2007

Speech intelligibility in cochlear implant simulations: Effects of carrier type, interfering noise, and subject experience

Nathaniel A. Whitmal, III, *University of Massachusetts - Amherst*
Sarah F. Poissant, *University of Massachusetts - Amherst*
Richard L. Freyman, *University of Massachusetts - Amherst*
Karen S Helfer, *University of Massachusetts - Amherst*

Speech intelligibility in cochlear implant simulations: Effects of carrier type, interfering noise, and subject experience

Nathaniel A. Whitmal III,[a] Sarah F. Poissant, Richard L. Freyman, and Karen S. Helfer
Department of Communication Disorders, University of Massachusetts, Amherst, Massachusetts 01003

(Received 7 November 2006; revised 26 July 2007; accepted 30 July 2007)

Channel vocoders using either tone or band-limited noise carriers have been used in experiments to simulate cochlear implant processing in normal-hearing listeners. Previous results from these experiments have suggested that the two vocoder types produce speech of nearly equal intelligibility in quiet conditions. The purpose of this study was to further compare the performance of tone and noise-band vocoders in both quiet and noisy listening conditions. In each of four experiments, normal-hearing subjects were better able to identify tone-vocoded sentences and vowel-consonant-vowel syllables than noise-vocoded sentences and syllables, both in quiet and in the presence of either speech-spectrum noise or two-talker babble. An analysis of consonant confusions for listening in both quiet and speech-spectrum noise revealed significantly different error patterns that were related to each vocoder's ability to produce tone or noise output that accurately reflected the consonant's manner of articulation. Subject experience was also shown to influence intelligibility. Simulations using a computational model of modulation detection suggest that the noise vocoder's disadvantage is in part due to the intrinsic temporal fluctuations of its carriers, which can interfere with temporal fluctuations that convey speech recognition cues.
 [DOI: 10.1121/1.2773993]

PACS number(s): 43.71.Es, 43.71.Ky, 43.66.Ts [AJO] Pages: 2376–2388

I. INTRODUCTION

Researchers have long sought to determine which acoustic features of speech aid speech recognition in favorable and unfavorable listening conditions. One frequently investigated factor is the degree of spectral resolution required to produce intelligible speech. Experiments assessing the effects of spectral resolution often use speech synthesized by a channel vocoder, comprised of sums of amplitude-modulated carriers (e.g., pure tones or band-limited noise). Most of these experiments have considered the effects of only one carrier type; few have compared two or more carrier types.

This study compares the effects of pure-tone carriers and noise-band carriers on vocoded speech intelligibility. The results of this work are particularly relevant for cochlear implant processing. Similarities between channel vocoders and front-end processing for cochlear implants have led several investigators (Shannon *et al.*, 1995; Dorman *et al.*, 1997; Dorman and Loizou, 1998; Dorman *et al.*, 1998; Fu *et al.*, 1998; Loizou *et al.*, 1999; Friesen *et al.*, 2001; Qin and Oxenham, 2003) to use channel-vocoded speech as a model for cochlear implant processed speech. Researchers comparing implant and simulator performance have consistently shown that intelligible speech can be produced by vocoders with a small number of channels under favorable listening conditions. Shannon *et al.* (1995) showed that experienced subjects could correctly recognize at least 90% of medial vowels, medial consonants, and sentences produced by a four-channel noise-band vocoder. Similarly, the subjects of Dorman *et al.* (1997) achieved recognition scores of 90% correct or better on similar listening materials processed by six-channel noise-band and pure-tone vocoders. Significant differences between the two vocoder types were found in only two cases: recognition of multi-talker vowels (favoring the sine-wave processor), and reception of place of articulation of consonants (favoring the noise-band processor).

The similarities in results of sine-wave and noise-band vocoder studies are somewhat unexpected, since the two types of vocoders produce signals with vastly different subjective characteristics. Given the practical challenges of synthesizing fricatives and bursts with only sine waves (George and Smith, 1997), one might expect to improve the performance of the vocoder by combining two carrier types as Dudley (1939) did in the first channel vocoder system. Toward this end, Whitmal *et al.* (2004) conducted a pilot experiment comparing the intelligibility of speech from six-band tone and noise-band vocoders with speech from a hybrid six-band vocoder using tone carriers in the lowest three bands and noise carriers in the highest three bands. This carrier allocation was intended to improve the quality of phonemes having primary spectral emphasis at either low frequencies (e.g., vowels) or high-frequencies (e.g., fricatives). Results indicated that sentences in quiet and in noise processed by the tone vocoder were more intelligible than those from the hybrid vocoder, which in turn were more intelligible than noise-band vocoded sentences. This surprising result (discussed later in the paper) suggests that the two carrier types may not be interchangeable for cochlear implant simulation experiments.

Two other recent studies have also reported performance advantages for sine-wave vocoders. In pilot testing for vowel recognition and gender identification tasks, Fu *et al.* (2004)

[a]Electronic mail: nwhitmal@comdis.umass.edu

found that normal-hearing subjects listening through one-band and four-band noise-band vocoders performed poorly relative to cochlear-implant users. When they replaced their noise carriers with tone carriers, their subjects obtained higher levels of performance that better resembled the performance of cochlear implant users. Gonzalez and Oliver (2005) directly compared tone-vocoded and noise-vocoded sentences in speaker and gender identification tasks and also measured significantly better performance with the tone-vocoded sentences. The superior performance of tone vocoding in these experiments was attributed to both an apparent preservation of pitch periodicity and the greater sensitivity of the auditory system to amplitude modulation of tones.

Similar vocoder simulations (Dorman and Loizou, 1998; Fu *et al.*, 1998; Friesen *et al.*, 2001; Qin and Oxenham, 2003) have also been used to understand how speech recognition performance for cochlear implants is affected by additive noise. These studies suggest that recognition performance improves as the spectral resolution (reflected by the number of channels) and/or the signal-to-noise ratio (SNR) of the stimuli increases. Qin and Oxenham (2003), in particular, found that competing speech and amplitude-modulated speech-spectrum noise were less efficient maskers for unprocessed sentences than speech-spectrum noise, and more efficient maskers for vocoded sentences using four, eight, or twenty-four bands. The authors suggested that their vocoders were removing cues to pitch perception that made it more difficult for subjects to discriminate between target speech and masking speech.

It is difficult to reconcile the large differences between tone and noise-band vocoders described earlier (Whitmal *et al.*, 2004; Fu *et al.*, 2004; Gonzalez and Oliver, 2005) with the finding of Dorman *et al.* that there were few significant differences between the vocoders. Moreover, while some simulation studies have explored the effects of noise on sine-wave vocoded speech (Dorman *et al.* 1998) and noise-band vocoded speech (Fu *et al.*, 1998; Friesen *et al.*, 2001; Qin and Oxenham, 2003), none has performed a direct comparison of the intelligibility of the two vocoder types for speech in noise. Hence, it is not clear whether any similarities between intelligibility scores for tone and noise-band vocoder-processed speech would be observed in adverse conditions. The purpose of this study was to determine whether tone- and noise-band vocoded speech signals were equally intelligible in both quiet and noisy conditions. The present study consists of four experiments. Experiments 1 and 2 compared the intelligibility of sentences and medial consonants (respectively) for the two vocoders in both quiet and noisy conditions. Experiment 3 tested closed-set consonant recognition extensively in both quiet and noisy conditions for two purposes: analyzing error recognition patterns for each vocoder type, and observing learning effects. Experiment 4 compared the sine-wave vocoder with three different types of noise-band vocoders to determine the effects of carrier envelope fluctuations on consonant recognition.

II. EXPERIMENT 1: INTELLIGIBILITY OF TONE-VOCODED AND NOISE-VOCODED SENTENCES IN QUIET AND IN BACKGROUND NOISE

The purpose of Experiment 1 was to determine whether tone-vocoded and noise-vocoded sentences were equally intelligible in two types of background noise.

A. Methods

1. Subjects

Twelve adult female listeners participated in Experiment 1. The subjects' ages ranged from 21 to 38 years (mean age=25.8 years). All subjects were native speakers of American English with normal hearing. None of the subjects had participated in previous simulation experiments. The subjects were compensated for their participation with either a cash payment or partial course credit.

2. Materials

Stimuli for Experiment 1 consisted of 360 sentences, each containing three key words (Helfer and Freyman, 2004). The key words were one- or two-syllable nouns or verbs from the Francis and Kucera (1982) list of most common words. The sentences were assigned to one of 24 topics (e.g., food, clothing, politics) used to help listeners direct their attention to the target speaker when sentences were heard in the presence of competing speakers. The sentences were uttered by a female speaker with an American English dialect and digitally recorded in a sound-treated booth (IAC 1604) with 16 bit resolution at a 22 050 Hz sampling rate. The COOL EDIT PRO software package (Syntrillium Software, Phoenix, AZ) was used to scale each sentence to the same overall rms level (−24 dB relative to the maximum level for 16 bit resolution). Recognition data acquired in speech-spectrum noise for a previous study (Helfer and Freyman, 2004) were then used to group the 360 sentences into 24 15-sentence equally intelligible lists with average *unprocessed* key word recognition of 50% correct.

3. Processing

Subjects listened to vocoded versions of the above-described sentence materials, presented either as recorded in quiet or with masking noise added at SNRs of 3, 8, 13, or 18 dB before vocoding. Six-channel vocoders were chosen because they approach asymptotic performance in vocoder simulations without imposing ceiling effects (Dorman *et al.*, 1997), correspond to the effective number of channels that many cochlear-implant listeners can access (Dorman *et al.*, 1998), and represent a configuration for which simulation results are similar to results from cochlear-implant systems (Dorman and Loizou, 1998; Friesen *et al.*, 2001).

Two types of masking noise were used: speech-spectrum noise (SSN) and two-talker babble (TTB). The SSN was developed by creating a 110 250 sample white-noise signal (i.e., 5 s at a 22 050 Hz sampling rate), passing it through a 50th-order all-pole filter matched (via Levinson's recursion) to the average autocorrelation function for the 360 sentences, and scaling the filter's output to the average rms level of the 360 sentences. The TTB (used in Freyman *et al.*, 2001) con-

TABLE I. Band parameters for the six-channel vocoder.

Band	1	2	3	4	5	6
Center freq (Hz)	180	446	885	1609	2803	4773
Bandwidth (Hz)	201	331	546	901	1487	2453

sisted of digital recordings (sampling rate: 22 050 Hz) of two college-aged female students speaking different sets of nonsense sentences that were syntactically correct but semantically anomalous. Pauses between sentences were removed to produce two recordings of continuous speech which were then matched in rms level and combined to produce TTB. A 35-s-long stream of the babble was extracted for use in the present investigation.

Maskers for each sentence were derived from segments of either the SSN or TTB. Each segment was the same length as the individual sentence recordings, which each contained a short silent period of variable length both before (estimated mean=0.064 s) and after (estimated mean=0.328 s) the sentence was spoken. The origins of the masker segments were selected at random to prevent the subjects from hearing the same phrases or sections repeatedly. The masker segment was scaled to produce the desired SNR for the particular trial, and the target and masker were summed together before being input to the vocoder under test.

The vocoders used in the experiment were implemented via custom MATLAB software (Mathworks, Natick, MA) using the configuration of Qin and Oxenham (2003). The vocoders filtered their inputs (using sixth-order Butterworth filters) into six contiguous frequency bands in the 80–6000 Hz range. The frequency range was divided equally in terms of the Cam (or equivalent rectangular bandwidth) scale (Glasberg and Moore, 1990), such that each band was approximately 4.65 Cams in width. Bandwidth and band center frequencies are shown in Table I. It should be noted that the center frequency and bandwidth of Band 1, while lower than that used by Dorman *et al.* (1997), are consistent with frequency table parameters for the Nucleus-22 cochlear implant using SPEAK processing (Fu and Shannon, 1999).

Envelopes for each frequency band were obtained from filtered speech by half-wave rectification followed by smoothing with a second-order Butterworth low-pass filter. The bandwidth of the smoothing filter was the smaller of either 300 Hz or half the analysis bandwidth. The resulting envelopes were used to amplitude-modulate one of two carriers: a sine wave (located at the band's center frequency), or white noise subsequently filtered by that band's bandpass filter. The modulated carriers were level matched to their original in-band input signal and summed to produce simulated implant-processed speech.

4. Procedure

Subjects listened to the 24 lists of sentences while seated in a double-walled sound-treated booth (IAC 1604) during one 75 min listening session. Subjects were given breaks between the 12th and 13th lists. Prior to testing, each of the 24 lists was assigned to one of two groups; one for SSN, and one for TTB. Carrier/list pairings for the 12 subjects were determined by two 12×12 Latin squares: one for the SSN group, the other for the TTB group. Each combination of the six SNRs and two carrier configurations was used in processing one of the twelve 15-sentence lists in a group. Sentences within each list were shuffled for presentation in random order. Each subject listened to all of the sentences. Half of the subjects listened to the SSN sentences before listening to TTB sentences; the remaining subjects listened to TTB sentences before listening to SSN sentences.

Custom MATLAB software (executed on a remote computer) was used to present the sentences to the subject and to score the number of key words correctly repeated by her. A laptop screen located inside the test booth prompted the subject with the word "Ready?" and the sentence topic exactly 2 s before the sentence was presented. The sentence was then retrieved from the remote computer's hard disk, converted to an analog signal by the computer's sound card (Analog Devices, SoundMax Integrated Digital Audio) using 16 bit resolution at a 22 050 Hz sampling rate, passed through an attenuator (Tucker-Davis PA4) and headphone amplifier (Tucker-Davis HB5), and presented diotically to the subject at 65 dBA through TDH-50P headphones. (TDH-series headphones have been used to present high-intelligibility vocoded speech in previous studies, e.g., Shannon *et al.*, 1995; 1998.) Presentation levels were calibrated daily using repeated loops of the speech-spectrum noise described earlier. Upon hearing each sentence, the subject repeated what she heard into a talk-back microphone monitored by a researcher, who then recorded the number of key words correctly repeated.

Training materials were limited to ten sentences per carrier type presented without feedback just before starting the experiment. Of the ten sentences, five were presented in quiet and five were presented in speech-spectrum noise at 8 dB SNR. The sentences used for training were not used in the main experiment.

B. Results

Intelligibility scores for Experiment 1 were derived from the percentage of correctly repeated key words per condition. Mean sentence intelligibility scores for SSN and TTB are shown in Fig. 1. Under all test conditions, tone-vocoded sentences were more intelligible than noise-vocoded sentences. Scores for masking with speech-spectrum noise were higher than scores for masking with babble; this is consistent with the results of Qin and Oxenham (2003).

Subject scores were converted to rationalized arcsine units (Studebaker, 1985) and input to a repeated-measures analysis of variance (ANOVA) of intelligibility scores. Within-subject factors for the ANOVA included carrier configuration, noise type, SNR, and carrier presentation order (i.e., first group or second group). Results of the ANOVA indicated that carrier configuration ($F[1,287]=240.62$, $p<0.0001$), noise type ($F[1,287]=21.43$, $p<0.0001$), SNR ($F[5,287]=377.39$, $p<0.0001$), and carrier order ($F[1,287]=11.50$, $p=0.0008$) were all significant main factors. Post hoc tests using Tukey's Honestly Significant Difference (HSD) criterion ($\alpha=0.05$) indicated that (a) tone-based voc-

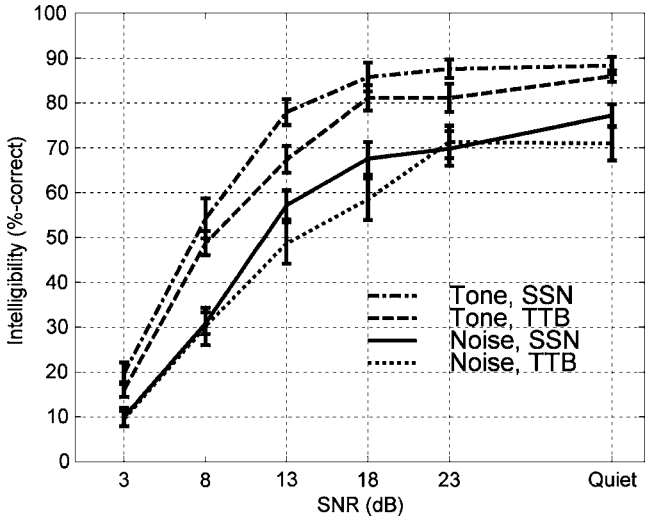

FIG. 1. Experiment 1: Percent-correct recognition scores for words in sentences (±1 SE) mixed with either speech-spectrum noise (SSN) or two-talker babble (TTB) when processed by tone and noise-band vocoders.

oding was significantly more intelligible than noise-band vocoding, (b) TTB was a significantly better masker than SSN, (c) scores at SNRs below 23 dB were significantly different from each other, and (d) average scores for the second carrier presented were significantly greater than average scores for the first carrier. A comparison of first-presented and second-presented means using Tukey's HSD criterion (α=0.05) showed significant gains of 9.5 rational arcsine units (RAU) for both carrier types at 18 dB SNR. Smaller gains ranging between 3.1 and 5.7 RAU were also observed for both carriers at 8 and 13 dB SNR.

III. EXPERIMENT 2: INTELLIGIBILITY OF TONE-VOCODED AND NOISE-VOCODED CONSONANTS IN BACKGROUND NOISE

The tone-vocoded stimuli of Experiment 1 provided approximately 13% higher average intelligibility scores in quiet and approximately 20% higher average scores in masking noise than did the noise-vocoded stimuli. This contrasts markedly with previous research (Dorman *et al.*, 1997). The purpose of Experiment 2 was to determine whether the advantage for tone-vocoded stimuli extends to vowel-consonant-vowel (VCV) syllables, which are free from the syntactic, semantic, or contextual cues provided by the sentences of Experiment 1.

A. Methods

1. Subjects

Six adult female listeners participated in Experiment 2. The subjects' ages ranged from 21 to 38 years (mean age =25.5 years). Five of the subjects were completely inexperienced, having never participated in simulation experiments, and met the screening requirements of Experiment 1. The sixth subject was a participant in Experiment 1 who had no previous experience listening to processed consonant stimuli. The subjects were compensated for their participation with either a cash payment or partial course credit.

2. Materials

Stimuli for Experiment 2 consisted of the 23 consonants /b d g p t k f θ v ð h s ʃ z ʒ ʧ ʤ m n w l j r/, uttered in /a/C/a/ format. The consonants were spoken by a female talker with an American English dialect and digitally recorded in a sound-treated booth (IAC 1604) with 16 bit resolution at a 22 050 Hz sampling rate. As in Experiment 1, consonants were scaled to the same overall rms level.

3. Processing

Subjects listened to vocoded versions of the VCV materials described earlier, presented either as recorded or with the speech-spectrum noise of Experiment 1 added at SNRs of 3, 8, 13, 18, or 23 dB before vocoding. Listening in TTB (the more efficient masker) was eliminated in order to provide a wider range of scores. Vocoders included the tone- and noise-band vocoders of Experiment 1.

4. Procedure

Subjects listened to the VCV syllables while seated in a double-walled sound-treated booth (IAC 1604) during two 1-h listening sessions. Custom MATLAB software was used to present the stimuli to the subjects and to score the number of correctly recognized consonants. The laptop screen displayed a 5×6 grid of buttons, 23 of which contained the English spelling of one of the tokens, e.g., "asha." The unvoiced and voiced consonants /θ ð/ were represented by "atha" and "aTHa," respectively. The screen also contained an indicator sign which prompted the subjects by changing color and presenting the messages "Idle," "Ready?," and "Guess!"

Each trial consisted of two repetitions of each of the 23 consonants, presented in random order. Trials were presented in blocks of two, matched for SNR, with either tone-vocoding first and noise-vocoding second (TN), or noise first and tone second (NT). Subjects listened to one 12-block sequence in each session, with block types alternated (e.g., TN/NT/TN/NT, etc.) Half of the subjects heard sequences beginning with TN in session 1 and NT in session 2; the other half heard NT first in session 1 and TN first in session 2. This alternation method was intended to prevent the subject from having a persistent experience advantage on either vocoder. In order to acclimate subjects to the task, the first block in a sequence was always presented in quiet; SNRs for the remaining blocks were assigned at random without replacement.

The subject was prompted to start each trial with a mouse click when ready. Two seconds prior to each VCV's presentation, the indicator sign turned red and presented "Ready?" to the subject. The VCV was then sent to the computer's sound card (Silicon Integrated Systems 7018) using 16 bit resolution at a 22 050 Hz sampling rate, passed through a custom-built impedance matching network, and presented to the subject's right ear over TDH-50P headphones at 65 dBA. The indicator turned green and presented "Guess?" to the subject. The subject double-clicked on the button labeled with the VCV that she believed she heard. The indicator state returned to "Idle" for 2 s before returning to the "Ready?" state for the next consonant in the trial.

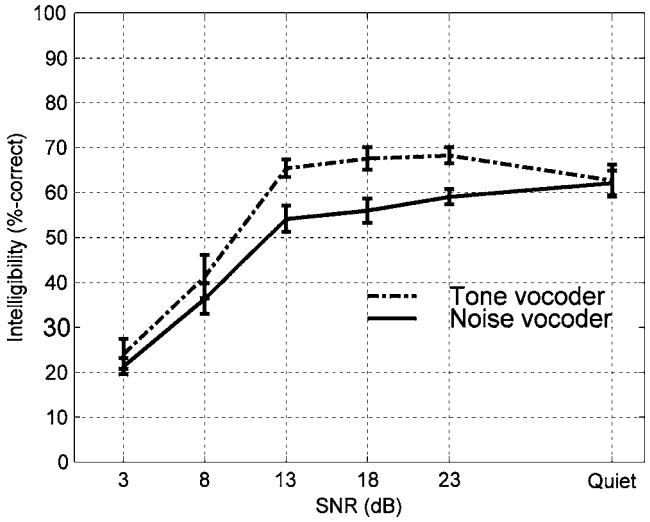

FIG. 2. Experiment 2: Percent-correct recognition scores for vowel-consonant-vowel (VCV) syllables (±1 SE) mixed with speech-spectrum noise when processed by tone and noise-band vocoders. Two blocks of VCVs were allotted to each listening condition.

Training was conducted in two parts using VCVs uttered by a male speaker. First, one repetition of the unprocessed VCV set was presented in quiet. This was intended to (a) familiarize the subject with the response process and (b) make sure that the subject could identify and distinguish the consonants. The lowest score for any of the six subjects was 21 of 23 correct, or 91% correct. Typical errors for these subjects were confusions between different nonstrident fricatives with relatively low intensity and spectral peaks outside the composite bandwidth of the processor/headphone combination (Kent and Read, 2002). Next, four processed VCV sets were presented to the subject to provide limited practice for the main experiment. Each processed set used a unique combination of vocoder (tone or noise) and noise condition (in quiet or at 13 dB SNR).

B. Results

Intelligibility scores for Experiment 2 were derived from the percentage of correctly identified consonants per condition. A test-retest analysis of first- and second-session scores indicated excellent reliability, with significant intersession Pearson product-moment correlations for both vocoders ($r=0.8281$ for noise-vocoded speech and $r=0.9254$ for tone speech with $p<0.0001$). It is important to note that these high correlations only suggest a predictable linear relationship between first- and second-session scores; they do *not* imply that second-session scores are replications of first-session scores (measurement error notwithstanding). On average, Session 2 scores (52.4% correct) were slightly higher than Session 1 scores (50.6% correct), suggesting the possibility of a learning effect.

Mean consonant intelligibility scores for tone-based vocoders and noise-based vocoders are shown in Fig. 2. As in Experiment 1, tone-vocoded speech was more intelligible than noise-band-vocoded speech at SNRs of 8, 13, 18, and 23 dB. Scores for the two vocoders at 3 dB SNR were approximately equal at 23% correct; this likely reflects a floor effect in the data. Scores in quiet were also approximately equal at 62% correct; this likely reflects a training effect (described in the following). Subject scores converted to RAU were input to a repeated-measures ANOVA with main factors of carrier configuration, SNR, and session number (i.e., first or second). Results indicated that carrier configuration ($F[1,143]=43.28$, $p<0.0001$) and SNR ($F[5,143]=187.28$, $p<0.0001$) were both significant main factors. The interaction between SNR and carrier (caused by the near-equality of scores in quiet and at 3 dB SNR) was also significant ($F[5,143]=3.38$, $p=0.0078$). No other interactions were significant. Session number approached (but did not reach) statistical significance ($F[1,143]=3.71$, $p=0.0573$). Post hoc tests using Tukey's HSD criterion ($\alpha=0.05$) indicated that (a) tone-based vocoding was again significantly more intelligible on average than noise-band vocoding, and (b) scores in quiet were significantly higher than those in noise at 3 or 8 dB SNR.

In Experiment 1, tone-based vocoding of quiet speech produced the most intelligible stimuli presented to the subjects. While the score of 62% correct in quiet for tone vocoding is a relatively high score in Experiment 2, it does not again represent the most intelligible of the eight conditions. One possible explanation for this interexperiment difference is that the higher intelligibility of the quiet speech was compromised by making the quiet speech the first signal presented in every session. This ordering would prevent the listener from applying any benefits of experience to the quiet speech. At the same time, the randomized presentation order used for speech in noise would allow listeners to apply varying amounts of experience to each trial. On average, then, vocoding in moderate SNRs like 18 dB might produce better performance than expected relative to scores for vocoding in quiet. At the same time, the combination of highly intelligible quiet speech in the first trial followed by less intelligible speech in successive trials might have confounded any existing learning effects. To explore these possibilities, a second ANOVA was conducted using only scores for speech in noise. Results indicated that removing quiet scores marginalized the SNR-carrier interaction ($F[4,119]=2.41$, $p=0.0573$) and made the advantage of Session 2 scores over Session 1 scores (now 50.9% correct to 48.4% correct) statistically significant ($F[1,119]=5.63$, $p=0.0205$).

IV. EXPERIMENT 3: INVESTIGATION OF ERROR PATTERNS AND LEARNING EFFECTS FOR TONE-VOCODED AND NOISE-VOCODED CONSONANTS IN QUIET AND IN NOISE

The data from Experiments 1 and 2 show an intelligibility advantage for tone-vocoded stimuli, both in quiet and in masking noise, and suggest the possibility of significant learning effects. The purpose of Experiment 3 was to explore the intelligibility advantage and learning effects in greater detail. Tone-vocoded and noise-vocoded VCVs at 8 dB SNR and in quiet were presented to the subjects of Experiment 2. The 8 dB SNR condition was chosen as a representative noise condition that was challenging for subjects but free from floor effects.

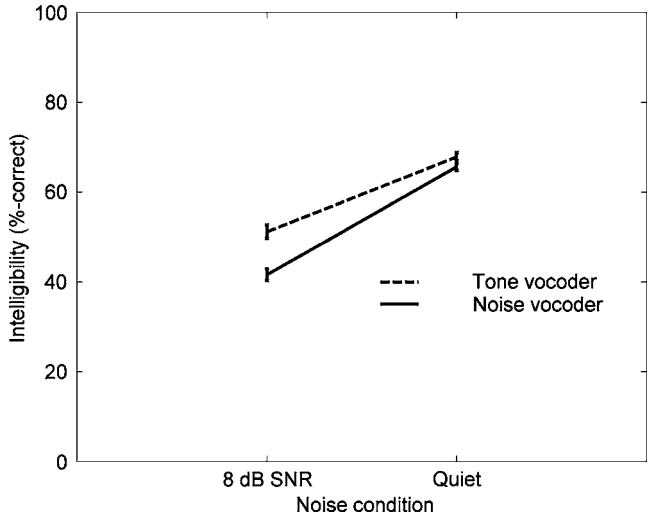

FIG. 3. Experiment 3: Percent-correct recognition scores (±1 SE) for Experiment 2 subjects listening to tone-vocoded and noise-vocoded VCV syllables, either in quiet or mixed with speech-spectrum noise at 8 dB SNR. Five blocks of VCVs were allotted to each listening condition.

A. Methods

1. Subjects

The six adult female listeners of Experiment 2 also participated in Experiment 3.

2. Materials

Subjects listened to tone- and noise-vocoded versions of Experiment 2 consonant recordings, presented either in quiet or in speech-spectrum noise at 8 dB SNR as in Experiments 1 and 2.

3. Procedures

Subjects listened to VCVs under the conditions of Experiment 2 in two listening sessions. The first session was dedicated to listening in speech-spectrum noise; subjects were later asked to return for listening in quiet. Each session consisted of two 10-block sequences, with block types alternated as in Experiment 2 (e.g., TN/NT/TN/NT, etc.). Subjects were given a break between sequences. This protocol allotted a total of five blocks (230 VCVs) to each condition (Experiment 2 allotted 92 VCVs), providing subjects with more rapid and extensive exposure to each condition than allowed in Experiment 2.

B. Results

1. Intelligibility scores

Intelligibility scores for Experiment 3 were derived as in Experiment 2. Mean consonant intelligibility scores for tone-based vocoders and noise-based vocoders are shown in Fig. 3. Two similarities between these data and those of Experiment 2 are noted. First, tone-vocoded speech was more intelligible than noise-vocoded speech. Differences between the carriers were most noticeable at 8 dB SNR, with average tone-vocoded scores of 51.2% correct and average noise-vocoded scores of 41.6% correct observed. Second, scores in quiet were closer in value, with average tone-vocoded scores of 67.9% correct and average noise-vocoded scores of 65.6% correct observed.

TABLE II. Linear regression model of subject learning effects, line parameters, correlation coefficients (r), and significance level (p).

Listening condition		Slope	Intercept	r	p
Tone vocoder	Quiet	0.70	64.07	0.77	0.009
	8 dB SNR	0.72	47.17	0.78	0.007
Noise vocoder	Quiet	0.56	62.57	0.64	0.045
	8 dB SNR	0.39	39.44	0.51	0.128

Intelligibility scores for each vocoder/SNR combination were found to increase in near-linear fashion as a function of trial number as the subjects gained experience. Regression lines fit to data for each vocoder/SNR combination (see Table II) indicated that scores for tone-vocoded consonants increased at the approximate rate of 7 percentage points over ten trials. In contrast, scores for the more difficult noise-vocoded consonants increased at an approximate rate of 4.75 points over ten trials. The regression lines for the tone vocoder also fit the data better ($r=0.77$ in quiet and 0.78 in noise, $p<0.01$) than the lines for the noise vocoder ($r=0.64$ for quiet, $p=0.045$; $r=0.51$ for noise, $p=0.128$).

Subject scores were converted to RAU and input to a repeated-measures ANOVA with main factors of carrier configuration, SNR, and trial number. Results of the ANOVA indicated that carrier configuration ($F[1,239]=60.48$, $p<0.0001$), SNR ($F[1,239]=714.98$, $p<0.0001$), and trial number ($F[9,243]=3.17$, $p=0.0016$) were all significant main factors. The interaction between SNR and carrier (caused by the near-equality of scores in quiet) was also significant ($F[1,243]=21.38$, $p<0.0001$). Post hoc tests using Tukey's HSD criterion ($\alpha=0.05$) indicated that (a) tone vocoding was again significantly more intelligible than noise-band vocoding, (b) scores in quiet were significantly higher than those in noise at 8 dB SNR, and (c) trials 9 and 10 had significantly higher scores than trial 1 and trial 10 had significantly higher scores than trial 2.

2. Consonant confusions

Consonant confusion scores for Experiment 3 were derived from contingency tables (or, confusion matrices) comparing *a priori* voicing, manner, and place of articulation classifications with those produced by the subjects' identification tasks. Consonant confusions for voicing identification (shown in Table III) indicate nearly perfect performance in quiet: 98.7% correct for the tone vocoder versus 97.2% correct for the noise-band vocoder. In noise, errors increased in a complementary fashion, with the tone vocoder causing more unvoiced-to-voiced conversions and the noise vocoder causing more voiced-to-unvoiced conversions. The higher proportion of voiced phonemes (1680 vs 1080) ensured that tone vocoding maintained a small advantage over noise vocoding in noise (91% correct vs 86% correct).

Consonant confusions for manner identification (shown in Table IV) indicate good performance in quiet for both systems, with correct manner identification for 92% of noise-

TABLE III. Voicing confusion analysis of responses to VCV stimuli in Experiment 3.

	Tone vocoder		Noise vocoder	
Responses in quiet				
Stimulus	Unvoiced	Voiced	Unvoiced	Voiced
Unvoiced	1070	10	1058	22
Voiced	24	1656	53	1627
% correct	98.70%		97.20%	
Responses in noise (8 dB SNR)				
Stimulus	Unvoiced	Voiced	Unvoiced	Voiced
Unvoiced	895	185	958	122
Voiced	77	1603	259	1421
% correct	90.50%		86.20%	

vocoded VCVs and 94% of tone-vocoded VCVs. The largest difference between the systems was observed for correct classification of semivowels: 87% correct for noise vocoding, 94% correct for tone vocoding. In noise, performance dropped to 75% correct for noise vocoding and 79% correct for tone vocoding, with distinctive interaction between manner type and vocoder type. Tone carriers provided better recognition for stops, nasals, and semivowels; noise carriers provided better recognition for fricatives only. It is likely that the particular strengths of each vocoder were related to its carrier's ability to produce fricative-like noise. This phenomenon was also reflected in the error patterns for each vocoder.

TABLE IV. Manner confusion analysis of responses to VCV stimuli in Experiment 3.

Stimulus	Stop	Fricative	Affricate	Nasal	Semivowel
Tone vocoder responses in quiet					
Stop	719	1	0	0	0
Fricative	111	936	33	0	0
Affricate	2	3	235	0	0
Nasal	0	0	0	240	0
Semivowel	3	12	0	13	452
Score for this condition: 93.55% correct					
Noise vocoder responses in quiet					
Stop	716	2	2	0	0
Fricative	96	930	28	9	17
Affricate	0	5	234	1	0
Nasal	0	0	0	240	0
Semivowel	3	1	0	55	421
Score for this condition: 92.07% correct					
Tone vocoder responses in noise (8 dB SNR)					
Stop	642	39	13	24	2
Fricative	293	678	39	58	12
Affricate	1	3	236	0	0
Nasal	0	4	0	191	45
Semivowel	25	20	0	2	433
Score for this condition: 78.99% correct					
Noise vocoder responses in noise (8 dB SNR)					
Stop	463	160	90	2	5
Fricative	101	883	53	3	40
Affricate	3	3	234	0	0
Nasal	11	35	1	109	84
Semivowel	15	49	2	38	376
Score for this condition: 74.82% correct					

TABLE V. Place confusion analysis of responses to VCV stimuli in Experiment 3.

Stimulus	Labial	Dental	Alveolar	Palatal	Velar
Tone vocoder responses in quiet					
Labial	575	6	127	11	1
Dental	193	42	2	1	2
Alveolar	285	27	383	9	16
Palatal	50	0	53	495	2
Velar	25	1	8	0	446
Score for this condition: 70.3% correct					
Noise vocoder responses in quiet					
Labial	521	6	171	18	4
Dental	148	43	17	1	31
Alveolar	187	59	455	4	15
Palatal	34	0	44	521	1
Velar	67	3	33	0	377
Score for this condition: 69.5% correct					
Tone vocoder responses in noise (8 dB SNR)					
Labial	396	14	142	35	133
Dental	142	17	7	2	72
Alveolar	170	14	321	11	204
Palatal	30	1	65	486	18
Velar	8	0	7	3	462
Score for this condition: 60.9% correct					
Noise vocoder responses in noise (8 dB SNR)					
Labial	284	50	167	87	132
Dental	118	42	21	2	57
Alveolar	165	69	355	31	100
Palatal	35	8	44	481	32
Velar	75	25	80	19	281
Score for this condition: 52.3% correct					

The tone vocoder caused more fricative-to-stop conversions than the noise vocoder, which in turn converted more stops, nasals, and semivowels to fricatives than the tone vocoder. Similar results for noise vocoding were reported by Drullman *et al.* (1994), who suggested that the increased duration and modified envelopes of their medial stops led their subjects to sometimes identify them as fricatives.

Consonant confusions for place identification (shown in Table V) indicate poorer performance than for either manner or voicing identification in quiet or in noise. In quiet, tone vocoding and noise-band vocoding provided nearly equal accuracy (70.3% correct for tones, 69.5% correct for noise), with vastly different error patterns. Tone carriers provided better recognition for labial consonants (defined here as /p b m w f v/) and velar/glottal consonants, most of which were sonorants or stops with a low-frequency spectral emphasis. Conversely, noise carriers provided better recognition for alveolar consonants, most of which were fricatives or stops with a high-frequency spectral emphasis. Dentals (/θ ð/) and palatals (consisting largely of strident fricatives/affricates) were reproduced nearly as well by both vocoders. In noise, place identification dropped to 52.3% correct for noise vocoding and 60.9% correct for tone vocoding. Error patterns resembled those observed in manner identification: Tone vocoding made more dental-to-labial, dental-to-velar, and alveolar-to-velar conversions than noise vocoding, while noise vocoding made more velar-to-dental and velar-to-

TABLE VI. Kappa coefficients for feature confusion matrices of Experiment 3, with a χ^2 analysis testing the equality of coefficients for each feature.

Feature	Quiet		8 dB SNR		χ^2 for equal kappa test	p
	Noise vocoder	Tone vocoder	Noise vocoder	Tone vocoder		
Voicing	0.94	0.97	0.72	0.80	481.9	<0.001
Manner	0.89	0.91	0.66	0.72	590.4	<0.001
Place	0.60	0.61	0.39	0.50	240.2	<0.001

alveolar conversions than tone vocoding and converted 42% of labials to dentals, alveolars, or palatals. Of the remaining 39% of labials identified correctly, half were due to identification of or confusions between the fricatives /f/ and /v/. It should be noted that the nonstandard labeling of "labial" and "dental" phonemes used here allowed each phoneme to be identified by its voicing, manner, and place without creating more than five place designations.

The robustness of manner and voicing cues and relative fragility of place cues in both quiet and noise is expected, and consistent with results for both natural (Miller and Nicely, 1955) and vocoded (Dorman *et al.*, 1997) speech. The exception to this trend is identification of the palatal consonants /ʃ ʒ ʧ ʤ r/, which were typically higher in level and/or longer in duration than either stops or nonstrident fricatives.

The chance-corrected agreement between stimuli and subject responses for each set of feature confusion matrices was quantified by computing the kappa coefficients (Cohen, 1960) shown in Table VI. Kappa coefficients are commonly used to describe the degree of agreement between two classification approaches, with values of kappa ranging between −1 (denoting no agreement) and +1 (denoting perfect agreement). Kappa values ranging from 0.8 to 1.0 denote near-perfect agreement between stimulus and response; from 0.6 to 0.8, substantial agreement; and from 0.4 to 0.6, moderate agreement (Landis and Koch, 1977). The computed kappa values are consistent with the intelligibility data in Figs. 1–3, with better stimulus-response agreement shown for quiet conditions and tone vocoding. Performance for voicing and manner features, in particular, was excellent for both vocoders in quiet, while performance for place identification in quiet was only moderately accurate for both vocoders. Differences between noise and tone vocoding were most evident in noise, with good performance shown for voicing and manner and only moderate performance shown for place identification. Chi-squared tests (see Table VI) were conducted for each feature to test a null hypothesis of equal kappa values; the hypothesis was rejected in each case, further supporting the advantage of tone vocoding over noise vocoding.

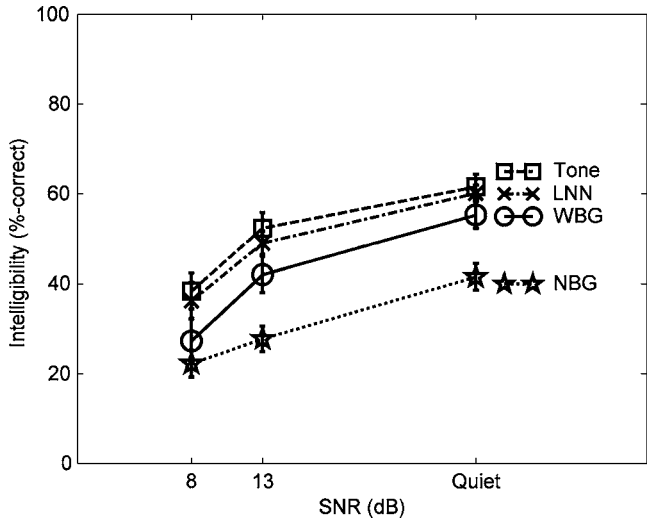

FIG. 5. Experiment 4: Percent-correct recognition scores for vowel-consonant-vowel (VCV) syllables (±1 SE) mixed with speech-spectrum noise when processed by vocoders using tones, low-noise noise bands, wideband Gaussian noise, and narrow-band Gaussian noise as carriers.

V. EXPERIMENT 4: EFFECTS OF CARRIER BANDWIDTH AND ENVELOPE ON INTELLIGIBILITY OF VOCODED CONSONANTS

The data of Experiments 1, 2, and 3 indicate a performance advantage for tone-vocoded stimuli relative to noise-vocoded stimuli. Other researchers reporting performance advantages for tone vocoders (Fu *et al.*, 2004; Gonzalez and Oliver, 2005) suggest that noise carrier envelope modulations reduce intelligibility by impairing detection of speech signal envelope modulations. This suggestion is supported by the computational model of Dau *et al.* (1999), which predicts higher modulation detection thresholds (MDTs) for noise carriers than for the tone carriers. The purpose of Experiment 4 was to investigate the relationship between the intelligibility of vocoded speech and predicted MDTs. The performance of the noise vocoder of Experiments 1 and 2 was compared to that of two additional noise-band vocoders using carriers examined by Dau *et al.* (1999). One of the two additional vocoders used narrow-band Gaussian noise carriers found to impair modulation detection; the other used "low-noise noise" carriers (Pumplin, 1985; Kohlrausch *et al.*, 1997) found to facilitate modulation detection. Predictions of detection thresholds for the carriers were computed and analyzed in association with intelligibility scores.

A. Methods

1. Subjects

Twelve adult female listeners participated in Experiment 4. The subjects' ages ranged from 19 to 25 years. None of the subjects had participated in previous simulation experiments. The subjects were compensated for their participation with either a cash payment or partial course credit.

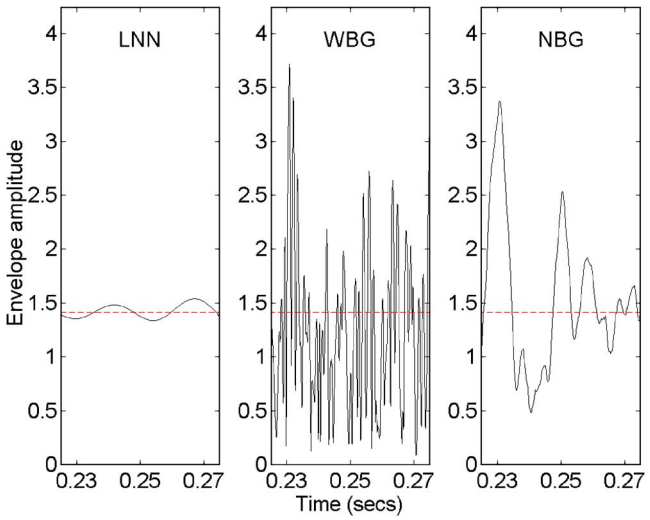

FIG. 4. (Color online) Experiment 4: 50 ms segments of low-noise noise, wideband Gaussian noise, and narrow-band Gaussian noise carrier envelopes in vocoder band 4 at equal rms levels. The envelope of a Band 4 tone carrier (at the same rms level) is also plotted for reference in each panel (dotted line).

2. Processing

Subjects listened to Experiment 2 consonant materials, presented in quiet or with masking noise added at SNRs of 8 or 13 dB before vocoding. These two SNRs were chosen on the basis of data from Experiments 1 and 2 suggesting that the two SNR values were challenging to subjects and less susceptible to either ceiling or floor effects than the other values used.

Four six-channel vocoder configurations were used in the experiment. Two of the vocoders were the tone and wideband Gaussian (WBG) noise vocoders used in the previous three experiments. Carriers for the third vocoder consisted of low-noise noise (LNN) signals (prepared as in Kohlrausch *et al.*, 1997) with bandwidths of 100 Hz and center frequencies matched to those of the vocoder's channels. The resulting carriers contained envelopes that were nearly flat (like the tone carriers' envelopes) with narrow-band spectra. Carriers for the fourth vocoder were narrow-band Gaussian (NBG) noise signals, created by passing white noise through a bank of fourth-order Butterworth filters with bandwidths of 100 Hz and center frequencies matched to those of the vocoder's channels. Representative noise carrier envelope wave forms are shown in Fig. 4. As expected, the LNN carrier presented envelopes with minimal temporal fluctuation, while the Gaussian carriers presented rapidly fluctuating envelopes with spectral energy at frequencies above the speech modulation frequency range (Houtgast and Steeneken, 1985).

3. Procedure

Subjects listened to the VCVs under the conditions described for Experiment 2, with two variations. First, the training session was used to identify viable subjects. As in Experiment 2, one repetition of the unprocessed consonant set was presented in quiet to each subject. Nine of the twelve subjects achieved scores of 87% correct or better. Error patterns for these subjects resembled those of subjects in Experiment 2. The remaining three subjects achieved low scores (65%, 78%, and 78% correct) that reflected poor consonant recognition, and were excused from further participation. Four processed consonant sets in quiet (one per vocoder type) were then presented to the remaining subjects for limited practice. Second, trial ordering for each subject followed an extended form of the ordering scheme of Experiment 2 in order to reproduce (as much as possible) that experiment's learning effects. The twelve subjects were first counted off into four groups of 3 to determine presentation order. Trials were presented in blocks of four (one block per vocoder, as in Experiment 2), matched for SNR. The ordering of the vocoders in each block was based on a fixed sequence of four numbers per subject, with the first number equal to the subject's group number, and the other three following in random order. (Example: A viable sequence for a subject in Group 1 might be {1 4 2 3}). Subjects listened to three 4-block sequences in each session, with the first and third sequences following their set order and the second sequence in reverse order. (Example: Using the above-noted sequence, Subject 5 would hear vocoders in the following order: 1 4 2 3 3 2 4 1 1 4 2 3.) The first block in a sequence was always presented in quiet; SNRs for the remaining blocks were assigned at random without replacement.

B. Results

1. Intelligibility scores

Consonant intelligibility scores for Experiment 4 were derived as in Experiment 2. Mean consonant intelligibility scores for the four vocoders are shown in Fig. 5. As before, tone vocoded speech was more intelligible than WBG noise-vocoded speech. Average scores for the LNN vocoder, however, were only 2.1 percentage points below those of tone-vocoded speech. Speech from the NBG vocoder was least intelligible of all, with its average scores in *quiet* (42% correct) measuring well below average scores at 13 dB SNR (58% correct) for the other three vocoders.

Subject scores were converted to RAU and input to a repeated-measures ANOVA with main factors of carrier configuration and SNR. Results indicated that carrier configuration ($F[3,48]=52.39$, $p<0.0001$) and SNR ($F[2,48]=119.03$, $p<0.0001$) were both significant main factors. No significant interactions were observed. Post hoc tests using Tukey's HSD criterion ($\alpha=0.05$) indicated that (a) results for tone-based vocoding and LNN-based vocoding were not significantly different, (b) speech from the tone and LNN vocoders was significantly more intelligible than the two Gaussian noise vocoders, (c) speech from the wide-band Gaussian noise vocoder was significantly more intelligible than the narrow-band Gaussian noise vocoder, and (d) average scores at all SNRs were significantly different from each other.

Figure 5 also reveals that intelligibility scores for tone and WBG noise vocoders in Experiment 4 were noticeably lower than comparable scores in Experiment 2. The largest discrepancies between the two experiments were observed at 13 dB, the SNR value least susceptible to either floor or

ceiling effects. An inspection of individual scores revealed that intelligibility for most subjects in Experiment 2 tended to increase monotonically with SNR, while several Experiment 4 subjects did not show monotonic increases and often performed far worse than their peers at either 8 or 13 dB. One subject in particular (JW4) produced an average score of 28.6% correct, with lower scores in nine conditions than the other eight Experiment 4 subjects (who averaged 42.8% correct) and lower scores (average: 26.8% correct) in all comparable conditions than the six Experiment 2 subjects (average: 53.68% correct) who also listened to tone and wideband Gaussian noise vocoders in quiet and at 8 and 13 dB SNR. Subject scores from Experiments 2 and 4 were subsequently input to a pair of pooled ANOVAs with main factors of carrier, SNR, and experiment number (i.e., 2 or 4), computed both with and without scores from JW4. Significant interexperiment average differences were seen in both cases: 6.3 RAU when JW4's scores were included, and 3.8 RAU when excluded. In the latter case, Tukey's HSD criterion showed no significant differences between the average scores of individual subjects. Moreover, the largest interexperiment differences were observed with noise-band vocoders; these may be attributed in part to subject experience. In Experiment 2, subject experience was divided between two different vocoders (tone and WBG). In Experiment 4, subject experience was divided among four different vocoders (tone, LNN, WBG, and NBG), with the two most intelligible vocoders sounding (and performing) similarly. The subjects' remaining attention was split between the WBG and the less intelligible NBG vocoder. Reducing the time that the subject spent listening to the WBG vocoder may (as suggested by Experiment 3) have prevented subjects from acclimating to the WBG vocoder over time as well as they may have acclimated to the tone vocoder.

2. Predicted modulation detection thresholds

MDT predictions were computed using the algorithm of Dau *et al.* (1999). Dau *et al.* (1997a, b) had modeled modulation detection in normal-hearing subjects with a signal processor consisting of four stages: peripheral auditory filtering, envelope detection, nonlinear amplitude compression, and a modulation-frequency bandpass filter bank. The processor's responses to probe signals were used as templates in a signal detection stage that simulated subject responses in modulation detection and modulation masking experiments. The authors postulated that MDTs would be influenced by carrier envelope fluctuations transmitted to the output of each modulation-frequency filter. Their resulting simulations (Dau *et al.*, 1997b) were in good agreement with their subjects' responses. In later work (Dau *et al.*, 1999), the authors retained only the envelope detector and modulation filter bank, and modeled MDTs as the ratio of the average ac output power in the filtered envelope to the average carrier power. As before, they found good agreement between their predicted and measured data.

In the present work, carrier levels for the predictions were determined by passing an array of 50 sentences (scaled to 65 dB SPL) through each of the four vocoders and measuring the rms levels prescribed for the carriers in each of the six bands. Five-hundred millisecond segments of the 24 carriers were then generated, scaled to their appropriate rms level, and processed in three stages. First, carrier envelopes were computed from each carrier's Hilbert transform. Each envelope function was then input to a first-order modulation-frequency bandpass filter with complex-valued coefficients (Dau *et al.*, 1997a) centered at one of several modulation frequencies (1, 2, 4, 5, 10, 20, 50, 70, or 100 Hz). The filtered-envelope-to-carrier power ratio was then computed as an estimate of the carrier's MDT. As in Dau *et al.* (1999), a constant value of 0.001 was added to the ratio to impose a threshold floor of −30 dB measured for sine carriers at moderate levels (Dau *et al.*, 1997a).

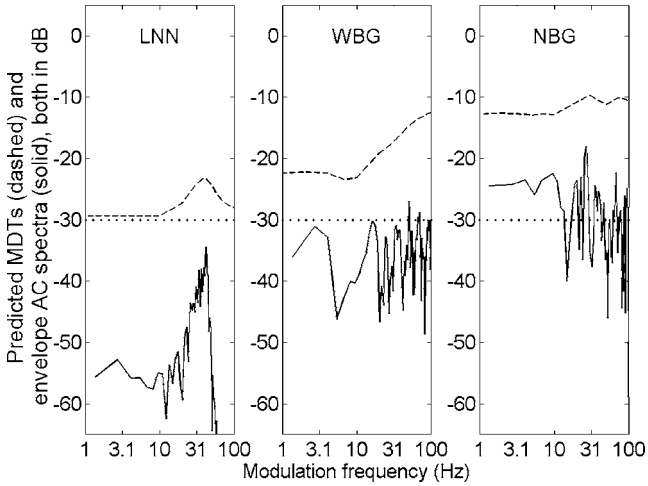

FIG. 6. Experiment 4: Predicted MDTs (upper dashed line) for low-noise noise (LNN), wideband Gaussian noise (WBG), and narrow-band Gaussian (NBG) noise carriers, plotted against modulation frequency. All carriers are limited to the range of vocoder band 4, a band exhibiting moderate differences between WBG and NBG thresholds. Thresholds are expressed as the ratio of the power in a sinusoidal modulation signal to the power of the carrier being modulated. The spectra of each carrier's unfiltered envelope fluctuations about the average envelope value are represented by curves with solid lines. For display purposes only, the envelope energy was normalized to a value of one to facilitate comparison with the MDTs. The MDT of a tone carrier (Dau *et al.*, 1999) is represented by a dotted line.

Representative predicted MDT values for each carrier in Band 4 are indicated in Fig. 6 by dashed lines. The LNN carrier's MDTs are nearly identical to the tone carrier's −30 dB value at speech modulation frequencies (indicated by a dotted line), with a substantial threshold shift only seen near 40 Hz. In contrast, both the WBG and NBG carriers produce higher predicted MDTs than the LNN carrier. Differences between the MDT predictions are due to corresponding differences in the spectra of ac coupled envelope wave forms. Examples of wave form spectra for each carrier are indicated in Fig. 6 by solid lines. Since all carriers are normalized to the same rms level, the narrower bandwidth of the NBG carrier serves to concentrate more carrier envelope power at low modulation frequencies, which in turn lowers the envelope-to-carrier power ratio. As a result, the WBG carrier produces lower predicted MDTs than the NBG carrier, with differences ranging between 1.8 dB (band 1) and 9.0 dB (band 6). This effect is particularly evident in bands 1 and 2, where the 100 Hz NBG carrier bandwidth is compa-

rable to the width of the vocoder band and differences between WBG and NBG thresholds are smallest.

The data of Figs. 5 and 6 suggest a negative association between intelligibility scores and predicted MDTs. Computed Pearson product-moment correlation coefficients for intelligibility scores in quiet and 4 Hz detection threshold levels in individual bands 4, 5, and 6 support this observation ($-1.00 < r < -0.97$, $p < 0.025$). Note that the 4 Hz modulation frequency was selected as a representative frequency for speech modulations (Houtgast and Steeneken, 1985).

VI. DISCUSSION

A. Differences between tone and noise carriers

The results of the present study suggest that tone-vocoded speech and noise-vocoded speech are not equally intelligible in either quiet or noisy listening conditions for subjects with limited training. In Experiment 1, subject scores for a sentence identification task were significantly higher for tone-vocoded speech than for noise-vocoded speech, with measured differences of 13 percentage points in quiet and 17 percentage points in masking noise. Likewise, in Experiment 2, subject scores for a consonant identification task in masking noise were 8 percentage points higher (a significant difference) for tone-vocoded speech than for noise-vocoded speech. The results of Experiment 4 suggest that these differences between vocoders may be due in part to extraneous envelope fluctuations. Tone carriers have no intrinsic envelope fluctuations, and facilitate better modulation detection than noise-band carriers (Dau *et al.*, 1999). As a result, tone carriers appear to be better at faithfully reproducing speech envelope fluctuations than noise-band carriers. This theory is supported by other studies (Cazals *et al.*, 1994; Fu, 2002) showing negative associations between intelligibility scores and MDTs in cochlear implant users. It may also explain why the hybrid vocoder of Whitmal *et al.* (2004) with noise-band carriers in high-frequency bands did not perform as well as the tone vocoder. Conversely, it is plausible that noise-band carrier fluctuations were occasionally mistaken for actual speech envelope fluctuations. This phenomenon may explain why the Experiment 3 vocoders produced manner-of-articulation errors consistent with the ability to produce fricative-like noise. Specifically, the noise vocoder converted 35% of stop consonants into fricatives or affricates, while the tone vocoder converted 27% of fricatives into stops.

Masker envelopes may provide another source of extraneous envelope fluctuations. Results from Experiments 1 and 2 showed that TTB is a more efficient masker of vocoded speech than speech-spectrum noise. Other researchers obtaining similar results with speech maskers (Kwon and Turner, 2001; Qin and Oxenham, 2003; Stickney *et al.*, 2004; Nelson and Jin, 2004; Gonzalez and Oliver, 2005) suggest that the limited spectral resolution of vocoded speech forces listeners to rely more heavily on temporal cues that become degraded by the intrinsic envelope fluctuations of the noise-band carriers. In particular, Qin and Oxenham (2003) note that the vocoders' poor spectral resolution at low frequencies hinders subjects from using target/masker pitch differences and/or comodulation masking release to separate the target from the masker.

Carrier sidebands are a second major source of difference between the tone and noise-band vocoders. Amplitude-modulated tone carriers have sidebands that (depending on auditory filter responses at carrier frequencies) can provide additional detection cues. The 300 Hz bandwidth of the present study's envelope smoothing filters is broad enough to pass pitch-related periodic temporal fluctuations into the modulation envelope. As a result, envelope spectrum components at the talker's average fundamental frequency (approximately 215 Hz) appear as carrier sidebands that are outside the passbands of auditory filters centered at the carrier frequencies for bands 1–5. These sidebands impose a periodic temporal structure on the vocoder's output, with the talker's pitch accurately replicated over a majority of voiced segments (de Cheveigné and Kawahara, 2002). Gonzalez and Oliver (2005), who used envelope smoothing filters with a minimum bandwidth of 160 Hz, observed similar replications of pitch structure in their tone-vocoded signals. In contrast, sidebands of noise-band carriers are masked by carrier spectra, such that no periodic temporal structure is produced.

B. Comparisons with previous vocoder implementations

The results of Experiments 1 and 2 contrast with the results of Dorman *et al.* (1997), who measured intelligibility scores above 90% correct and perfect manner transmission for their six-band tone and noise vocoders. In addition, place transmission was significantly better for their noise vocoder (76% correct) than for their tone vocoder (70% correct). They attributed the noise vocoder's advantage to its contiguous bands, which they argued were better suited than tones for transmitting some place cues. The differences between the data of Dorman *et al.* (1997) and the data in this study may be attributed in part to differences in vocoder spectral resolution and emphasis. Previous work (Dorman *et al.*, 1998; Loizou *et al.*, 1999) suggests that higher spectral resolution in the 900–2500 Hz region can improve perception of vocoded consonants and second formants. Shannon *et al.* (1998) reduced the intelligibility of speech for a four-band vocoder (Shannon *et al.*, 1995) from 95% correct to 80% correct by increasing the tonotopic width of the band containing 1500 Hz (a selected boundary point) from 9.7% of basilar membrane length (Greenwood, 1990) to 12.8% and 14.7%. Likewise, the six-band Dorman *et al.* (1997) vocoder used tonotopic bandwidths of 8.6–9.4% in the 900–2500 Hz region, whereas the vocoder in the present work used bandwidths ranging between 10.6% and 11%. The front ends of the Shannon *et al.* (1995) and Dorman *et al.* (1997) vocoders also featured 1200 Hz high-pass pre-emphasis filters for limiting upward spread of masking. Their combinations of pre-emphasis and spectral resolution provide high-frequency emphasis that (in informal listening) produces a clearer, more intelligible signal that may account for the best scores of the vocoders of both this study and of Qin and Oxenham (2003) being significantly lower than the 90%-correct scores reported by Dorman *et al.* (1997). (It should be noted that Qin

and Oxenham did not publish their raw data; performance here is inferred from their published two-parameter cumulative Gaussian functions to which they fit individual subject data
(mean=SRT, SD=slope) for listening in either SSN or to a single-talker masker.)

C. Effects of subject training

The extent of the subjects' training may also have influenced the results of the present experiments. Other studies using more extensive training regimens reported higher intelligibility scores. Shannon *et al.* (1995) reported that their subjects practiced listening to the stimuli for 8–10 h, with reported scores above 90% correct reflecting stabilized performance. Dorman *et al.* (1997) provided subjects with two passes through all items with visual feedback, ran a sample test with visual feedback given after each answer, and then presented test conditions in increasing order of difficulty (i.e., nine channels first, followed by eight channels, then seven, etc.) in order to better familiarize subjects with the tasks. Their subjects also achieved sentence recognition scores greater than 90% correct. Loizou *et al.* (1999) later measured sentence intelligibility of only 63% correct with the same vocoder using target sentences uttered by 135 different talkers. The authors subsequently argued that the single-talker targets used in previous studies helped elevate intelligibility scores by eliminating the need for subjects to learn to adapt to varying stimuli. Stone and Moore (2003) presented subjects with five counterbalanced blocks of 45 sentences each in a sentence intelligibility task and observed a 41%-correct increase in average intelligibility between the first and fifth block presentations. Since most of the increase occurred within the first two blocks, they suggested that investigators provide between 30 and 60 min of training for subjects in future studies. Given the effects of training and experience, it is possible that a comparison of sine-wave and noise-band processed stimuli conducted without extensive training may be more sensitive to differences in intelligibility than that of Dorman *et al.*, particularly in noise where useful recognition cues are further obscured.

D. Implications for cochlear implant simulation

Finally, the choice of carrier type for optimal simulation of cochlear implant performance requires some discussion. The speech recognition capabilities of subjects in cochlear implant simulation studies using both tone carriers (Dorman *et al.*, 1998) and noise-band carriers (Fu *et al.*, 1998; Friesen *et al.*, 2001) have resembled those of the best cochlear implant users in these studies' patient populations. However, many contemporary cochlear implant processors excite electrodes with amplitude-modulated pulse train carriers, rather than tone or noise carriers. As a result, there are often substantial differences between auditory nerve responses to implant stimuli and acoustic stimuli from a vocoder (Litvak *et al.*, 2001) that prevent either scheme from precisely modeling implant performance.

One rationale for using tone vocoders as cochlear implant simulators comes from Dorman *et al.* (1997), who reported that cochlear implant users perceived individual channel stimulation as tone-like percepts, rather than noise-like percepts. The gender/speaker identification studies of Fu *et al.* (1998) and Gonzalez and Oliver (2005) further suggest that the perception of tone-vocoded speech is better and more similar to that of cochlear implant users than noise-band vocoded speech. At the same time, the fine spectral resolution that makes speech from tone vocoders more intelligible prevents the vocoders from accurately modeling channel interaction effects. This is illustrated by the work of Fu and Nogaki (2005), who showed that noise-band vocoders with broadly overlapping channel filters were better matched to cochlear implant performance than vocoders with steeply sloping channel filters. The authors noted similarities between four-channel vocoders with steeply sloped filters and eight-channel vocoders with overlapping filters, suggesting that there may be a range of acceptable channel bandwidth/channel overlap simulator combinations, none of which can be realized with a tone vocoder. Further research will be required to develop simulation methods that properly incorporate the salient features of both vocoder types.

The results of the present study are consistent with those of Fu *et al.* (2004) and Gonzalez and Oliver (2005) favoring the fidelity of tone vocoders. However, this fidelity difference does not address the question of whether one vocoder is a better simulator of implant performance. Moreover, the small intersubject performance differences observed for either of the present vocoders do not accurately model the large intersubject performance differences observed within or across studies of actual implant users (Friesen *et al.*, 2001; Fu *et al.*, 1998; Stickney *et al.*, 2004). Differences observed across studies may be in part attributable to implant/vocoder parameter differences (e.g., frequency range, number/width of bands, ceiling/floor effects, frequency allocation tables). Given these differences, it is not necessarily possible to assess conditions within one study that closely match the processing characteristics of all active CI recipients. There is, however, much to be learned from vocoder experiments, particularly in light of the fact that researchers may recruit large numbers of subjects, evaluate parameters not easily manipulated in actual cochlear implant systems, and then select the most important and sensitive parameters for use in studies with actual cochlear implant recipients.

VII. SUMMARY AND CONCLUSION

Channel vocoders employing either tone or bandpass noise carriers are often used to simulate cochlear implant processing in normal-hearing listeners. Previous research has suggested that the two types of carriers provide similarly high levels of performance in vocoders with as few as four bands. The present work compared tone and noise vocoders with six bands in both quiet and noisy listening conditions with subjects who have not undergone extensive training. In all four experiments, vocoders using tone carriers produced more intelligible speech than vocoders using noise carriers. An analysis of consonant confusions indicated that recognition error patterns for the two types of vocoders were also significantly different. These differences in performance

were attributed in part to the noise carriers' intrinsic fluctuations, which can impair detection of envelope fluctuations produced by speech, and in part to sidebands imparting a periodic temporal structure in voiced speech segments. Differences between the present results and those of previous studies (Shannon *et al.*, 1995; Dorman *et al.*, 1997) were attributed in part to differences in vocoder parameters and subject training protocols. These factors typically vary widely from experiment to experiment. Future research directed at understanding the effects of these factors may result in improved models of cochlear implant processing and perception.

ACKNOWLEDGMENTS

We would like to thank Christine Alexander, Gail Brown, Kristina Curro, Beth Ann Jacques, Katelyn McLaughlin, Heather Nunes, and Natalie Sitko for help in testing subjects. We would also like to thank Associate Editor Andrew Oxenham and two anonymous reviewers for their helpful suggestions. Funding for this research was provided by the National Institutes of Health (NIDCD Grant Nos. R01 DC01625 and R03 DC7969-01).

Cazals, Y., Pelizzone, M., Saudan, O., and Boex, C. (**1994**). "Low-pass filtering in amplitude modulation detection associated with vowel and consonant identification in subjects with cochlear implants," J. Acoust. Soc. Am. **96**, 2048–2054.

Cohen, J. (**1960**). "A coefficient of agreement for nominal scales," Educ. Psychol. Meas. **20**, 37–46.

Dau, T., Kollmeier, B., and Kohlrausch, A. (**1997a**). "Modeling auditory processing of amplitude modulation. I. Detection and masking with narrow-band carriers," J. Acoust. Soc. Am. **102**, 2982–2905.

Dau, T., Kollmeier, B., and Kohlrausch, A. (**1997b**). "Modeling auditory processing of amplitude modulation. II. Spectral and temporal integration," J. Acoust. Soc. Am. **102**, 2906–2919.

Dau, T., Verhey, J., and Kohlrausch, A. (**1999**). "Intrinsic envelope fluctuations and modulation-detection thresholds for narrow-band noise carriers," J. Acoust. Soc. Am. **106**, 2752–2760.

de Cheveigné, A., and Kawahara, H. (**2002**). "YIN, a fundamental frequency estimator for speech and music," J. Acoust. Soc. Am. **111**, 1917–1930.

Dorman, M. F., and Loizou, P. C. (**1998**). "The identification of consonants and vowels by cochlear implant patients using a 6-channel continuous interleaved sampling processor and by normal-hearing subjects using simulations of processors with two to nine channels," Ear Hear. **19**, 162–166.

Dorman, M. F., Loizou, P. C., Fitzke, J., and Tu, Z. (**1998**). "The recognition of sentences in noise by normal-hearing listeners using simulations of cochlear-implant signal processors with 6–20 channels," J. Acoust. Soc. Am. **104**, 3583–3585.

Dorman, M. F., Loizou, P. C., and Rainey, D. (**1997**). "Speech intelligibility as a function of the number of channels of stimulation for signal processors using sine-wave and noise-band outputs," J. Acoust. Soc. Am. **102**, 2403–2411.

Drullman, R., Festen, J. M., and Plomp, R. (**1994**). "Effect of temporal envelope smearing on speech reception," J. Acoust. Soc. Am. **95**, 1053–1064.

Dudley, H. (**1939**). "Remaking speech," J. Acoust. Soc. Am. **11**, 169–177.

Francis, W. N., and Kucera, H. (**1982**). *Frequency Analysis of English Usage: Lexicon and Grammar* (Houghton Mifflin, Boston).

Freyman, R. L., Balakrishnan, U., and Helfer, K. S. (**2001**). "Spatial release from informational masking in speech recognition," J. Acoust. Soc. Am. **109**, 2112–2122.

Friesen, L. M., Shannon, R. V., Baskent, D., and Wang, X. (**2001**). "Speech recognition in noise as a function of the number of spectral channels: Comparison of acoustic hearing and cochlear implants," J. Acoust. Soc. Am. **110**, 1150–1163.

Fu, Q.-J. (**2002**). "Temporal processing and speech recognition in cochlear implant users," NeuroReport **13**, 1635–1639.

Fu, Q.-J., Chinchilla, S., and Galvin, J. J. (**2004**). "The role of spectral and temporal cues in voice gender discrimination by normal-hearing listeners and cochlear implant users," J. Assoc. Res. Otolaryngol. **5**, 253–260.

Fu, Q.-J., and Nogaki, G. (**2005**). "Noise susceptibility of cochlear implant users: The role of spectral resolution and smearing," J. Assoc. Res. Otolaryngol. **6**, 19–27.

Fu, Q.-J., and Shannon, R. V. (**1999**). "Effects of electrode configuration and frequency allocation on vowel recognition with the Nucleus-22 cochlear implant," Ear Hear. **20**, 321–331.

Fu, Q.-J., Shannon, R. V., and Wang, X. (**1998**). "Effects of noise and spectral resolution on vowel and consonant recognition: Acoustic and electric hearing," J. Acoust. Soc. Am. **104**, 3586–3596.

George, E. B., and Smith, M. J. T. (**1997**). "Speech analysis/synthesis and modification using an analysis-by-synthesis/overlap-add sinusoidal model," IEEE Trans. Speech Audio Process. **5**, 389–406.

Glasberg, B. R., and Moore, B. C. J. (**1990**). "Derivation of auditory filter shapes from notched-noise data," Hear. Res. **47**, 103–138.

Gonzalez, J., and Oliver, J. C. (**2005**). "Gender and speaker identification as a function of the number of channels in spectrally reduced speech," J. Acoust. Soc. Am. **118**, 461–470.

Greenwood, D. D. (**1990**). "A cochlear frequency-position function for several species—29 years later," J. Acoust. Soc. Am. **87**, 2592–2605.

Helfer, K. S., and Freyman, R. L. (**2004**). "Development of a topic-related sentence corpus for speech perception research," J. Acoust. Soc. Am. **115**, 2601–2602.

Houtgast, T., and Steeneken, H. J. M. (**1985**). "A review of the MTF concept in room acoustics and its use for estimating speech intelligibility in auditoria," J. Acoust. Soc. Am. **77**, 1069–1077.

Kent, R. D., and Read, C. (**2002**). *The Acoustic Analysis of Speech* (Delmar/Thomson Learning, Albany, NY).

Kohlrausch, A., Fassel, R., van der Heijden, M., Kortekaas, R., van de Par, S., Oxenham, A., and Puschel, D. (**1997**). "Detection of tones in low-noise noise: Further evidence for the role of envelope fluctuations," Acust. Acta Acust. **83**, 659–669.

Kwon, B. J., and Turner, C. W. (**2001**). "Consonant identification under maskers with sinusoidal modulation: Masking release or modulation interference?," J. Acoust. Soc. Am. **110**, 1130–1140.

Landis, J. R., and Koch, G. G. (**1977**). "The measurement of observer agreement for categorical data," Biometrics **33**, 159–174.

Litvak, L., Delgutte, B., and Eddington, D. (**2001**). "Auditory nerve fiber responses to electric stimulation: Modulated and unmodulated pulse trains," J. Acoust. Soc. Am. **110**, 368–379.

Loizou, P. C., Dorman, M., and Tu., Z. (**1999**). "On the number of channels needed to understand speech," J. Acoust. Soc. Am. **106**, 2097–2103.

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

Nelson, P. B., and Jin, S.-H. (**2004**). "Factors affecting speech understanding in gated interference: Cochlear implant users and normal-hearing listeners," J. Acoust. Soc. Am. **115**, 2286–2294.

Nilsson, M., Soli, S., and Sullivan, J. (**1994**). "Development of the hearing in noise test for the measurement of speech reception thresholds in quiet and in noise," J. Acoust. Soc. Am. **95**, 1085–1099.

Pumplin, J. (**1985**). "Low-noise noise," J. Acoust. Soc. Am. **78**, 100–104.

Qin, M. K., and Oxenham, A. J. (**2003**). "Effects of simulated cochlear-implant processing on speech reception in fluctuating maskers," J. Acoust. Soc. Am. **114**, 446–454.

Shannon, R. V., Zeng, F.-G., Kamath, V., Wygonski, J., and Ekelid, M. (**1995**). "Speech recognition with primarily temporal cues," Science **270**, 303–304.

Shannon, R. V., Zeng, F.-G., and Wygonski, J. (**1998**). "Speech recognition with altered spectral distribution of envelope cues," J. Acoust. Soc. Am. **104**, 2167–2476.

Stickney, G. S., Zeng, F.-G., Litovsky, R., and Assmann, P. (**2004**). "Cochlear implant speech recognition with speech maskers," J. Acoust. Soc. Am. **116**, 1081–1091.

Stone, M. A., and Moore, B. C. J. (**2003**). "Effect of the speed of a single-channel dynamic range compressor on intelligibility in a competing speech task," J. Acoust. Soc. Am. **114**, 1023–1034.

Studebaker, G. A. (**1985**). "A 'rationalized' arcsine transform," J. Speech Hear. Res. **28**, 455–462.

Whitmal, N. A., Poissant, S., Freyman, R. L., and Helfer, K. S. (**2004**). "Effect of combining different carriers across bands on speech intelligibility in cochlear implant simulation," 27th Annual Mid-Winter Meeting of the Association of Research in Otolaryngology, Daytona Beach, FL.