signaling the presence of fricative noise or stop closures, are absent. Some high-frequency acoustic landmarks are present in the corrupted signal (e.g., see fricative segment at $t=1.1$ s in the middle panel of Fig. 3) but these landmarks are not always reliable. In some cases, extraneous and perhaps distracting landmark information (e.g., see $t=1.8$ s and $t=2.1$ s) might be introduced from the masking noise. Hence, it is clear that when listeners were presented with stimuli processed using $F_C \leq 1000$ Hz (see bottom panel in Fig. 3), they had a clear idea on the location of the low-frequency acoustic landmarks and in some instances, but not reliably so, access to high-frequency landmarks. From the findings of experiment 2 and, in particular, from the data obtained with $F_C \leq 1000$ Hz, we may deduce that speech recognition in continuous noise backgrounds is extremely difficult due to the absence of reliable acoustic landmarks.

## IV. GENERAL DISCUSSION

The outcome from experiment 1 demonstrated that the information carried by obstruent consonants contributes significantly to speech recognition in noisy backgrounds. Listeners were able to integrate successfully the noise-corrupted information present in the sonorant sounds with the information present in the clean obstruent consonants (e.g., stops). This integration enabled them to identify more words in the (corrupted) speech stream. Furthermore, experiment 2 demonstrated that listeners did not need access to the full spectrum of the obstruent consonants to obtain significant improvements in intelligibility relative to the unprocessed (noisy) stimuli. The findings from the two experiments raised the following question: What is the underlying mechanism(s) responsible for the large improvements (by a factor of 3 in some cases) in performance observed when listeners had full or partial access to the information provided by the obstruent consonants? Listeners had understandably access to multiple temporal and/or spectral cues (e.g., F1/F2 formant transitions) when presneted with clean obstruent sounds (full spectrum) and corrupted sonorant sounds (experiment 1). The number of cues, however, was limited when listeners were presented with stimuli processed with $F_C \leq 1000$ Hz (experiment 2). Listeners had good access to voicing and low-frequency acoustic landmark information, along with partial, and sometimes unreliable high-frequency landmark information. Yet, the improvement in performance was significant, particularly on the sentence recognition task. What contributed to this improvement? We contend that it is the enhanced access to acoustic landmarks, present in the low and high frequencies, that enabled listeners to identify more words in the noisy speech stream. The acoustic landmarks provided by the obstruent consonants aided the listeners in identifying word (or syllable) boundaries, which are often blurred or smeared in extremely noisy conditions (−5 dB SNR in our study). Next, we discuss in more depth the importance of acoustic landmarks in the context of speech perception models and in the context of speech perception in fluctuating maskers by normal-hearing and HI listeners.

### A. Importance of acoustic landmarks in noise: Implications for lexical-access models

Many speech recognition models (e.g., Stevens, 2002; Pisoni and Sawusch, 1975; Cutler and Norris, 1988; Klatt, 1979) assume that speech is first segmented at reliable points ("islands of reliability") of the acoustic signal followed by a classification of the segmented units into a sequence of phonetic units (e.g., syllables and words). The identified phonetic units are then matched against the items in the lexicon to select the best word sequence intended by the speaker. Evidence of the segmentation stage in the speech recognition process was provided in the study by Cutler and Norris

(1988). Listeners detected words embedded in nonsense bisyllables more slowly when the bisyllables had two strong syllables than when the bisyllables had a strong and a weak syllable. Cutler and Norris (1988) interpreted their data to suggest that strong syllables trigger the segmentation process. Aside from the strong-syllable segmentation model (Cutler and Norris, 1988), others proposed that segmentation is done at word onsets (e.g., Gow *et al.*, 1996) or distinct acoustic landmarks partitioning vowels and consonants (Stevens, 2002).

The data from our study lend support to the segmentation and lexical-access models proposed by Stevens (2002). The prelexical stage of this model consists of three steps. In the first step, the signal is segmented into acoustic landmarks based on detection of peaks and spectral discontinuities in the signal. These landmarks define the boundaries of the vowels, consonants, and glide segments. The second step involves extraction of acoustic cues from the vicinity of the landmarks signifying which articulators are active when the vowel, glide, or consonant landmarks are created and how these articulators are shaped or positioned. This step produces a set of articulator-based features (e.g., [high] tongue body and [anterior] tongue blade), which includes the place of articulation and voicing features. The third step consolidates, taking context into account, all the cues collected in step 2 to derive a sequence of features for each of the landmarks detected in step 1. In this speech perception model, each word in the lexicon is represented by a set of features that specify the manner of articulation, the position and shape of the articulators, as well as information about the syllable structure, such as the position of the consonant in the syllable (e.g., onset and rime) and information about the vocalic nuclei in the syllable in as far as being stressed or reduced.

It is clear that the first step in Steven's (2002) model (i.e., the segmentation into acoustic landmarks) is crucial to the lexical-access model. If the acoustic landmarks are not detected accurately by the listeners or if the landmarks are perceptually not clear or distinct owing to corruption of the signal by external noise, this will affect the subsequent stages of the model in the following ways. First, errors will be made in step 2 in terms of identifying the shape and position of the articulators used in each landmark. Indeed, errors were made in our study in the transmission of voicing and place of articulation feature information (see Fig. 5) when the listeners did not have a clear idea about the location of the acoustic landmarks in the noisy speech stream. Restoring the acoustic landmarks helped improve performance significantly (see Fig. 5). In particular, when the low-frequency landmarks were restored (see experiment 2 with $F_C \leqslant 1000$ Hz), the transmission of voicing information improved significantly (see Fig. 5). The word recognition scores also improved significantly by 21% points at −5 dB SNR when listeners had better access to voicing cues. Second, the absence of reliable landmarks can disrupt the syllable structure, which is known to be important for determining word boundaries in fluent speech. This is so because the onset of a word is always the onset of a syllable. Therefore, not knowing when the syllable starts (i.e., the syllable onset) makes word boundary determination very difficult. Some posit that word onsets are critically important in the speech recognition process as they trigger segmentation and lexical access (Gow *et al.*, 1996; Owren and Cardillo, 2006). In quiet, the word-initial (also syllable-initial) consonants contain several salient features that make them easier to identify (even in the absence of contextual information) than other phonemes that occur later in words (e.g., Owren and Cardillo, 2006). For instance, the voice-onset timing cue is more reliable for signifying voicing in syllable-initial stops than in syllable-final stops, which may not be released. Also, many word-initial consonants benefit from the fact that they are stressed, and stressed syllables occur quite more often (by a factor of 3:1) in the English language compared to weak syllables (Cutler and Carter, 1987). External noise can degrade the salient cues present in syllable-initial consonants. These cues are present in the vicinity of the acoustic landmarks, hence identifying or somehow enhancing access to these landmarks ought to aid in identifying word boundaries and consequently improving word recognition. To further corroborate this, we analyzed the phoneme errors made on the sentence recognition task and examined whether these errors occurred in word-initial position or in other (e.g., medial) positions. We thus compared the errors made in the unprocessed (control) stimuli with those made in the $F_C = 1000$ Hz condition. The error analysis revealed that the word-initial phoneme errors were reduced by 32% when listeners had access to the low-frequency ($F_C \leqslant 1000$ Hz) acoustic landmarks. In contrast, there was only a small ($<10\%$) reduction in the phoneme errors made in the noninitial position. This phoneme error analysis confirmed that access to low-frequency acoustic landmarks facilitates segmentation and aids listeners in determining word boundaries in noisy conditions.

The present study focused on the contribution of acoustic landmarks introduced by obstruent sounds on speech recognition in noise. The vowels and glides also introduce landmarks in the signal (Stevens, 2002), but were not studied in this paper. We make no argument that these landmarks do not play a significant role or are not as important for speech recognition. As mentioned earlier, we restricted our attention to the landmarks introduced by obstruent sounds, for the main reason that these sounds are more susceptible to noise than the sonorant sounds. Nevertheless, the contribution of the glide and vowel landmarks on speech recognition in noise deserves further study.

### B. Speech perception in noise: Normal-hearing and hearing-impaired listeners

The masker (20-talker babble) used in the present study is considered to be continuous, in that it lacks temporal envelope dips. A continuous type of masker was used in this study as it was expected that the contribution of acoustic landmarks to speech recognition would be most prominent. It is generally accepted that listeners are able to recognize speech in modulated or fluctuating maskers with higher accuracy than in continuous (steady-state) noise (e.g., Festen and Plomp, 1990). Several factors contribute to the masking release (see review in Assmann and Summerfield, 2004) in-

cluding segregation of the target on the basis of F0 differences (between the target and masker) and the ability to glimpse the target during the portions of the mixture in which the SNR is favorable, i.e., during periods in which the temporal envelope of the masker reaches a low. The data from the present study suggest that the contributing factor for the masking release may not be the favorable SNR *per se*, but rather the favorable timing of the masker dips, i.e., occurring during periods in which the target signal envelope is low (e.g., during stop closures or generally during obstruent consonants). We thus contend that during the favorable (timewise) periods that the masker envelope reaches a dip, listeners have better access to acoustic landmarks. In contrast, the steady-state maskers do not provide such favorable timing of masker dips as they are continuously "on." Hence, we can deduce that having better access to acoustic landmarks is yet another contributing factor to the release of masking observed with fluctuating maskers.

As shown in Fig. 3 (middle panel), some of the high-frequency acoustic landmarks, such as those signaling the onset of frication noise, are relatively more visible or accessible compared to the low-frequency landmarks. The cues provided by these high-frequency landmarks are particularly important, for instance, for the accurate perception of /s/ (Stelmachowicz *et al.*, 2001), especially for speech tokens produced by female talkers. While normal-hearing listeners may be able to perceive /s/ or other high-frequency consonants, it is questionable whether HI listeners will be able to do so. A number of studies (e.g., Ching *et al.*, 1998; Hogan and Turner, 1998) have shown that some listeners with sensorineural HL (with $\geqslant$55 dB HL at or above 4 kHz) receive limited benefit from amplification of the high-frequency region. While normal-hearing listeners might utilize multiple cues (including low-frequency cues) for the perception of /s/, some studies (e.g., Zeng and Turner, 1990) have reported that HI listeners do not make use of the low-frequency formant transitions, but rather rely on fricative noise. Hence, the present data suggest that some HI listeners might not be able to perceive the high-frequency landmarks present in the signal, despite the high-frequency amplification provided. We can thus infer that the difficulty that HI listeners experience in perceiving speech in noise can be attributed, at least partially, to their inability to perceive high-frequency landmarks, such as those present at the onset of fricatives.

### C. Landmark-detection algorithms

In the context of developing signal processing algorithms for hearing aids or cochlear implant devices, the data from the present study implicate that the obstruent segments need to be treated differently than the sonorant segments of the signal. More precisely, techniques are needed that will make the acoustic landmarks perceptually more distinct and accessible to HI listeners. At the very least, the high-frequency landmarks need to be made accessible to HI listeners. Equivalently, algorithms specially designed to suppress noise in the corrupted obstruent consonants can be applied differentially to the obstruent consonants. In practice, such algorithms need to somehow detect the location of the acoustic landmarks in the signal. Several such landmark-detection algorithms have been proposed and found to perform quite well, at least in quiet (Mermelstein, 1975; Liu, 1996; Junega and Espy-Wilson, 2008). Further research is thus needed to extend and perhaps redesign some of the existing landmark-detection algorithms to perform well in noisy conditions.

## V. SUMMARY AND CONCLUSIONS

The present study assessed the contribution of the information carried by obstruent sounds on speech recognition in noise. Sentence recognition scores improved significantly when the clean obstruent consonants were introduced in the noise-corrupted sentences, and the improvement was largest (nearly threefold increase) at −5 dB SNR (experiment 1). Experiment 2 was designed to assess the factors contributing to the large improvement in performance. This was done by introducing systematically clean obstruent spectral information, up to a cutoff frequency $F_C$, while leaving the remaining region ($\geqslant F_C$ Hz) of the spectrum corrupted (experiment 2). By controlling the value of $F_C$ we were able to control the amount of spectral information available to the listeners during the obstruent segments of the signal. For small values of $F_C$, listeners had access to F1 and voicing information, as well as to low-frequency acoustic landmarks signaling the onset of stops or fricatives. The improvement from transmission of voicing information was quite substantial amounting to roughly 50% of the overall improvement on sentence recognition. The importance of acoustic landmarks to speech recognition in continuous type of noise was cast in a lexical-access framework (Stevens, 2002). Consistent with Steven's (2002) lexical-access model, the data from experiment 2 showed that the absence of reliable (low and high frequency) acoustic landmarks, owing to the corruption of the signal by external noise, can produce voicing and place of articulation errors (see Fig. 5) and can disrupt the syllable structure, known to be important for determining word boundaries in running speech. Analysis of the phoneme errors made on the sentence recognition task confirmed that the errors made in the word-initial position were reduced by 32% when listeners had access to the low-frequency ($F_C \leqslant 1000$ Hz) acoustic landmarks. This outcome provided support for the view that acoustic landmarks can facilitate segmentation and aid listeners determine word boundaries in noisy conditions.


## ACKNOWLEDGMENTS

This research was supported by Grant No. R01 DC007527 from the National Institute of Deafness and other Communication Disorders, NIH. The authors would like to thank the Associate Editor, Dr. Mitchell Sommers, and the two anonymous reviewers for providing valuable comments that improved the manuscript.


## APPENDIX

The NCM (Hollube and Kollmeier, 1996) was used to assess the intelligibility of sentences processed using $F_C \leqslant 1000$ Hz. This measure is similar in some respects to the STI (Steeneken and Houtgast, 1980) in that it computes the

STI as a weighted sum of transmission index (TI) values determined from the envelopes of the probe and response signals in each frequency band (Goldsworthy and Greenberg, 2004). Unlike the traditional STI measure, however, which quantifies the change in modulation depth between the probe and response envelopes using the modulation transfer function, the NCM measure is based on the covariance between the probe and response envelope signals. The NCM measure was modified in our study as follows. The stimuli were first bandpass filtered into nine bands spanning the range 100–1000 Hz. The envelope of each band was computed using the Hilbert transform and downsampled to 50 Hz, thereby limiting the envelope modulation frequencies to 0–25 Hz. Let $x_i(t)$ be the downsampled envelope in band $i$ of the clean (probe) signal and let $y_i(t)$ be the downsampled envelope of the processed (response) signal. The normalized covariance in the $i$th frequency band is computed as

$$r_i = \frac{\Sigma_i(x_i(t) - \mu_i)(y_i(t) - \nu_i)}{\sqrt{\Sigma_i(x_i(t) - \mu_i)^2}\sqrt{\Sigma_i(y_i(t) - \nu_i)^2}}, \quad (1)$$

where $\mu_i$ and $\nu_i$ are the mean values of the $x_i(t)$ and $y_i(t)$ envelopes, respectively. The SNR in each band is computed as

$$\mathrm{SNR}_i = 10 \log_{10}\left(\frac{r_i^2}{1 - r_i^2}\right), \quad (2)$$

and subsequently limited to the range of −15 to 15 dB (as done in the computation of the articulation index). The TI in each band is computed by linearly mapping the SNR values between 0 and 1 as follows:

$$\mathrm{TI}_i = \frac{\mathrm{SNR}_i + 15}{30}. \quad (3)$$

Finally, the TIs are averaged across all low-frequency bands (<1 kHz) to produce a low-frequency NCM measure: $\mathrm{NCM} = \frac{1}{9}\Sigma_{i=1}^{9}\mathrm{TI}_i$. No weighting is applied to the individual TI values (i.e., equal emphasis is placed to all envelopes).

[1]No ramping of the signal and no window overlap was used in order to avoid smearing (temporally) the boundaries between obstruent and sonorant segments. Such smearing would affect the timing of the acoustic landmarks, which are the subject of the present study. Choosing the appropriate duration of the ramping signal was found to be problematic since the duration of the obstruent consonants can vary from as little as 10 to 200 ms. Due to the absence of ramping and window overlap, some small distortion in the signal was noted sometimes due to transient effects. Pilot data indicated that this signal distortion was not substantial or significant in terms of reducing intelligibility.

[2]The intelligibility scores obtained by the first group of subjects listening to the processed VCV stimuli produced by six speakers were slightly lower (<10%) than the corresponding intelligibility scores obtained by the second group of subjects listening to the processed VCV stimuli produced by two speakers. The trend in performance (as shown in Fig. 4), however, was the same. The intelligibility scores obtained by the two groups of listeners in the various conditions were compared using nonparametric two-sample Kolmogorov–Smirnov tests. These tests revealed nonsignificant ($p > 0.05$) difference in scores. The comparison, for instance, of the scores obtained by the two groups of listeners on the $F_C = 1000$ Hz condition at −5 dB SNR revealed nonsignificant ($p = 0.074$) differences.


Assmann, P., and Summerfield, Q. (**2004**). "The perception of speech under adverse conditions," in *Speech Processing in the Auditory System*, edited by S. Greenberg, W. Ainsworth, W. Popper, and R. Fay (Springer, New York), pp. 231–308.

Bashford, J. A., Jr., and Warren, R. M. (**1987**). "Multiple phonemic restorations follow rules for auditory induction," Percept. Psychophys. **42**, 114–121.

Blumstein, S., and Stevens, K. (**1979**). "Acoustic invariance in speech production: Evidence from some measurements of the spectral characteristics of stop consonants," J. Acoust. Soc. Am. **66**, 1001–1017.

Ching, T. Y., Dillon, H., and Byrne, D. (**1998**). "Speech recognition of hearing-impaired listeners: Predictions from audibility and the limited role of high-frequency amplification," J. Acoust. Soc. Am. **103**, 1128–1140.

Cole, R., Yan, Y., Mak, B., Fanty, M., and Bailey, T. (**1996**). "The contribution of consonants versus vowels to word recognition in fluent speech," Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP'96), Atlanta, GA, May 1996, pp. 853–856.

Cutler, A., and Norris, D. (**1988**). "The role of strong syllables in segmentation for lexical access," J. Exp. Psychol. Hum. Percept. Perform. **14**, 113–121.

Cutler, A., and Carter, D. M. (**1987**). "The predominance of strong initial syllables in the English language," Comput. Speech Lang. **2**, 133–142.

Delgutte, B., and Kiang, N. (**1984**). "Speech coding in the auditory nerve: IV. Sounds with consonant-like dynamic characteristics," J. Acoust. Soc. Am. **75**, 897–907.

Festen, J., and Plomp, R. (**1990**). "Effects of fluctuating noise and interfering speech on the speech-reception threshold for impaired and normal hearing," J. Acoust. Soc. Am. **88**, 1725–1736.

Goldsworthy, R., and Greenberg, J. (**2004**). "Analysis of speech-based speech transmission index methods with implications for nonlinear operations," J. Acoust. Soc. Am. **116**, 3679–3689.

Gow, D., Melvold, J., and Manuel, S. Y. (**1996**). "How word onsets drive lexical access and segmentation: Evidence from acoustics, phonology, and processing," in *Proceedings of 1996 International Conference on Spoken Language Processing* (University of Delaware and Alfred I. DuPont Institute, Philadelphia, PA), pp. 66–69.

Hogan, C., and Turner, C. (**1998**). "High-frequency audibility: Benefits for hearing-impaired listeners," J. Acoust. Soc. Am. **104**, 432–441.

Hollube, I., and Kollmeier, K. (**1996**). "Speech intelligibility prediction in hearing-impaired listeners based on a psychoacoustically motivated perception model," J. Acoust. Soc. Am. **100**, 1703–1715.

IEEE Subcommittee (**1969**). "IEEE recommended practice for speech quality measurements," IEEE Trans. Audio Electroacoust. **AU-17**, 225–246.

Junega, A., and Espy-Wilson, C. (**2008**). "A probabilistic framework for landmark detection based on phonetic features for automatic speech recognition," J. Acoust. Soc. Am. **123**, 1154–1168.

Kawahara, H., Masuda-Katsuse, I., and de Cheveigné, A. (**1999**). "Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction: Possible role of a repetitive structure in sounds," Speech Commun. **27**, 187–207.

Kewley-Port, D., Burkle, Z., and Lee, J. (**2007**). "Contribution of consonant versus vowel information to sentence intelligibility for young normal-hearing and elderly hearing-impaired listeners," J. Acoust. Soc. Am. **122**, 2365–2375.

Klatt, D. H. (**1979**). "Speech perception: A model of acoustic-phonetic analysis and lexical access," J. Phonetics **7**, 279–312.

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, S. (**1996**). "Landmark detection for distinctive feature-based speech recognition," J. Acoust. Soc. Am. **100**, 3417–3430.

Loizou, P. (**2007**). "Speech Enhancement: Theory and Practice," (CRC Press, Boca Raton, FL).

Mermelstein, P. (**1975**). "Automatic segmentation of speech into syllabic units," J. Acoust. Soc. Am. **58**, 880–883.

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

Mines, M., Hanson, B., and Shoup, J. (**1978**). "Frequency of occurrence of phonemes in conversational English," Lang Speech **21**, 221–241.

Owren, M., and Cardillo, G. (**2006**). "The relative role of vowels and consonants in discriminating talker identity versus word meaning," J. Acoust. Soc. Am. **119**, 1727–1739.

Parikh, G., and Loizou, P. (**2005**). "The influence of noise on vowel and consonant cues," J. Acoust. Soc. Am. **118**, 3874–3888.

Phatak, S., and Allen, J. (**2007**). "Consonants and vowel confusions in speech-weighted noise," J. Acoust. Soc. Am. **121**, 2312–2326.

Pisoni, D., and Sawusch, J. (**1975**). "Some stages of processing in speech

perception," in *Structure and Process in Speech Perception*, edited by A. Cohen and S. Nooteboom (Springer-Verlag, Berlin), pp. 16–34.
Seneff, S., and Zue, V. (**1988**). "Transcription and alignment of the TIMIT database," Proceedings of the Second Symposium on Advanced Man-Machine Interface Through Spoken Language, Oahu, HI, 20–22 November 1988.
Shannon, R., Jensvold, A., Padilla, M., Robert, M., and Wang, X. (**1999**). "Consonant recordings for speech testing," J. Acoust. Soc. Am. **106**, L71–L74.
Shriberg, L., and Kent, R. (**2003**). *Clinical Phonetics*, 3rd ed. (Allyn and Bacon, Boston, MA).
Smith, R. (**1979**). "Adaptation, saturation and physiological masking of single auditory-nerve fibers," J. Acoust. Soc. Am. **65**, 166–178.
Steeneken, H., and Houtgast, T. (**1980**). "A physical method for measuring speech transmission quality," J. Acoust. Soc. Am. **67**, 318–326.
Stelmachowicz, P., Pittman, A., Hoover, B., and Lewis, D. (**2001**). "Effect of stimulus bandwidth on the perception of /s/ in normal- and hearing-impaired children and adults," J. Acoust. Soc. Am. **110**, 2183–2190.
Stevens, K. (**2002**). "Toward a model for lexical access based on acoustic landmarks and distinctive features," J. Acoust. Soc. Am. **111**, 1872–1891.
Zeng, F.-G., and Turner, C. W. (**1990**). "Recognition of voiceless fricatives by normal and hearing-impaired subjects," J. Speech Hear. Res. **33**, 440–449.