Research Article

Spectral and Temporal Envelope Cues for Human and Automatic Speech Recognition in Noise

GUANGXIN HU,¹ SARAH C. DETERMAN,¹ YUE DONG,¹ ALEC T. BEEVE,¹ JOSHUA E. COLLINS,¹ AND YAN GAI¹

¹Biomedical Engineering Department, Saint Louis University, 3007 Lindell Blvd Suite 2007, St Louis, MO 63103, USA

Received: 15 March 2019; Accepted: 16 September 2019; Online publication: 22 November 2019

ABSTRACT

Acoustic features of speech include various spectral and temporal cues. It is known that temporal envelope plays a critical role for speech recognition by human listeners, while automated speech recognition (ASR) heavily relies on spectral analysis. This study compared sentence-recognition scores of humans and an ASR software, Dragon, when spectral and temporal-envelope cues were manipulated in background noise. Temporal fine structure of meaningful sentences was reduced by noise or tone vocoders. Three types of background noise were introduced: a white noise, a time-reversed multi-talker noise, and a fake-formant noise. Spectral information was manipulated by changing the number of frequency channels. With a 20-dB signal-to-noise ratio (SNR) and four vocoding channels, white noise had a stronger disruptive effect than the fake-formant noise. The same observation with 22 channels was made when SNR was lowered to 0 dB. In contrast, ASR was unable to function with four vocoding channels even with a 20-dB SNR. Its performance was least affected by white noise and most affected by the fake-formant noise. Increasing the number of channels, which improved the spectral resolution, generated non-monotonic behaviors for the ASR with white noise but not with colored noise. The ASR also showed highly improved performance with tone vocoders. It is possible that fake-formant noise affected the software's performance by disrupting spectral cues, whereas white noise affected performance by compromising speech seg-

mentation. Overall, these results suggest that human listeners and ASR utilize different listening strategies in noise.

Keywords: noise vocoding, tone vocoding, speech recognition, formants, spectral, temporal, automated speech recognition, speech segmentation

INTRODUCTION

Natural speech contains redundant acoustic features and information (Stilp 2011; Beekhuizen et al. 2013). The redundancy is reflected in the fact that human listeners can often recognize a word or phoneme that is poorly presented, e.g., when certain acoustic features are absent or distorted. Speech features can be largely divided into two categories: spectral and temporal. Temporal cues include the slow-varying envelope, which is the instantaneous sound amplitude, and temporal fine structure, which is the fast-changing structures associated to the center frequency of a filter (Rosen 1992). To evaluate the two cues in speech perception, vocoding techniques using noise or tone carriers have been applied to extract the temporal envelope cue (Shannon et al. 1995). A general finding is that normal-hearing listeners can correctly recognize speech in an easy-listening condition (e.g., no background noise) based on the temporal-envelope cue provided in as few as four channels (Shannon et al. 1995; Dorman et al. 1997a; Loizou et al. 1999; Friesen et al. 2001; Qin and Oxenham 2003); having more than four channels generates only a minor improvement. In other words, temporal envelope information was largely sufficient for listeners to perceive speech in quiet. Apart from the vocoding studies, other studies also highlight the importance of temporal-envelope cues in speech recognition and signal detection. For example, envelope cues

Guangxin Hu and Sarah C. Determan are both first authors with equal contributions.

Correspondence to: Yan Gai · Biomedical Engineering Department · Saint Louis University · 3007 Lindell Blvd Suite 2007, St Louis, MO 63103, USA. Telephone: (314) 977-9739; email: yan.gai@slu.edu

become increasingly important for consonant identifications as the center frequency moves to high frequencies (Ardoint et al. 2011); models based on envelope cues can best account for detections in reproducible noise under both diotic and dichotic conditions (Mao and Carney 2014, 2015; Mao et al. 2015).

Spectral cues include spectral shapes, formant spacing, and formant transitions. In the context of vocoded speech, spectral cues refer to the energy distribution among different frequency channels and are highly affected by the number of channels. With 22 frequency channels, a rough spectral profile of speech can be derived, and most of the formant frequencies can be resolved (Roberts et al. 2011). When the number of channels decreases to four, formant frequencies may not be resolved. Now one may ask the following questions: If humans heavily relied on spectral cues, how would their speech-recognition behaviors change? What are the advantages of using temporal cues, especially when background noise is present and fine-structure information is compromised? It should be noted that the above-mentioned speech cues are not independent. Specifically, envelope cues may be partially recovered from broadband (Zeng et al. 2004) and narrowband (Gilbert and Lorenzi 2006; Heinz and Swaminathan 2009) fine-structure speech. Ghitza (2001) showed that, even after the envelope has been smoothed, the cochlear output can still reconstruct fast-changing envelope information. In addition, with decreased channel numbers, envelope and fine-structure cues may also vary, depending on the vocoding strategy (e.g., the channel bandwidths).

Here, we use artificial intelligence to provide insights on the question related to spectral and temporal cues for speech-in-noise detection. Automated speech recognition (ASR) heavily relies on spectral cues, while incorporating some temporal cues mostly for the purpose of speech segmentation (Atal and Hanaver 1971; Makhoul 1975; Davis and Mermelstein 1990; Ali 1999; Rao and Kumaresan 2000; Juneja and Espy-Wilson 2008). The present study compared the intelligibility of noise-vocoded speech between human subjects and an ASR software. Three types of vocoders were implemented: a “narrowband vocoder” using the human equivalent-rectangular bandwidths (ERBs) for each frequency channel, a “continuous vocoder” that broadened the bandwidths as the number of channels decreased, and a sinusoidal “tone vocoder.” We are especially interested in the noise resiliency of the temporal vs. spectral cues under various types of background noise profiles. Here, vocoding served two purposes—it reduced temporal fine structure detail and also allowed us to control the spectral resolution. It has been shown that more than four frequency channels are needed for speech recogni-

tion in background noise (Eisenberg et al. 2000; Friesen et al. 2001; Shannon et al. 2004), suggesting that better spectral cues or more channels of narrow-band envelopes are required when listening in noise. On the other hand, speech-recognition algorithms that depend on spectral profiles using linear prediction or other procedures are found to be unreliable with noise-induced energy fluctuation (Allen 1995; Rao and Kumaresan 2000). To address this disparity, we presented vocoded speech in white noise and two types of spectrally colored noise. The variation of the noise’s spectro-temporal features and the corresponding effects on sentence recognition by humans and ASR can help examine the hypothesis that humans and ASR utilize different listening strategies in noise.

METHODS

Sound Waveforms

Calandruccio and Smiljanic (2012) designed a list of 500 meaningful sentences, each having four keywords. Those sentences were intended to be less linguistically biased and more applicable for non-native, as well as native, English speakers. Two sample sentences are *The PARK OPENS in ELEVEN MONTHS*, and *He LOST his WHITE HAT TODAY*. Our study mostly used the first 400 sentences, with one condition using the last 100, from their list. Figure 1(A) (*leftmost*) shows the waveform and spectrogram of a sample sentence for the first 1.2 s. Although we have also obtained their audio recordings, we recruited two native English speakers to reread the sentences, because each speaker had to read an additional training story to establish a unique profile with the ASR software (described later). The sentences were read in a double-walled sound booth (8' × 8' × 8.1'; Noise Barriers LLC, Illinois) and recorded by a digital speech recorder at a sampling frequency of 44.1 kHz (Philips DVT3000). The female speaker read the first 200 sentences, and the male speaker read the second 200. The sentence durations varied between 1.2 and 2.9 s, with a mean of 1.9 s.

The vocoding technique based on Shannon et al. (1995) was applied to the quiet or noisy speech with different vocoder types. The number of channels was 22 or 4 for the human subjects and was systematically varied for the ASR (described later) because previous studies have extensively examined the effect of varying the number of channels with human subjects (Shannon et al. 1995; Dorman et al. 1997b; Loizou et al. 1999; Friesen et al. 2001; Shannon et al. 2004). The center frequencies (CFs) were evenly spaced on a logarithmic scale from CF₁ to 10 kHz, with CF₁ being 300 or 500 Hz. Three types of vocoders were examined. The bandwidth of the “ERB vocoder” was

FIG. 1. Sound stimuli. *A*, temporal waveforms and spectrograms of a 1.2-s partial sentence, “The stranger seemed” (leftmost), a 1.2-s segment of reversed multi-talker noise (RMN; middle), and a 1.2-s segment of fake-formant noise (FFN; rightmost). *B*, spectrograms of

the vocoded speech with different numbers of channels and different vocoder types. In the 4-channel condition, the 4th frequency channel, CF₄, was set as 10 kHz, and thus located beyond the range of the frequency axis

$24.7 + 0.1 \cdot CF$ (Hz), according to the ERB of human listeners with normal hearing (Glasberg and Moore 1990). For a large number of frequency channels (e.g., 22), the spectral information was continuous along the frequency axis (Fig. 1(B), leftmost). The details of the spectrum were smeared due to channel interaction, but some general features were retained. With fewer channels (e.g., 4), the ERB vocoder had spectral holes in between the channels (Fig. 1(B), middle). Therefore, we also tested a “continuous vocoder,” in which the cutoff frequencies of neighboring channels were continuous. For example, the cutoff frequencies were 326 and 767 Hz for CF₁ = 500 Hz, and 767 and 4805 Hz for CF₂ = 1357 Hz. The cutoff frequencies were evenly spaced in the logarithmic scale. Figure 1(B) (rightmost) shows the spectrogram of the 4-channel continuous vocoder. This procedure was most similar to what was used in previous studies (Shannon et al. 1995; Loizou et al. 1999). Last, we tested the “tone vocoder” with pure tones centering at each CF for the ASR.

Human Subjects

The experimental protocol was approved by the Institutional Review Board of Saint Louis University. A total of 25 human subjects (21 to 27 years old, 11

female and 14 male college students) participated in multiple conditions, with each condition containing five subjects. We frequently recruited new subjects for new conditions because each sentence could only be presented to each subject once. All listeners passed a screening test, judged by pure tone thresholds of ≤ 20 dB HL at 0.25, 0.5, 1, 2, and 4 kHz.

Dragon ASR Software

The ASR software was Dragon NaturallySpeaking Premium (Version 13; Nuance Communications, Inc.). Dragon uses a probabilistic function of a Markov process as a general model (Baker 1975). Its ability to learn and adapt is based on Deep Neural Networks (see Dragon’s website: <https://whatsnext.nuance.com>). It needs training stories to establish a unique user profile for each individual speaker. After the training, it transcribes sentences from the same speaker using the established profile. In the present study, the software was trained with vocoded stories that had been processed the same way as the test sentences. The goal was to train the ASR to recognize speech features (such as compromised spectral and fine-structure cues) that were critical for perceiving vocoded sentences, if possible. In two conditions (Fig. 5(A)), we also trained the ASR with

intact/non-vocoded stories, but did not observe a difference in performance.

The audio input, including originally recorded and/or vocoded sentences or training stories, was digitally fed into the ASR without going through speakers or microphones. For each testing condition, a new user profile was created by training the ASR with a 10-min story (*President Kennedy's Inaugural Address*, selected from a list of stories provided by Dragon). The ASR compared the audio of the story with the correct text to establish a user profile. Next, the original or vocoded sentences, with or without background noise, and identical to what had been used in the behavioral studies, were transcribed by the ASR.

Once the sentences had been transcribed, we graded the keywords in the same way as we graded the human results with one exception. Because homophone detection was always a problem for speech-recognition software, we considered it correct as long as the right pronunciation was produced. This issue rarely occurred with the human subjects except in the worst condition (i.e., four channels with background noise) when only broken sentences could be perceived; in that case, we also accepted homophones.

Experimental Conditions

Several common conditions were applied to the human listeners and ASR software in the same manner. Because previous studies have intensively explored the human behavior, we tested more conditions with the ASR and compared performance with previous human studies (in Discussion). Table 1 summarizes the conditions based on the vocoder type, CF₁, and the type of background noise.

The human subjects completed experiments I-a, I-b, and I-c. The experiments were conducted in the same sound booth used for sentence recordings. Sound was presented to the listeners via headphones (Sennheiser HD 280 PRO). Both signal and noise were identical for

the two ears. On each trial, a new sentence was randomly selected. The listeners wrote down as many keywords as they had detected. The listeners controlled the pace of the sentence presentation.

In experiment I-a, both 22 channels and 4 channels were implemented with the ERB vocoder. The first 100 sentences (i.e., $N = 400$ keywords) were used for testing the signal alone under the 22-channel condition; the second 100 sentences were used for testing the signal alone under the 4-channel condition.

For the white-noise condition, the third and fourth 100 sentences were used for testing the 22- and 4-channel speech, respectively. Here, a broadband noise (0.4–15 kHz, unless otherwise specified) was added to the sentences before being processed by the vocoding algorithm. The noise was 6-s long, with the sentence presentation roughly centered at the time window.

The third and fourth 100 sentences were also used with multi-talker babble with a new set of five listeners (subjects 6–10) recruited. During the 6-s presentation window, eight sentences (four spoken by the male reader, and four by the female reader) were randomly selected from the pool of 400 sentences with staggered timing, so that at a given moment, there were two or three sentences presented simultaneously. To prevent the ASR from transcribing the words contained in the babble, this 6-s multi-talker noise was presented in a time-reversed manner. Therefore, we call it reversed multi-talker noise (RMN). As shown in the temporal waveform, the RMN amplitude fluctuated over time with fewer silent moments compared with the clean speech (Fig. 1(A), *middle*). The spectrogram may have recognizable formants, but the timing of the formants was unsynchronized unlike in the clean speech (Fig. 1(A), *leftmost*). The RMN was added to the sentences before being processed by the vocoding algorithm.

The third and fourth 100 sentences were again tested with fake-formant noise (FFN; Fig. 1(C)) with another set of new subjects (11–15). In the spectrogram of the sample non-vocoded speech shown in Fig. 1(A) (*leftmost*), the formants start around two to three hundred Hz (i.e., the first formant) and expand to several thousand Hz (Gelfer

TABLE 1

Experimental conditions for the human subjects and ASR software

	Human	ASR
Vocoder type	I-a: ERB (22-ch, 4-ch), SNR = 20 dB I-b: Continuous (4-ch), SNR = 20 dB I-c: ERB (22-ch), SNR = 0 dB & 4-BPF, SNR = 0 or 20 dB	II-a: ERB (22-ch, 4-ch), SNR = 20 dB II-b: ERB (22-ch), varying SNR II-c: Continuous, varying # of channels, SNR = 20 dB II-d: Tone, varying # of channels, SNR = 20 dB
Lowest CF	500 Hz (22-ch, 4-ch)	500 Hz (22-ch, 4-ch) 300 Hz (22-ch)
Noise type	No noise White noise FFN RMN	No noise White noise FFN RMN

and Mikos 2005; Schnupp et al. 2012). The FFN stimuli were generated by passing white noise through six narrowband filters whose center frequencies were around 300, 1000, 2000, 3000, 4000, and 5000 Hz with a bandwidth of 100 Hz. This is to simulate the fact that the first formant frequencies for human speech are typically below 500 Hz with higher formants roughly spaced 1 kHz apart (Gelfer and Mikos 2005). We then added randomness to the timing and frequency of every 5000 points (i.e., 113 ms) of the narrowband noise by randomly compressing or stretching the waveform by $\leq 10\%$ (uniform distribution). This simple manipulation created shifts in both frequency and timing to disrupt frequency cues (Fig. 1(A), *rightmost*). FFN was added to the sentences before being processed by the vocoding algorithm.

In experiment I-b, subjects 16–20 repeated the 4-channel conditions with the continuous vocoder. The first 100 sentences were used for the clean speech, the second 100 sentences were used for the white-noise, the third 100 sentences for the FFN, and the fourth 100 sentences for the RMN conditions. Except the vocoder type, everything was the same as described above. We did not test the 22-channel condition with the continuous vocoder since it was similar to the ERB vocoder with 22 channels. In both I-a and I-b, the lowest channel frequency CF_1 was 500 Hz.

The above two conditions used the same high SNR for human subjects that was also used for the ASR. As will be shown later, this high SNR did not create significant errors with the human subjects and 22 vocoding channels. In experiment I-c, additional conditions were performed with subjects 21–25. First, the SNR was lowered to 0 dB (with both speech and noise set as 47 dB SPL) for the 22-channel condition to create enough errors. With the errors, correlations can be made between human and ASR on a word-by-word basis. Next, the subjects were tested with four ERB filters but no vocoding at both 0- and 20-dB SNR. In experiment I-c, the subjects were presented with the 400 sentences as used above, as well as the additional 100 sentences in the pool for the five conditions. For the above conditions, we should also point out that variations in the tested sentences may have contributed to variability in listeners' performance.

The following conditions were applied to the ASR. In experiment II-a, we first repeated the same conditions that had been applied to the human subjects as in I-a using the ERB vocoder. Unlike humans, the ASR can be exposed to the same set of sentences, because whenever a new user profile was created, memories of the previous transcriptions were erased. Therefore, instead of using 100 sentences for the 22-channel and another 100 for the 4-channel conditions, we used the same 200 sentences (i.e., $N = 800$ keywords) for all the conditions. This avoided performance differences that may have been caused

by variations in the sentence sets when comparing between different numbers of channels. After testing with the common conditions, the number of channels and CF_1 were varied to improve the ASR performance. The SNR prior to the vocoding process was fixed at 20 dB. The SNR after vocoding will be described in details in the following subsection. In experiment II-b, SNRs were systematically varied for the ERB vocoder. Since the ASR failed to be trained with the 4-channel condition using any vocoder, the number of channels was fixed at 22 in II-b. (And with 22 channels, the ERB and continuous vocoders behave similarly.) For each condition, the SNR was varied so that it covered the range from 0% to its maximum performance.

In experiment II-c, the SNR was fixed at 20 dB for the continuous vocoder, which was located near the center of the dynamic range revealed in II-b. The number of channels was then varied, i.e., 4, 8, 16, 22, 32, and 44. In experiment II-d, the procedure was repeated for the tone vocoder. We did not test with the ERB vocoder, because (1) it can be considered as somewhere in between the continuous and tone vocoders and (2) the filter bandwidths would have to be reduced for larger number of channels to avoid overlap.

Sound Level and SNR

Sound-level calibration with the headphones was performed with a sound-level meter (Extech, MA). The average speech level was approximately 50 dB SPL, varying across sentences. For most of the speech-in-noise conditions (except II-b), the SNR was set as 20 dB prior to vocoding. We used such a high SNR because the ASR showed low performance even at this high level. We also obtained the near-ear SNR using the sound-level meter. For each condition, we performed 8 measurements and computed the averages. We found that the averaged SNRs slightly varied with noise and vocoder types. Using the 22-channel ERB vocoder, the near-ear SNRs were 18.0, 18.7, and 18.0 dB for the white noise, FFN, and RMN, respectively. Using the 4-channel ERB vocoder, the near-ear SNRs were 24.8, 23.6, and 20.0 dB for the white noise, FFN, and RMN. Using the 4-channel continuous vocoder, the near-ear SNRs were 19.0, 20.0, and 18.5 dB for the white noise, FFN, and RMN.

Statistical Analysis

Given a correct percentage, p , the confidence interval, Θ , can be computed as

$$\Theta = p \pm z^* \sqrt{\frac{p(1-p)}{N}} \quad (1)$$

(Lock et al. 2017). For a 99% confidence interval, $z = 2.57$. The value of θ was thus solely determined by the correct rate and number of keywords, N . With large N values ($N = 400$ for human and 800 for the ASR), the θ 's obtained in this study were small.

The study also involves frequent comparisons of two or more proportions (p_1 and p_2) between human and ASR performance or between various conditions. Since the decision variable was either correct or wrong and the errors were not normally distributed, ANOVA was unsuitable here. Instead, a two-proportion z -test (Lock et al. 2017) was calculated as

$$z = \frac{p_1 - p_2}{\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}}, \quad (2)$$

where n_1 and n_2 are the number of keywords, with the null hypothesis $H_0: p_1 = p_2$. When repeated comparisons were made (e.g., p_1 vs. p_2 , and p_2 vs. p_3), a Bonferroni correction was applied to adjust the significance values.

RESULTS

Human Performance with the Narrowband Vocoder

Experiment I-a examined human performance with the ERB vocoder. Figure 2(A) shows the correct rates

of detected keywords with subjects 1–5 when vocoded sentences were presented alone. Here, the 22- and 4-channel conditions each used a different set of 100 sentences ($N = 400$ keywords). With 22 channels, performance ranged 98–100 % for all five subjects (Fig. 2(A), blue). When the channel number was reduced to four, performance decreased to a range of 80–89% (red). The decrease was statistically significant ($p < 0.01$, $z = 5.7$ – 9.4) for every subject; nonetheless, the majority of keywords could be correctly recognized.

Next, vocoded speech was presented with white noise. When the white noise (0.4 to 15 kHz) was added to the sentences prior to the vocoding process, the 22-channel performance obtained with the same five listeners stayed in the range of 99–100% (Fig. 2(B), blue). When the channel number decreased to four, performance dropped to 48–54% (Fig. 2(B), red). For each subject, the 4-channel performance with noise was significantly lower ($p < 0.01$, $z = 17.6$ – 20.7) than the 22-channel performance with noise (Fig. 2(B), blue) as well as the 4-channel performance without noise (Fig. 2(A), red; $p < 0.01$, $z = 8.3$ – 14.7).

In Fig. 2(C), the RMN was added to clean sentences before being processed by the vocoder. With subjects 6–10, the same sentences used in Fig. 2(B) were tested. In contrast to the performance in the white-noise condition (Fig. 2(B)), the five new subjects showed highly variable performance with the RMN (Fig. 2(C)) especially under the 4-channel condition. The averaged subject performance with RMN and 4-channel speech

FIG. 2. Experiment I-a with human subjects using the ERB vocoder. A, intelligibility of human subjects for vocoded sentences without background noise. B–D, intelligibility of human subjects for vocoded sentences with white noise (B), RMN (C), and FFN (D). *The 4-channel performance was significantly lower than the 22-channel performance ($p < 0.01$, z test, $N = 400$). Error bars are 95%

confidence intervals. SNR = 20 dB. (Using the 22-channel ERB vocoder, the near-ear SNRs were 18.0, 18.7, and 18.0 dB for the white noise, FFN, and RMN, respectively. Using the 4-channel ERB vocoder, the near-ear SNRs were 24.8, 23.6, and 20.0 dB for the white noise, FFN, and RMN.)

(Fig. 2(C), red; mean = 50%) was not significantly different from the averaged performance with the white noise (Fig. 2(B), red; mean = 51%).

The last condition in I-a was obtained with subjects 11–15 using the FFN. In contrast to the RMN result, the averaged 4-channel performance with the FFN (Fig. 2(D), red; mean = 72%) was significantly better than the 4-channel white-noise (Fig. 2(B), red; $p < 0.01$, $z = 6.3$) and RMN performances (Fig. 2(C), red; $p < 0.01$, $z = 6.7$), although it was still significantly lower than performance when no noise was present (Fig. 2(A), red; $p < 0.01$, $z = 5.1$). A Bonferroni correction was applied to all multiple comparisons.

Human Performance with the Continuous Vocoder

As shown in Fig. 1(B), the ERB vocoder with 4 channels had missing spectral information. Therefore, we repeated the above conditions using the continuous vocoder which has much broader bandwidths. Figure 3 shows performance of a new set of subjects on the sentence recognition for vocoded speech alone, as well as the vocoded speech with the three types of noise. Similar to the observation using the ERB vocoder, the white-noise performance was the lowest (cyan). It was significantly lower ($p < 0.01$, $z = 2.7$ – 8.9) than any other condition except for the last subject.

When comparing across the two vocoders in I-a and I-b, the average of the white-noise performance was higher for the continuous vocoder (Fig. 3, cyan) than the ERB vocoder (Fig. 2(B), red) with a moderate significance level ($p < 0.05$, $z = 2.3$). The average of the RMN performance was also higher for the continuous vocoder (71%; Fig. 3, yellow) than the ERB vocoder (58%; Fig. 2B, red) with a high significance level ($p <$

0.01, $z = 3.8$). For FFN and the no-noise conditions, the continuous vocoder had lower performances ($p < 0.05$, $z = 2.5$ – 8.2). It was unclear why the no-noise performance using the continuous vocoder was low (66%), sometimes lower than the RMN and FFN performances with the same vocoder (Fig. 3). Note that the listeners performing the ERB-vocoding condition had been trained with 22 and 4 channels, while the listeners performing the continuous-vocoding condition were only tested with 4 channels. The amount of training may have influenced the performance (Swaminathan et al. 2014).

Human Performance Under Additional Conditions

As shown above, the human performance with 22 channels and SNR = 20 dB reached a ceiling level. It was unclear whether the observations with the three types of noise and four channels can be generalized for 22 channels at a lower SNR. Figure 4(A) shows the performance when SNR was lowered to 0 dB with five new subjects in experiment I-c. Except for one subject (subject 24), performance with the FFN was always significantly better than the other two noise types ($p < 0.01$, $z = 3.6$ – 10.7). On average, white noise performance was the lowest.

Figure 4(B) shows the subject performance with four ERB filters but no vocoding at 20-or 0-dB SNR in white noise. Recall that the subjects showed low performance with the 4-channel ERB vocoder in white noise for SNR = 20 dB (Fig. 2(B)), and the ERB filters created spectral holes (Fig. 1(B)). Here, the high performance with the bandpass filters but no vocoding (Fig. 4(B), blue) indicates that the lower performance observed earlier was not due to spectral holes but was a result of noise vocoding. Meanwhile,

FIG. 3. Experiment I-b with human subjects using the continuous vocoder and 4 channels. The red horizontal lines mark statistical differences between the white-noise condition and any other condition ($p < 0.01$, z test). $N = 400$. SNR = 20 dB. (Using the 4-channel continuous vocoder, the near-ear SNRs were 19.0, 20.0, and 18.5 dB for the white noise, FFN, and RMN.)

FIG. 4. Experiment I-c with human subjects under additional conditions. *A*, the 22-channel ERB vocoder was tested with SNR lowered to 0 dB for the three types of noise. The red horizontal lines mark statistical differences between conditions ($p < 0.01$, z test). $N =$

400. *B*, only bandpass filtering but no vocoding was implemented for the 4-channel ERB condition in white noise for two SNR values. A significant reduction was observed for all five subjects with reduced SNRs ($p < 0.01$, z test). $N = 400$

lowering the SNR to 0 dB reduced the performance significantly ($p < 0.01$, $z = 11.7$ – 29.1).

Basic Performance of the ASR

Fig. 5(A) shows the correct rates of transcribed keywords by the ASR for speech presented alone.

First, non-vocoded sentences were used as a control to evaluate the software and the result was near perfect (Fig. 5(A), black). When vocoded sentences using the ERB vocoder were presented, performance decreased significantly ($p < 0.01$, $z = 19.3$; Fig. 2(D), blue). Only 60% of the 800 keywords were correctly transcribed, which was also significantly lower ($p < 0.01$, $z = 15.8$)

Exp II-a: ASR, ERB Vocoder

FIG. 5. Experiment II-a with the ASR and ERB vocoder: *A* and *B*, speech alone without noise. In *A*, $CF_1 = 500$ Hz. *F*, the ASR software failed to be trained. In *B*, the number of channels were increased to 44 (44-ch) or CF_1 was decreased to 300 Hz (22-ch-low). *C*, speech with white noise. In *A* and *C*, “4-BPF” is the condition when only bandpass filtering but no vocoding was implemented. *D*, the same

manipulations as in *B* were performed with white noise. *E*, with $CF_1 = 300$ Hz, white-noise performance was compared with RMN and FFN performances. *Performance increased or decreased significantly ($p < 0.01$, z test, $N = 800$, with Bonferroni corrections). Error bars are 95% confidence intervals

than the average human performance using 22 channels (Fig. 2(C), blue). For the 4-channel condition, the ASR reported that the vocoded training story was insufficient for establishing a working profile (Fig. 2(D), red). In fact, the 4-channel condition never elicited a non-zero performance in any of the following tests. For example, using the continuous vocoder, the ASR could be trained, but less than ten words have been correctly transcribed out of 200 sentences, and none of them were keywords.

We also tried training the ASR with non-vocoded stories. With 22 channels, performance was 58%. With 4 channels (now that the ASR can always be trained), only 1.5–2% of keywords could be correctly detected. Therefore, the low performance above was not caused by training with vocoded stories. In the following tests, the training story was always processed the same way as the test sentences.

Next, we varied the vocoding parameters to see if ASR performance could be improved. We first increased the number of channels from 22 to 44. To avoid channel overlap (e.g., there would have been a 40-Hz overlap for the first two of the 44 channels positioned at 500 and 536 Hz), we also reduced the bandwidth of each bandpass filter to half of the human ERB. This manipulation should have enhanced the spectral representation of speech sounds. We found that having more channels significantly improved ASR performance for speech alone from 60 to 92% (Fig. 5(B), blue and yellow; $p < 0.01$, $z = 16.0$). We then kept the number of channels as 22 but decreased CF_1 from 500 to 300 Hz. The other CFs were also adjusted so that they remained evenly spaced on a logarithmic scale between 0.3 and 10 kHz. For speech alone, the software performance increased significantly to 75% (Fig. 5(B), magenta; $p < 0.01$, $z = 6.45$), although not as high as the 44-channel result (Fig. 5(B), yellow).

When speech and white noise were presented to the ASR, the correct rate was only 14% using the 22-channel ERB vocoder (Fig. 5(C), blue). Performance was significantly lower ($p < 0.01$, $z = 21.6$) than the 22-channel quiet condition for ASR (Fig. 5(A), blue). Note that, when non-vocoded speech was presented to the ASR, the presence of the white noise did not affect the intelligibility (Fig. 5(C), black). Here, the SNR was 20 dB; we varied the SNR in experiment II-b to see when the noise started affecting the perception of non-vocoded speech. We also tested the ASR with the same four ERB channels as bandpass filters but no vocoding. Performance for speech alone was also near 100% (Fig. 5(A), gray), and performance for speech in white noise was 90% (Fig. 5(C), gray). The result indicates that the failure of the ASR to be trained with the 4-channel vocoded speech was not due to the loss of spectral information, but was due to the loss of fine-

structure information as well as the limited frequency channels.

Although doubling the number of channels improved speech intelligibility in quiet (Fig. 5(B), yellow), in the presence of white noise, the larger number of channels significantly impaired recognition (Fig. 5(D), yellow; $p < 0.01$, $z = 7.2$). To explore this nonlinear effect, the number of channels was systematically varied in experiment II-c. We again kept the number of channels as 22 and decreased CF_1 . When white noise was introduced, we extended the lower cutoff frequency of the noise from 400 to 200 Hz so that the lowest-frequency channel ($CF_1 = 300$ Hz, with a bandwidth of 54.7 Hz) remained masked by noise. The 22-channel performance with lowered frequencies showed a significantly improved correct rate from 14 to 46% (Fig. 5(D), magenta; $p < 0.01$, $z = 14.8$). Combining Fig. 5(B and D), increasing the number of channels only improved the software performance in quiet, whereas extending the lowest channel frequency improved performance for both speech in quiet and in white noise.

Fig. 5(E) shows speech intelligibility of the ASR with the 22-channel vocoded sentences in the presence of the spectrally-colored noise. Because ASR performance with the white noise and $CF_1 = 500$ Hz was low (Fig. 4(C), blue; 14%), we used the 300 Hz condition (Fig. 5(D), magenta; 46 %) as a reference for comparisons. Performance was re-plotted in Fig. 5(E) (magenta). Then, white noise was replaced with the RMN or FFN while everything else remained the same. With the RMN, ASR performance decreased from 46 to 31% (Fig. 5(E), green), which was significantly lower ($p < 0.01$, $z = 14.8$). Interestingly, when the RMN was further replaced with FFN, performance again significantly decreased to 13% ($p < 0.01$, $z = 18.9$; Fig. 5(E), cyan). A Bonferroni correction was applied to all the multiple comparisons. We should add that, similar to results for white noise, when non-vocoded speech was tested with the ASR in the presence of FFN or RMN, a correct rate of $\geq 97\%$ was achieved (not shown). That is, the strong effect of noise presented in this study only applied to vocoded speech. The observation was also unrelated to the recorded SNR variations after vocoding, because the near-ear SNR of the RMN was actually lower than the SNR of the FFN (see “Methods” section).

To better examine the hypothesis that humans and ASR use different detection strategies, we performed an error analysis using correlations between the human and ASR detections on a word-by-word basis. We chose to compare the human 22-channel performance at 0-dB SNR (Fig. 4(A)) with the ASR’s 22-channel performance at 20-dB SNR (Fig. 5(E)), as both conditions generated significant amounts of

errors. For white noise, the Pearson's correlation coefficient between every two human subjects ranged between 0.21 and 0.44, which was significant for a total of 400 keywords (the significance level was $r = 0.13$ for $df = 398$). The correlation between the ASR and any human subject ranged between 0.03 and 0.14. For the RMN, the correlation among humans ranged between 0.12 and 0.23, whereas the correlation between the ASR and human ranged between 0 and 0.18. For the FFN, the correlation among humans ranged between 0.17 and 0.42, whereas the correlation between the ASR and human ranged between 0.12 and 0.28. Although the correlation method was an indirect measure of listening strategies, it seemed very likely that humans and ASR used different strategies.

Effect of SNR

The above ASR performance was obtained with a pre-coding SNR value of 20 dB. Figure 6 shows ASR performance over a large range of SNRs using the ERB vocoder and 22 channels. For non-vocoded speech with white noise, the highest performance was near 100% and the sloping area occurred at small SNRs (Fig. 6, black). When speech was vocoded, the sloping area was shifted > 15 dB towards higher SNRs and the maximum performance never approached 100%.

Similar to what was observed in II-a, the 22-channel condition with $CF_1 = 500$ showed the lowest performance (Fig. 6, blue). Decreasing CF_1 to 300 Hz shifted the sloping area by 5–10 dB and boosted the maximum performance by 16% (magenta). For one condition, we repeated the transcription eight times with the same set of sentences to get an idea of the variability (errorbar). The increase of performance was significant at all common SNRs ($p < 0.01$, $z = 7.0$ –15.6). The 500-Hz slope (blue) was also shallower than all the 300-Hz slopes. When CF_1 was fixed at 300

Hz, adding the FFN (Fig. 6, cyan) had the largest adverse effect on speech perception. RMN performance was significantly better than FFN performance for $SNR \leq 20$ dB ($p < 0.01$, $z = 8.9$ –10.4). The comparison of the RMN condition (green) and white-noise condition (magenta) varied with SNRs.

Effect of the Number of Channels

In II-c and II-d, SNRs were fixed at 20 dB with the number of channels varied for the continuous and tone vocoders. We did not test with the ERB vocoder, since the fixed ERB bandwidths would have generated overlap for a large number of channels. Surprisingly, the tone vocoder showed significantly better performance ($p < 0.01$, $z = 6.2$ –53.3) than the continuous vocoder for every noise type and any number of channels, except the 4-channel condition when all of them showed zero or near-zero performances.

With the tone vocoder, all noise types showed a small non-monotonic behavior with increasing number of channels (Fig. 7, dashed lines). Performance with 44 channels was lower than the peak performance for the white noise (both CF_1 values) and the FFN with a significant level of $p < 0.01$ ($z = 4.0$ –9.6), and lower for the RMN with a significant level of $p < 0.05$ ($z = 2.3$). With the continuous noise vocoder (Fig. 7, solid lines), significant non-monotonicity ($p < 0.01$, $z = 11.1$ –13.7) was observed with the white noise, whereas the FFN and RMN conditions showed monotonic behaviors. For all the numbers of channels except four, RMN performances (green solid lines) were significantly better ($p < 0.01$, $z = 4.7$ –9.1) than FFN performances (cyan solid lines). The same trend could also be observed with the tone vocoder with a lower significance level: RMN performances (green dashed lines) were significantly better ($p < 0.05$, $z = 2.3$ –11.8) than FFN performances (cyan dashed lines) for all but four channels.

FIG. 6. Experiments II-b with ASR and varying SNRs. The 22-channel ERB vocoder was used here. $N = 800$ for each data point. At $SNR = 20$ dB (vertical line), the values are the same as in Fig. 4. The error bar at 20-dB SNR shows the standard deviation of one condition when eight repetitions were made to estimate the variability of the ASR

FIG. 7. Experiments II-c and II-d with ASR and varying number of vocoding channels. Solid lines represent the continuous-noise vocoder, and dashed lines represent the tone vocoder. SNR = 20 dB. $N = 800$ for each data point

DISCUSSION

Speech Cues and Vocoding

Major acoustic features of speech include the spectral profile, temporal envelope, and temporal fine structure. This study compared recognition of vocoded speech obtained with human subjects and an ASR software, Dragon. By using vocoded speech with different vocoders and numbers of channels, temporal fine structure was reduced and the spectral resolution was varied. A focus of this study was to investigate how spectral and temporal envelope cues are influenced by different types of noise profiles and whether humans and ASR behave similarly under the same listening condition. We found different masking effects of the noise profiles for humans and ASR, suggesting different sensitivities to the manipulation of speech cues.

Note that the human experiments suffered from several potential limitations. First, the noise types were not equivalently spectrally shaped. Thus, it may not be easy to discriminate masking differences caused by fluctuations in temporal envelopes via dip listening vs. differences in the energetic profile of long-term spectra. Second, the same sentences were used to test the same conditions rather than randomizing or counter-balancing across conditions. Third, different subjects were tested on different conditions. Using fewer sentences per condition with more subjects to increase power may have been a better design.

ASR has been used previously to evaluate vocoded speech and cochlear-implant processing (Liu and Fu 2007; Do et al. 2010; Rader et al. 2015). Our results suggest that this type of evaluation can be problematic because humans and ASR apparently rely on different speech features even under the same listening condition. However, we should be cautious when interpreting the ASR results. First, vocoded-speech recognition may depend on speech materials, such as the IEEE vs. HINT sentences (Zeng et al. 2005). Second, the ASR was trained with normal speech

before it left the factory. Better performance may be achieved with extensive training using vocoded speech. Although we did not find an effect when the 10-min training story was vocoded vs. non-vocoded, more extensive training may make a difference. Last, some of our findings may only apply to the Dragon software and may fail to generalize to all ASR algorithms.

Nevertheless, there are commonly used acoustic cues by most of the speech-recognition algorithms, such as the short-time spectrum, formants, and zero-crossings (Reddy 1976). Since vocoding replaces the fine structure with noise or tone carriers, zero-crossings related to the speech signal are mostly disrupted. Therefore, we could assume that spectral cues dominate the ASR processing for vocoded speech perception. However, the story is not that simple as indicated in the next subsection.

Number of Channels and Speech Cues

The fundamental frequencies for men and women are typically in the range of 100 to 221 Hz (Baken and Orlikoff 2000). The first formant frequencies are below 500 Hz with higher formants roughly spaced 1 kHz apart (Gelfer and Mikos 2005). Formant information of vocoded speech can be extracted by comparing the relative amplitude between adjacent frequency channels (Roberts et al. 2011). In the present study, the spectral profile of speech can be largely represented with 22 channels (CF = 500, 577, 665 Hz, etc.) to resolve most formant frequencies and spacing, except for the first formant. With four channels (CF = 500, 1357, 3684, and 10000 Hz), formant frequencies cannot be resolved from the vocoded speech.

When vocoded speech was presented alone using the ERB vocoder, human listeners showed near-perfect speech recognition with 22 vocoding channels and a SNR of 20 dB. The listeners detected 80–89% words with the 4-channel ERB vocoder, slightly lower

than what was previously reported with continuous vocoders (Shannon et al. 1995; Loizou et al. 1999; Friesen et al. 2001). Using the 4-channel continuous vocoder, performance range was 62–71% (the effect of vocoder type will be described in a following subsection). Our lower performance was most likely caused by setting the lowest channel frequency to 500 Hz, much higher than what was previously used.

In contrast, the ASR cannot be trained to function under any 4-channel vocoding condition even without noise. Training the ASR with non-vocoded stories only generated near-zero performance. At first glance, this seems to agree with the fact that the algorithm relies heavily on spectral cues. However, when bandpass-filtered speech using the same four ERB filters without vocoding was presented, the ASR showed high performances even in the presence of white noise (Fig. 5(A and C), gray). Therefore, we conclude that missing fine-structure information due to vocoding and failing to resolve the formants contributed to the poor performance of the ASR more than the spectral holes. A similar observation was made with the human subjects using 4 channels for SNR = 20 dB in white noise (Fig. 4(B), blue). That is, the previously observed low performance with the human subjects using 4 channels (Fig. 2(B)) was due to noise vocoding rather than spectral holes. In addition, poor speech segmentation with noise carriers may also play an important role in the ASR's low performance (discussed later).

The extraction of silence, voicing, and stress by automated algorithms is often based on energy and fundamental frequencies (perceived as pitch) of speech (Reddy 1976). For the ASR, when we decreased CF_1 from 500 to 300 Hz, an increased performance was observed for both speech in quiet and in white noise with 16- and 22-channel noise vocoders (Figs. 5 and 7). Because the 300-Hz filter would allow the first formant (< 500 Hz) and a small amount of pitch information (100–221 Hz) to come through, better feature extraction and speech segmentation could be achieved.

The above discussions focused on channel numbers no larger than 22. The non-monotonic effect with larger number of channels will be discussed with respect to different vocoder types as follows.

SNR and Noise Profiles

Three types of background noise were tested with vocoded speech. White noise had relatively flat temporal and spectral properties. The FFN had fluctuating spectral distributions; there were also some temporal-envelope changes within each bandpass filter because even with the narrow ERB filters ($ERB = 24.7 + 0.1 \cdot CF$), the frequency shift of up to 10% was within the filter bandwidth. The RMN had large temporal-envelope and spectral fluctuations.

With 22 vocoding channels and a SNR of 20 dB, the perception of vocoded speech in white noise or spectrally colored noise by human listeners remained near perfect, similar to previous studies on vocoded speech perception in multiple types of noise (Eisenberg et al. 2000; Friesen et al. 2001; Shannon et al. 2004). With four channels, adding white noise decreased human performance to 51% with the ERB vocoder and 52 % with the continuous vocoder. One possibility is that a better spectral profile was required to recognize vocoded speech in noise, as in the 22-channel condition. If that were the case, replacing the white noise with FFN, which was intended to disrupt spectral cues, would have made human performance worse. Instead, average correct rates of 70% and 71% were achieved using the 4-channel ERB and continuous vocoders, respectively, for the FFN. The result cannot be explained by variations of the near-ear ERBs (see “Methods” section). Similarly, when the SNR was lowered to 0 dB, all but one human subjects showed significantly higher performance with the FFN using 22 channels. Those observations indicate that the listeners did not rely on spectral cues for vocoded-speech recognition in noise.

The listeners must have adopted certain strategies to counteract the masking effect of spectrally colored noise. With time-varying noise, one common strategy is “listening in dips” (Cooke 2006) in which the brain takes advantage of the moments that happen to have large signal-to-noise ratios. Because white noise has relatively constant temporal envelopes and evenly distributed energy across frequencies, this strategy does not apply. For the FFN and RMN conditions, it is possible that tonotopically arranged neurons in the brain can perform the dip-listening strategy within limited frequency bands. Another explanation for the better performance with FFN was related to the phenomenon called “streaming” (Bregman and Pinker 1978). Streaming refers to the perception that connects the sound of changing frequencies over time generated by a single source. The smaller the variation of frequencies, the higher the chance of perceiving the sound as a single object. With up to 10% frequency variations, the FFN may be streamed into separate sources. Its interfering effect may thus be reduced for a release from masking to occur. However, this is only our speculations. To explore these speculations, future experiments that directly test humans' segregation abilities under the same sound condition should be explored. For instance, listeners may be asked to report whether target speech is present, and performance can be compared with the ASR working at an algorithmic level.

Most cross-subject behaviors we recorded were similar for a given listening condition, especially at high SNRs. One exception is that, when the back-

ground noise was the time-reversed multi-talker babble, listeners exhibited quite variable abilities in overcoming the noise (Figs. 2 and 3) with 4 channels. Some listeners certainly had better abilities to ignore the RMN than other listeners, which may reflect their abilities in utilizing the dip-listening strategy, since the RMN had time-varying amplitudes. When SNR was lowered to 0 dB (Fig. 4), the inter-subject variability was larger, which was likely a result of various signal-in-noise detection thresholds across listeners.

At a 20-dB SNR, the ASR showed the highest correct rate with the white noise among all three types of noise, with the worst performance being the FFN. This is reasonable since, when the noise level was the same, the white noise least disturbed spectral cues with energy evenly distributed across frequency channels. The FFN had clear frequency stripes that could be mistaken as formants. Meanwhile, the ASR is unable to perform “streaming” as humans do, which was the rationale for using the FFN in this study.

However, we should point out that, because the slopes of the psychometric functions with respect to SNR were steep and noisy (Fig. 6) with $CF_1 = 300$ Hz, the effect of noise profile should be taken with caution. In addition, the above-mentioned noise effect only applied to vocoded speech that had compromised fine-structure information. When non-vocoded speech was presented to the ASR in the presence of white noise, FFN, or RMN (not shown), the correct rates were close to 100%. The interfering effect of the noise was a combined result of noise vocoding and external noise.

With varying SNRs, the dynamic ranges and slopes of the ASR psychometrical functions changed for different CF_1 and noise conditions (Fig. 6). When $CF_1 = 500$ Hz, the slope was much shallower than the slopes with $CF_1 = 300$ Hz. The sloping range for non-vocoded speech occurred mostly within a SNR of -10 to 10 dB. For the three types of noise with $CF_1 = 300$ Hz, the sloping ranges were around 10 to 30 dB. The range was further increased to 20 – 50 dB with $CF_1 = 500$ Hz. In previous human studies, human subjects usually had their sloping range around -10 to 10 (Friesen et al. 2001) or 0 to 20 dB SNR (Whitmal et al. 2007) with tone or noise vocoders. In short, when compared with human performance, the ASR showed a shift of sloping range to higher SNRs and reduced maximum performances using noise vocoders (Fig. 6). The ability of the ASR to recognize speech with compromised fine-structure information was highly inferior to human performance.

Number of Channels and Vocoder Types

For human listeners, previous studies usually kept the number of channels low (≤ 22). An exception is the

chimaeric-sound study by Smith et al. (2002) that extended the number to 32 for speech perception and 64 for melody recognition. In that study, when envelope cues were used, no non-monotonic behaviors were observed. This is probably because human listeners are more efficient in performing speech segmentation (such as using context or experience) than the ASR. Therefore, our human listeners would most likely show no non-monotonic behavior with increased number of channels, unlike the ASR, assuming the 32-channel result applies to our experimental paradigm. However, caution should be taken when comparing that study with the present study because of differences in the methodologies.

Previous noise-vocoding studies mostly used the continuous vocoder (e.g., the same cutoff frequency in two adjacent passbands) (Shannon et al. 1995; Loizou et al. 1999; Qin and Oxenham 2003). We did most of the conditions using the ERB filters and also repeated some of the conditions using the continuous filters. To test how well these two vocoding approaches represent the original speech, we measured the cross-correlation between non-vocoded and vocoded speech waveforms in very narrow bands (fixed at 50 Hz) with the same channels and CFs as in the vocoding procedure. We found that the average correlation value over multiple sentences between the 22-channel vocoded speech (without noise) and non-vocoded speech approximately was 0.70 for both vocoders. The average correlation value using the 4-channel ERB vocoder was 0.67, while the average correlation using the 4-channel continuous vocoder was only 0.4. In other words, although broadening the vocoding bandwidths can cover a larger frequency spectrum, temporal envelopes (when analyzed in narrow frequency bands) are compromised with broadened vocoding bandwidths. In this sense, we obtained lower performances from our human listeners using the 4-channel ERB vocoded speech, not because the spectral holes but likely because missing the speech information below 500 Hz. It also agrees with the fact that human performance with the 4-channel continuous vocoder was worse than performance with the ERB vocoder for speech alone.

Regarding the tone vocoder, some previous human studies showed similar speech perception performance using tone and noise vocoders (Dorman et al. 1997a), whereas others showed better performance with the tone vocoder (Whitmal et al. 2007). Even for those showing better performance, the difference in percent correct was no more than 20%. Previous studies suggest that better performance could be caused by a more faithful representation of speech envelope by tone carriers than noise carriers, since noise carriers naturally have fluctuations that can be mistaken as the speech envelopes (Whitmal et al. 2007).

In the present study, the ASR showed an improvement of performance as large as 74% with tone vocoders (Fig. 7). Apart from the above-mentioned argument, another possible explanation may relate to speech segmentation. ASR algorithms highly depend on segmentations of phonemic boundaries and word boundaries (Reddy 1976). It is critical to identify basic speech features, such as silence, voicing, and stress, before each word can be recognized. Because noise vocoding replaced the speech carriers with noise, it could be difficult to segment the speech when background noise was filling the silent gaps, i.e., to separate the noise carriers from background noise. The argument is further supported by the fact that, when spectral cues were improved by increasing the number of channels, the ASR showed a highly non-monotonic psychometric function for the white noise, but not for the RMN or FFN with the continuous-noise vocoder (Fig. 7). Specifically, when the ASR could separate the noise carriers (created by filtering white noise) from the colored noise, having better spectral resolution led to better recognition. When the ASR could not easily separate the noise carriers from the white-noise background, increasing the number of channels perhaps led to decreased filter bandwidths and longer ringing effects. As the filters get narrower, the inherent fluctuations in the noise get larger. Those effects could all play a role in limiting speech segregation. (Note that these are speculations, and we do not have clear evidence to support them.) In contrast, it is easier to separate speech with tone carriers embedded in noisy backgrounds, and thus leading to higher performance for the tone vocoder. To test the hypothesis that poor speech segmentation is the cause of the low ASR performance in noise, future studies can add extra silent periods between adjacent words (i.e., use gated noise around each word and increase the gaps between words) to see if ASR performance can be improved.

Another explanation is that the underlying phoneme identification model of the ASR likely contains a pitch detection algorithm that would be disrupted by noise vocoding. Errors introduced by noise vocoding disrupting pitch detection would lead to errors in terms of voiced/unvoiced consonant classification. Such disruption would be less for tone vocoders since they encode periodicity more reliably than noise voicing. Because the ASR was treated as a black box in this study, we cannot comment on whether segmentation errors or voiced/unvoiced errors occurred more frequently. In future studies, it is necessary to open the algorithm to determine if errors indeed occur during target-speech segmentation as opposed to phoneme confusions.

ACKNOWLEDGMENTS

We thank L. Carney for offering significant input on the manuscript. We thank L. Calandruccio for providing the sentences. We also thank the reviewers for providing tremendous help and insights to the manuscript.

COMPLIANCE WITH ETHICAL STANDARDS

Conflict of Interest The authors declare that they have no conflict of interest.

REFERENCES

- ALI A (1999) Auditory-based acoustic-phonetic signal processing for robust continuous speech recognition. PhD thesis, University of Pennsylvania
- ALLEN JB (1995) How do humans process and recognize speech? In: Ramachandran RP, Mammone RJ (eds) *Modern methods of speech processing*. Springer US, Boston, pp 251–275
- ARDOINT M, AGUS T, SHEFT S, LORENZI C (2011) Importance of temporal-envelope speech cues in different spectral regions. *J Acoust Soc Am* 130:EL115–EL121
- ATAL BS, HANAUER SL (1971) Speech analysis and synthesis by linear prediction of the speech wave. *J Acoust Soc Am* 50:637–655
- BAKEN R, ORLIKOFF R (2000) *Clinical measurement of speech and voice*, 2nd edn. Singular Publishing Group Thomson Learning, San Diego
- BAKER J (1975) The DRAGON system—an overview. *IEEE Transactions on Acoustics, Speech, and Signal Processing* 23:24–29
- BEEKHUIZEN B, BOD R, ZUIDEMA W (2013) Three design principles of language: the search for parsimony in redundancy. *Lang Speech* 56:265–290
- BREGMAN AS, PINKER S (1978) Auditory streaming and the building of timbre. *Can J Psychol* 32:19–31
- CALANDRUCCIO L, SMILJANIC R (2012) New sentence recognition materials developed using a basic non-native English lexicon. *J Speech Lang Hear Res* 55:1342–1355
- COOKE M (2006) A glimpsing model of speech perception in noise. *J Acoust Soc Am* 119:1562–1573
- DAVIS SB, MERMELSTEIN P (1990) Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. In: Alex W, Kai-Fu L (eds) *Readings in speech recognition*. Morgan Kaufmann Publishers Inc., pp 65–74
- DO CT, PASTOR D, GOALIC A (2010) On the recognition of cochlear implant-like spectrally reduced speech with MFCC and HMM-based ASR. *IEEE Transactions on Audio, Speech, and Language Processing* 18:1065–1068
- DORMAN MF, LOIZOU PC, RAINEY D (1997A) 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
- DORMAN MF, LOIZOU PC, RAINEY D (1997B) Simulating the effect of cochlear-implant electrode insertion depth on speech understanding. *J Acoust Soc Am* 102:2993–2996
- EISENBERG LS, SHANNON RV, MARTINEZ AS, WYGONSKI J, BOOTHROYD A (2000) Speech recognition with reduced spectral cues as a function of age. *J Acoust Soc Am* 107:2704–2710
- FRIESEN LM, SHANNON RV, BASKENT D, 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
- GELFER MP, MIKOS VA (2005) The relative contributions of speaking fundamental frequency and formant frequencies to gender identification based on isolated vowels. *J Voice* 19:544–554
- GHITZA O (2001) On the upper cutoff frequency of the auditory critical-band envelope detectors in the context of speech perception. *J Acoust Soc Am* 110:1628–1640
- GILBERT G, LORENZI C (2006) The ability of listeners to use recovered envelope cues from speech fine structure. *J Acoust Soc Am* 119:2438–2444
- GLASBERG BR, MOORE BC (1990) Derivation of auditory filter shapes from notched-noise data. *Hear Res* 47:103–138
- HEINZ MG, SWAMINATHAN J (2009) Quantifying envelope and fine-structure coding in auditory nerve responses to chimaeric speech. *J Assoc Res Otolaryngol* 10:407–423
- JUNEJA A, 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
- LIU C, FU QJ (2007) Estimation of vowel recognition with cochlear implant simulations. *IEEE Trans Biomed Eng* 54:74–81
- LOCK RH, LOCK PF, MORGAN KL, LOCK EF, LOCK DF (2017) *Statistics: unlocking the power of data*, 2nd edn. Wiley, NJ
- LOIZOU PC, DORMAN M, TU Z (1999) On the number of channels needed to understand speech. *J Acoust Soc Am* 106:2097–2103
- MAKHOUL J (1975) Linear prediction: a tutorial review. *Proc IEEE* 63:561–580
- MAO J, CARNEY LH (2014) Binaural detection with narrowband and wideband reproducible noise maskers. IV. Models using interaural time, level, and envelope differences. *J Acoust Soc Am* 135:824–837
- MAO J, CARNEY LH (2015) Tone-in-noise detection using envelope cues: comparison of signal-processing-based and physiological models. *J Assoc Res Otolaryngol* 16:121–133
- MAO J, KOCH KJ, DOHERTY KA, CARNEY LH (2015) Cues for diotic and dichotic detection of a 500-Hz tone in noise vary with hearing loss. *J Assoc Res Otolaryngol* 16:507–521
- QIN MK, OXENHAM AJ (2003) Effects of simulated cochlear-implant processing on speech reception in fluctuating maskers. *J Acoust Soc Am* 114:446–454
- RADER T, ADEL Y, FASTL H, BAUMANN U (2015) Speech perception with combined electric-acoustic stimulation: a simulation and model comparison. *Ear Hear* 36:e314–e325
- RAO A, KUMARESAN R (2000) On decomposing speech into modulated components. *IEEE Trans Speech Audio Process* 8:240–254
- REDDY DR (1976) Speech recognition by machine: a review. *Proc IEEE* 64:501–531
- ROBERTS B, SUMMERS RJ, BAILEY PJ (2011) The intelligibility of noise-vocoded speech: spectral information available from across-channel comparison of amplitude envelopes. *Proc Biol Sci* 278:1595–1600
- ROSEN S (1992) Temporal information in speech: acoustic, auditory and linguistic aspects. *Philos Trans R Soc Lond B Biol Sci* 336:367–373
- SCHNUPP J, NELKEN I, KING AJ (2012) *Auditory neuroscience: making sense of sound*. MIT Press, Cambridge
- SHANNON RV, ZENG FG, KAMATH V, WYGONSKI J, EKELID M (1995) Speech recognition with primarily temporal cues. *Science* 270:303–304
- SHANNON RV, FU QJ, GALVIN J, 3RD (2004) The number of spectral channels required for speech recognition depends on the difficulty of the listening situation. *Acta Otolaryngol Suppl*:50–54
- SMITH ZM, DELGUTTE B, OXENHAM AJ (2002) Chimaeric sounds reveal dichotomies in auditory perception. *Nature* 416:87–90
- STILP CE (2011) The redundancy of phonemes in sentential context. *J Acoust Soc Am* 130:EL323–EL328
- SWAMINATHAN J, REED CM, DESLOGE JG, BRAIDA LD, DELHORNE LA (2014) Consonant identification using temporal fine structure and recovered envelope cues. *J Acoust Soc Am* 135:2078–2090
- WHITMAL NA, POISSANT SF, FREYMAN RL, HELFER KS (2007) Speech intelligibility in cochlear implant simulations: effects of carrier type, interfering noise, and subject experience. *J Acoust Soc Am* 122:2376–2388
- ZENG FG, NIE K, LIU S, STICKNEY G, DEL RIO E, KONG YY, CHEN H (2004) On the dichotomy in auditory perception between temporal envelope and fine structure cues. *J Acoust Soc Am* 116:1351–1354
- ZENG FG, NIE K, STICKNEY GS, KONG YY, VONGPHOE M, BHARGAVE A, WEI C, CAO K (2005) Speech recognition with amplitude and frequency modulations. *Proc Natl Acad Sci U S A* 102:2293–2298