The role of periodicity in perceiving speech in quiet and in background noise

Kurt Steinmetzger and Stuart Rosen

Citation: The Journal of the Acoustical Society of America **138**, 3586 (2015); doi: 10.1121/1.4936945
View online: http://dx.doi.org/10.1121/1.4936945
View Table of Contents: http://asa.scitation.org/toc/jas/138/6
Published by the Acoustical Society of America

The role of periodicity in perceiving speech in quiet and in background noise

Kurt Steinmetzger and Stuart Rosen[a)]
Speech, Hearing and Phonetic Sciences, University College London, Chandler House, 2 Wakefield Street, London WC1N 1PF, United Kingdom

(Received 27 January 2015; revised 1 November 2015; accepted 15 November 2015; published online 11 December 2015)

The ability of normal-hearing listeners to perceive sentences in quiet and in background noise was investigated in a variety of conditions mixing the presence and absence of periodicity (i.e., voicing) in both target and masker. Experiment 1 showed that in quiet, aperiodic noise-vocoded speech and speech with a natural amount of periodicity were equally intelligible, while fully periodic speech was much harder to understand. In Experiments 2 and 3, speech reception thresholds for these targets were measured in the presence of four different maskers: speech-shaped noise, harmonic complexes with a dynamically varying *F*0 contour, and 10 Hz amplitude-modulated versions of both. For experiment 2, results of experiment 1 were used to identify conditions with equal intelligibility in quiet, while in experiment 3 target intelligibility in quiet was near ceiling. In the presence of a masker, periodicity in the target speech mattered little, but listeners strongly benefited from periodicity in the masker. Substantial fluctuating-masker benefits required the target speech to be almost perfectly intelligible in quiet. In summary, results suggest that the ability to exploit periodicity cues may be an even more important factor when attempting to understand speech embedded in noise than the ability to benefit from masker fluctuations.
[http://dx.doi.org/10.1121/1.4936945]

[EB] Pages: 3586–3599

I. INTRODUCTION

The production of any speech sound can be described by the interplay of a sound source and a vocal tract filter (e.g., Fant, 1960). Normally, either the periodically vibrating vocal cords (voiced speech) or aperiodic noise arising from constrictions in the vocal tract (voiceless speech) serve as source, although the two may occasionally overlap, such as in voiced fricatives. Clearly, the regular periodic pattern of voiced sounds stands in sharp acoustic contrast to noisy unvoiced sounds, and this contrast is also linguistically relevant since only the complex tones of voiced speech possess a pitch and thus allow the unambiguous signaling of intonation (Rosen, 1992). The component tones of voiced speech sounds stand in a harmonic relation and are not perceived individually. "All components point to a single source and meaning" (Rasch and Plomp, 1999, p. 95) and hence harmonicity can be said to add coherence to a sound stream (e.g., Oxenham, 2008). It thus seems reasonable to posit that periodicity in both target and masker helps to segregate a speech target from a background noise or an interfering talker.

On the other hand, de Cheveigné and colleagues (de Cheveigné *et al.*, 1995; de Cheveigné *et al.*, 1997b) found that listeners benefit from harmonicity in the masker, but not the target speech. In these studies artificial steady-state vowels were used as both target and masker. Inharmonic vowels were much more effective in masking the target vowel than harmonic ones, while harmonicity of the target vowel did not significantly affect performance. The results were taken to show that the auditory system seems to be able to cancel a harmonic masker out of the signal mixture. This so-called harmonic cancellation was also observed when unprocessed complex sentences were used as targets and the harmonicity of complex tone maskers was either blurred by modulating the masker *F*0 or further compromised by additionally reverberating the maskers (Deroche and Culling, 2011). Furthermore, Deroche and colleagues also provided evidence for spectral glimpsing in between resolved masker harmonics as an additional mechanism explaining the masking release found with harmonic complex maskers (Deroche *et al.*, 2014a,b). In sum, these findings emphasize the importance of periodicity in the masker, but not the target speech. However, these studies have computationally manipulated the harmonicity of the materials and so have not investigated the role of periodicity by contrasting voiced and unvoiced sounds as they occur in natural speech.

Although a lot of research in recent years has been devoted to the study of speech perception in noise and in particular, the ability of listeners to "glimpse" small sections of target speech in the troughs of an amplitude-modulated masker (Miller and Licklider, 1950), the role of periodicity information in this context has not been investigated thoroughly. It has been claimed that the ability to perceive the temporal fine-structure (TFS) in a target speech signal (i.e., any temporal information in speech, including periodicity information, apart from the slower envelope modulations) is essential in order to benefit from the dips of a fluctuating masker (Gnansia *et al.*, 2009; Hopkins and Moore, 2009;

a)Electronic mail: s.rosen@ucl.ac.uk

Hopkins *et al.*, 2008; Lorenzi *et al.*, 2006). However, it is unclear to date whether TFS information plays a special role in glimpsing or is just as important for steady maskers (Moore, 2012).

Generally, normal-hearing listeners have been found to show rather large benefits in response to fluctuating maskers such as amplitude-modulated noise (e.g., Festen and Plomp, 1990; Bacon *et al.*, 1998; Nelson *et al.*, 2003; Fastl and Zwicker, 2007, p. 352) or interfering talkers (e.g., Festen and Plomp, 1990; Cullington and Zeng, 2008). Studies with hearing-impaired subjects (Festen and Plomp, 1990; Bacon *et al.*, 1998; Peters *et al.*, 1998) or spectrally degraded stimuli (Peters *et al.*, 1998; Oxenham and Simonson, 2009) on the other hand, tend to find reduced fluctuating masker-benefits (FMBs), while studies with cochlear implant (CI) users and CI simulations find hardly any FMB (Nelson *et al.*, 2003; Fastl and Zwicker, 2007, p. 352; Cullington and Zeng, 2008) or even a worsening of performance (referred to as "modulation interference"—Stickney *et al.*, 2004; Kwon *et al.*, 2012).

However, an important confound that has been pointed out by Bernstein and Grant (2009) is that the FMB is generally smaller at higher signal-to-noise ratios (SNRs). Freyman *et al.* (2012) illustrate this point with the typical shape of the psychometric functions (PFs), which are steeper for steady as compared to fluctuating maskers but converge at higher SNRs. Since any form of hearing-impairment or stimulus degradation will generally lead to increased SNRs, the ability to glimpse in these contexts might have been significantly underestimated in previous experiments.

Few studies to date have explicitly investigated the role of periodicity in the perception of speech in noise. Freyman *et al.* (2012) compared unprocessed speech to naturally produced whispered speech and found no substantial differences in terms of the FMB obtained in steady and fluctuating speech-shaped noise, although the intelligibility of whispered speech was much lower. The authors concluded that for normal-hearing listeners, periodicity in the target speech has little effect on the ability to glimpse. However, due to the acoustic distinctiveness of whispered speech, which includes an altered consonant–vowel intensity ratio, it remains unclear whether the role of periodicity is similarly limited in normally articulated speech. Vestergaard and Patterson (2009), using artificially created "whispered" speech, report that only the absence of periodicity cues in both target and masker (i.e., a combination of whispered targets and maskers) negatively affects performance. Third, a study by Rosen *et al.* (2013) has recently compared speech reception thresholds (SRTs) of unprocessed and noise-vocoded target speech obtained in the presence of multi-talker babble, noise-vocoded babble, and speech-modulated noise. The most effective masker was in both cases the one that most closely resembled the target speech, which again argues against the hypothesis that periodicity helps to segregate competing speech signals.

The present study attempted to go beyond previous work by systematically investigating the role of periodicity using normally articulated speech only. Possible confounding factors such as the spectral resolution and intelligibility of the target speech were controlled for and informational masking effects were ruled out by using non-speech maskers only.

The amount of periodicity in the target speech was varied using different types of vocoders. While unvoiced speech can be reproduced adequately using a noise-vocoder that uses noise as source (Shannon *et al.*, 1995), vocoders with periodic sources have been used less often in the literature (Faulkner *et al.*, 2000). However, as originally described by Dudley (1939) and more recently by Loizou (2013, p. 54), voiced speech can be simulated efficiently with a vocoder using a pulse train carrier whose frequency follows the natural *F*0 contour of the original speech. The effects of periodicity in the masker were assessed by comparing aperiodic speech-shaped noise maskers to harmonic complex maskers with dynamically varying *F*0-contours based on real speech.

Experiment 1 tested whether the intelligibility of speech presented in quiet is affected by the amount of periodicity. In Experiments 2 and 3 the amount of periodicity in both target and masker was varied and SRTs were measured in steady and fluctuating maskers. Experiments 2 and 3 differed only regarding the intelligibility of the target speech materials in quiet. Hence, the results can be presented in the same figures.

II. EXPERIMENT 1

A. Short introduction and rationale

Experiment 1 investigated the role of periodicity in the perception of speech in quiet testing conditions by parametrically varying the amount of periodicity in the target speech along with the spectral resolution (i.e., the number of bands in the vocoder).

Aperiodic noise-vocoded speech has been used extensively in simulations of CIs (e.g., Shannon *et al.*, 1995; Fu and Nogaki, 2005; Whitmal *et al.*, 2007) and has become a popular tool for reducing the intelligibility of speech signals in neuroscience (e.g., Scott *et al.*, 2000; Obleser and Weisz, 2012). However, it has never been examined whether the absence of periodicity itself leads to a decrease in intelligibility. More generally, despite its salience it is unclear to date whether periodicity information is a beneficial cue in the absence of competing talkers or maskers.

In addition to completely unvoiced noise-vocoded speech and vocoded speech with a natural mix of voiced and unvoiced sections, the current experiment included completely voiced vocoded speech. The latter condition sounds very unnatural and is expected to be less intelligible in quiet. However, since periodicity is assumed to aid stream segregation, this condition will be of particular interest in the presence of background noise (Experiments 2 and 3). An additional purpose of the current experiment was to identify conditions with similar intelligibility rates across the three processing conditions.

Experiment 1 consisted of 18 processed speech conditions as well as unprocessed speech as an additional condition. Participants were presented with noise-vocoded speech (henceforth referred to as the Nx), Dudley-vocoded speech (Dudley, 1939) with a natural mix of periodicity and

aperiodicity (FxNx), and completely periodic *F*0-vocoded speech (Fx) with an *F*0 contour interpolated through unvoiced segments. These three types of stimuli also varied in the number of frequency bands used in their synthesis (6, 7, 8, 10, 12, or 16), and hence their intelligibility. An example sentence with eight bands for all three processing conditions is shown in Fig. 1, along with the unprocessed version of the same sentence.

B. Methods

1. Subjects

Eleven normal-hearing listeners (six females) were tested. Their ages ranged from 19 to 35 with a mean of 27.3 yrs. All participants were native speakers of British English and had audiometric thresholds of less than 20 dB hearing level (HL) at frequencies between 125 and 8000 Hz.

2. Stimuli

The targets used in this experiment were recordings of the IEEE sentences (Rothauser *et al.*, 1969) spoken by an adult male Southern British English talker with a mean *F*0 of 121.5 Hz that were normalized to a common root-mean-square (rms) level. The IEEE sentence corpus consists of 72 lists with 10 sentences each and is characterized by similar phonetic content across the lists and overall low semantic predictability. Every sentence contains five keywords.

3. Signal processing

All stimulus materials were processed prior to the experiment using a channel vocoder implemented in MATLAB R2012b (Mathworks, Natick, MA). For all three processing conditions (Nx, FxNx, and Fx) the original recordings of the IEEE sentences were first bandpass filtered into 6, 7, 8, 10, 12, or 16 bands using zero-phase-shift sixth-order Butterworth filters. The filter spacing was based on equal basilar membrane distance (Greenwood, 1990) across a frequency range of 0.1–11 kHz. The output of each filter was full-wave rectified and low-pass filtered at 30 Hz (zero-phase-shift fourth-order Butterworth) in order to extract the amplitude envelope. The low cutoff value was chosen in order to ensure that no temporal periodicity cues were present. The final waveforms were low-pass filtered at 10 kHz (sixth-order elliptic).

For the noise-vocoded condition (Nx), the envelope from each band was then multiplied with a wide-band noise carrier. The resulting signal was again bandpass filtered using the same sixth-order Butterworth filters as in the first stage of the process. Before the signal was summed together, the output of each band was adjusted to the same rms level as found in the original bands. For the Dudley-vocoded condition (FxNx), the envelope from each band was multiplied with either a wide-band noise carrier where the original speech was unvoiced, or a pulse train following the natural *F*0 contour when the original speech was voiced.

The *F*0 contours of each sentence were generated using ProsodyPro version 4.3 (Xu, 2013) implemented in PRAAT (Boersma and Weenink, 2013). The *F*0 extraction sampling rate was set to 100 Hz. The results were hand-corrected and the resulting values used to generate the pulse trains for the vocoder software described above.

Based on these pulse files, additional *F*0 contours were created by interpolation through unvoiced sections and

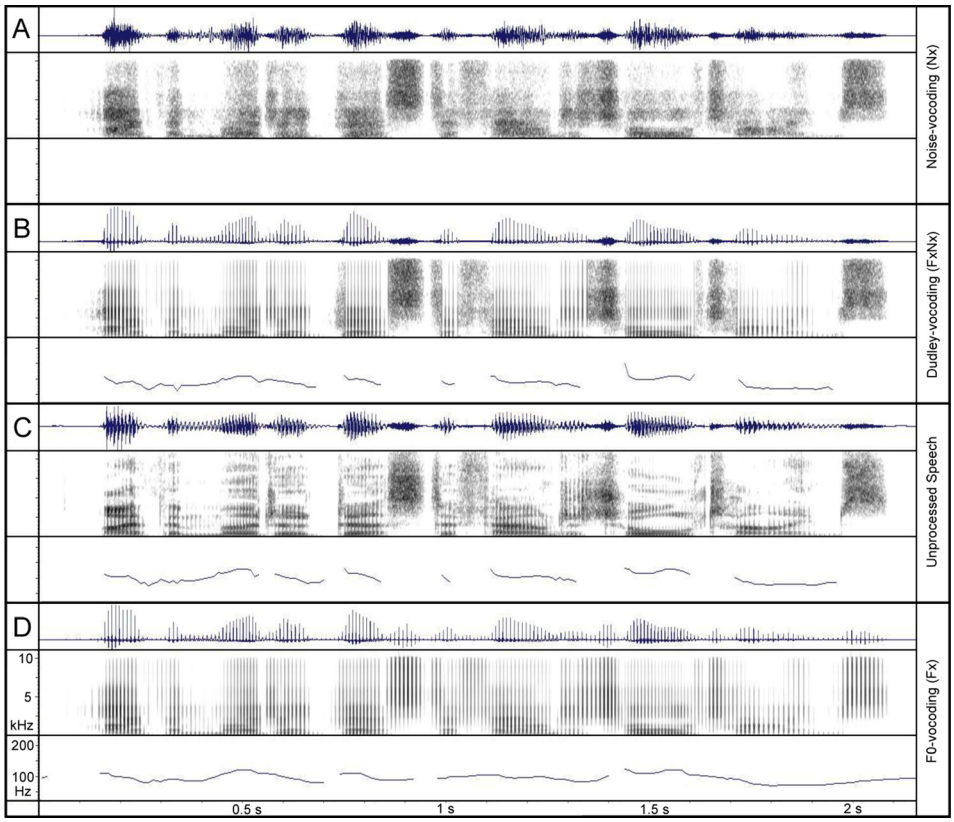

FIG. 1. (Color online) Target speech conditions. Waveforms, wideband spectrograms, and *F*0 contours for one example sentence (either mud or dust are found at all times) processed to have (A) an aperiodic (noise-vocoding, Nx), (B) mixed (Dudley-vocoding, FxNx), or (D) periodic source excitation (F0-vocoding, Fx). Panel (C) shows the unprocessed version of the same sentence for the purpose of comparison. The three processed sentences were all vocoded with eight frequency bands.

periods of silence in order to synthesize fully periodic vocoded speech (Fx). The interpolation was done using piecewise cubic Hermite interpolation in logarithmic frequency. The start and end points of each contour were anchored to the median frequency of the sentence.

4. Procedure

Every participant listened to two full IEEE lists (i.e., 20 sentences) per processing condition and was asked to repeat as many words as possible after every sentence. The verbal responses were logged by the experimenter before the next sentence was played (in terms of which of the roots of the five key words in each sentence were correctly identified, so-called loose key word scoring). No feedback was given following the responses. The presentation and logging of the responses was carried out using locally developed MATLAB software. The experiment consisted of 19 conditions (3 vocoding conditions × 6 degrees of spectral resolution, and 1 additional condition with unprocessed target speech). Hence every participant was presented with 380 sentences in total. The order of the 19 processing conditions was fully randomized using a Latin Square design and the order of the IEEE lists was also randomized. Before being tested the subjects were familiarized with the materials by listening to 2 example sentences of each of the 18 processed conditions. Here every sentence was directly followed by its unprocessed counterpart. The total testing time, including hearing screening and familiarization, was about 1 h and the subjects were allowed to take breaks whenever they wished to. The experiment took place in a double-walled sound-attenuating booth, with the computer signal being fed through the wall onto a separate monitor. The stimuli were converted with 24-bit resolution and a sampling rate of 22.05 kHz using an RME Babyface soundcard (Haimhausen, Germany) and presented over Sennheiser HD650 headphones (Wedemark, Germany) at a level of about 80 dB sound pressure level (SPL) over a frequency range of 75 Hz–10.0 kHz as measured on an artificial ear (type 4153, Brüel & Kjær Sound & Vibration Measurement A/S, Nærum, Denmark).

C. Results and discussion

The proportion correct scores obtained are shown in Fig. 2. The Dudley-vocoded condition (FxNx) with a natural mix of periodicity and aperiodicity led to the highest percentage of correctly repeated key words but is closely followed by the noise-vocoded (Nx) condition, irrespective of the number of frequency bands. Fully periodic *F*0-vocoded speech (Fx) on the other hand was found to result in much lower intelligibility rates, with only 84% correctly repeated key words even with as much spectral detail as 16 frequency bands. Unprocessed speech was found to have an almost perfect intelligibility level with 99.6% correct key words, proving that the IEEE materials as such do not impose excessive memory demands despite their complexity.

The data were analyzed using a generalized linear mixed effects model with a logistic link function that included target periodicity and spectral resolution as fixed factors and subjects as a random factor. The main effects of

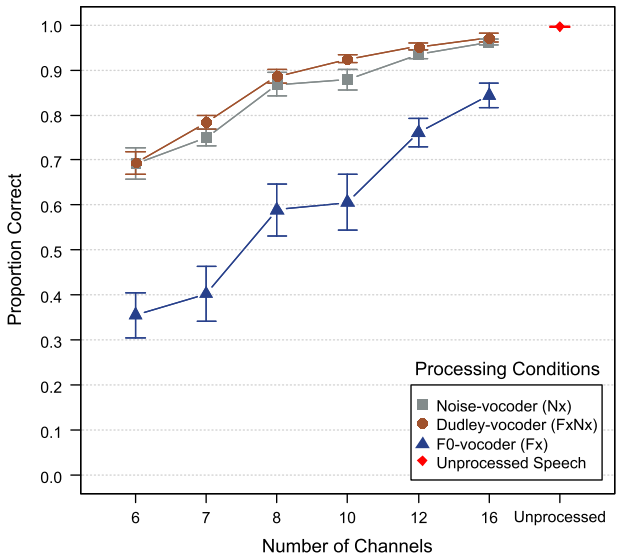

FIG. 2. (Color online) Proportion correct scores in experiment 1 plotted as a function of the number of frequency bands for the three different vocoding conditions: Noise-vocoding (Nx, aperiodic source), Dudley-vocoding (FxNx, mixed source), and *F*0-vocoding (Fx, periodic source). The score for unprocessed speech is included for the purpose of comparison. The error bars show the standard error of the mean.

target periodicity [$F(2,180) = 114.0$, $p < 0.001$] and spectral resolution [$F(5,180) = 113.5$, $p < 0.001$] were found to be highly significant, but there was no interaction of the two [$F(10,180) = 0.5$, $p = 0.89$]. The fixed coefficients furthermore showed that performance with *F*0-vocoded speech (-1.5, $p < 0.001$), but not Dudley-vocoded speech (0.4, $p = 0.24$), was significantly different from performance with noise-vocoded speech.

The fact that performance with noise-vocoded speech was not significantly worse than that with Dudley-vocoded speech suggests that the absence of any periodicity information, and hence intonation, is of minor importance in quiet testing conditions. Although voice pitch information is not essential for understanding English declarative sentences, it is still surprising that a cue as salient as periodicity transmits mostly redundant information. However, despite some important acoustic differences, noise-vocoded speech to some extent resembles whispered speech. Hence, listeners are likely to be at least implicitly familiar with this type of speech. Noise-vocoded speech also enables the listeners to use weaker correlated cues like intensity to distinguish between voiced and unvoiced consonants. Additionally, the spectral shape, which is well coded in a vocoder, gives strong cues to voicing, even in the absence of periodicity. Voiced speech is heavily weighted toward low frequencies, while voiceless excitation is typically weighted to the high frequencies.

The unnatural periodic energy in the *F*0-vocoded condition, especially in the frequency region above 4 kHz, on the other hand, might have substantially interfered with the listener's ability to correctly identify the individual sounds of the presented sentences. Since periodicity is such a dominant cue, weaker cues like intensity differences may not have been noticed. Similarly, for unvoiced fricatives like /s/ and /ʃ/, for example, aperiodic energy at high frequencies is

missing as a cue for identification and replaced by periodic energy in the *F*0-vocoded condition, making the information transmitted contradictory. In addition, listeners are confronted with "false" intonation contours due to the interpolation of the natural *F*0-contours, which is likely to have lowered intelligibility rates even further.

Taken together, the results of experiment 1 show that in quiet testing conditions listeners did not benefit from natural periodicity information, while additional unnatural periodicity cues lead to substantially poorer speech intelligibility rates.

III. EXPERIMENT 2

A. Short introduction and rationale

Experiment 2 presented the three classes of target speech described in experiment 1 in a variety of background noises. The maskers used were either aperiodic speech-shaped noises or fully periodic harmonic complexes with a dynamically varying *F*0 contour (similar to those used in Green and Rosen, 2013). Both types of maskers were presented in a steady or 10 Hz sinusoidally amplitude-modulated version. This design allowed for a systematic variation of periodicity in both target and masker, and also allowed the examination of the role of amplitude fluctuations in the masker.

Performance was assessed via an estimation of the SRT (Plomp and Mimpen, 1979). Importantly, recent studies have emphasized that the difference in SRTs between conditions with steady and amplitude-modulated maskers (i.e., the FMB) is highly dependent on the SNRs at which they are measured. As Bernstein and Grant (2009) show, there is a strong negative relationship between the SNR found in a steady noise background and the FMB, both for normal-hearing and hearing-impaired listeners. To control for this confound, Bernstein and Brungart (2011) introduced a technique that adjusts the word-set size in each experimental condition in order to equate the performance levels in steady noise. However, an equalization procedure that is based on similar performance levels in steady noise would itself be biased by a possible effect of periodicity in the target speech. Since it appears likely that, for instance, the absence of any periodicity cues makes it particularly difficult to segregate noise-vocoded speech from a steady noise masker, we took a different approach and used the results obtained in experiment 1 to adjust for the different performance levels in quiet.

This approach is based on the assumption that varying the spectral resolution of the target speech in the presence of a masker has no other effect than to determine its intelligibility. While a degraded spectrum is likely to interfere with the segregation of target and masker when both signals are processed together, as is the case in CI simulations, the spectrum of the maskers in Experiments 2 and 3 was always intact. As demonstrated by Apoux *et al.* (2015), it is differences in TFS *per se* that appear to be crucial for segregating target and masker. Thus, the critical point in the current experiments is that two separate carriers were present throughout. Nevertheless, it should be noted that degrading the spectrum of the target speech with a vocoder also introduces changes to the modulation spectrum, such as a greater similarity of the individual channel envelopes with fewer channels in the vocoder.

Conditions which were found to have very similar intelligibility rates in quiet were: Nx7, FxNx7, and Fx12, as well as Nx12, FxNx10, and Fx24 (see Table I). These 6 target conditions were combined with the 4 different maskers, adding up to 24 conditions. Note that the Fx24 condition was not part of experiment 1, but included in the current one. For convenience, results are presented together with those of experiment 3 that had a similar design but in which the intelligibility of the target speech in quiet was at ceiling.

B. Methods

1. Subjects

Twelve normal-hearing listeners (five females) were tested. Their ages ranged from 18 to 45 yrs with a mean age of 25.9. All participants were native speakers of British English, had audiometric thresholds of less than 20 dB HL at frequencies between 125 and 8000 Hz, and did not participate in experiment 1.

2. Stimuli

The target materials used in experiment 2 were the same recordings of the IEEE sentence corpus as in experiment 1. The harmonic complex maskers were based on *F*0 contours extracted from recordings in the EUROM database of English speech in which different speakers read five- to six-sentence passages (Chan *et al.*, 1995). Sixteen different male talkers with Southern British English accents, and a similar speaking rate and voice quality to that of the target talker were chosen. The median *F*0 frequency of these 16 passages was 122.9 Hz and the first and third quartiles ranged from 107.0 to 144.1 Hz. The median *F*0 of the IEEE target sentences was 117.2 Hz with the first and third quartiles ranging from 103.4 to 136.1 Hz. Thus, the median *F*0 frequency of the target sentences was about 6% lower, but due to the large interquartile range of the *F*0 contours of both masker complexes and target speech, frequent *F*0 contour crossings are guaranteed.

Both the noise and harmonic complex maskers were presented either in a steady-state version or were sinusoidally amplitude-modulated at a rate of 10 Hz with a modulation depth of 100%. For each trial of the experiment, a random portion of the noise or complex maskers was picked and presented along with the target sentence. For the harmonic complex maskers, the order of the talkers on which the contour was based was also randomized so that all 16 were used before any of them was repeated. The onset of all

TABLE I. Target speech conditions in experiment 2. Two sets of three processing conditions with similar percentage correct scores were chosen. The numbers following the abbreviation of the processing conditions indicate the number of frequency bands.

Processing condition	Nx7	FxNx7	Fx12	Nx12	FxNx10	Fx24
Percentage correct score	75.0	78.4	76.1	93.5	92.5	91.2

the maskers was 600 ms before that of the targets and they continued for another 100 ms after the end of the target sentence. An onset and offset ramp of 100 ms was applied to the mixture of target and masker. Waveforms, wide-band spectrograms, and *F*0 contours of an example of all four maskers are shown in Fig. 3.

3. Signal processing

All target stimulus materials were again processed prior to the experiment. The same channel vocoder software as described in the first experiment was used to create the six target speech conditions. The noise maskers were based on a 24-s passage of white noise that was filtered [finite impulse response filter, Greenwood filter spacing, 1-octave smoothing, filter order 1024, fast Fourier transform (fft) window size of 512 samples] to have the same long-term average speech spectrum (LTASS) as the target speech. The LTASS of the unprocessed target speech was determined by computing the power spectral density of the concatenated waveforms using Welch's method (window size 512 samples, 50% overlap, fft length 512 samples). The resulting spectrum was smoothed over 1 octave. *F*0 contours for the harmonic complex maskers were created by interpolating through unvoiced and silent periods using a piecewise cubic Hermite interpolation in logarithmic frequency. The waveforms were synthesized on a period-by-period basis using the Liljencrants-Fant model (Fant *et al.*, 1985), which closely approximates a typical adult male glottal pulse [see Green and Rosen (2013) for details], and matched in spectrum to the long-term average of the target using the same filtering procedure as for the noise maskers.

4. Procedure

The experimental setting and general procedure were the same as in experiment 1. The current experiment consisted of 24 processing conditions presented in background noise (3 vocoding conditions × 2 degrees of spectral resolution × 4 maskers) and 1 additional condition presented in quiet (*F*0-vocoded speech with 24 bands, Fx24). Each condition consisted of 20 sentences, adding up to 500 trials in total. Participants were familiarized with the materials by listening to five sentences of each of the six target speech conditions and two additional example sentences in each of the four background noises.

The SRT for every processing condition was determined by tracking the SNR necessary in order to repeat 50% of the key words in a sentence correctly. The initial SNR was set to +10 dB and adjusted up or down by 11 dB before the first reversal, 7 dB before the second reversal, and 3 dB after that. If the subject got less than half of the key words correct in the first sentence, the SNR was set to +24 dB and the procedure started over again. The SRT was calculated by taking the mean of the largest even number of reversals with 3-dB step size. Throughout the experiment the level of the target and masker together was fixed at about 80 dB SPL over a frequency range of 75 Hz–10 kHz as measured on an artificial ear (type 4153, Brüel & Kjær Sound & Vibration Measurement A/S, Nærum, Denmark).

Psychometric functions were obtained by fitting a single logistic function to the averaged responses of all listeners for each combination of target and masker following the procedure described by Wichmann and Hill (2001). While intercept and slope were estimated without any restrictions, the

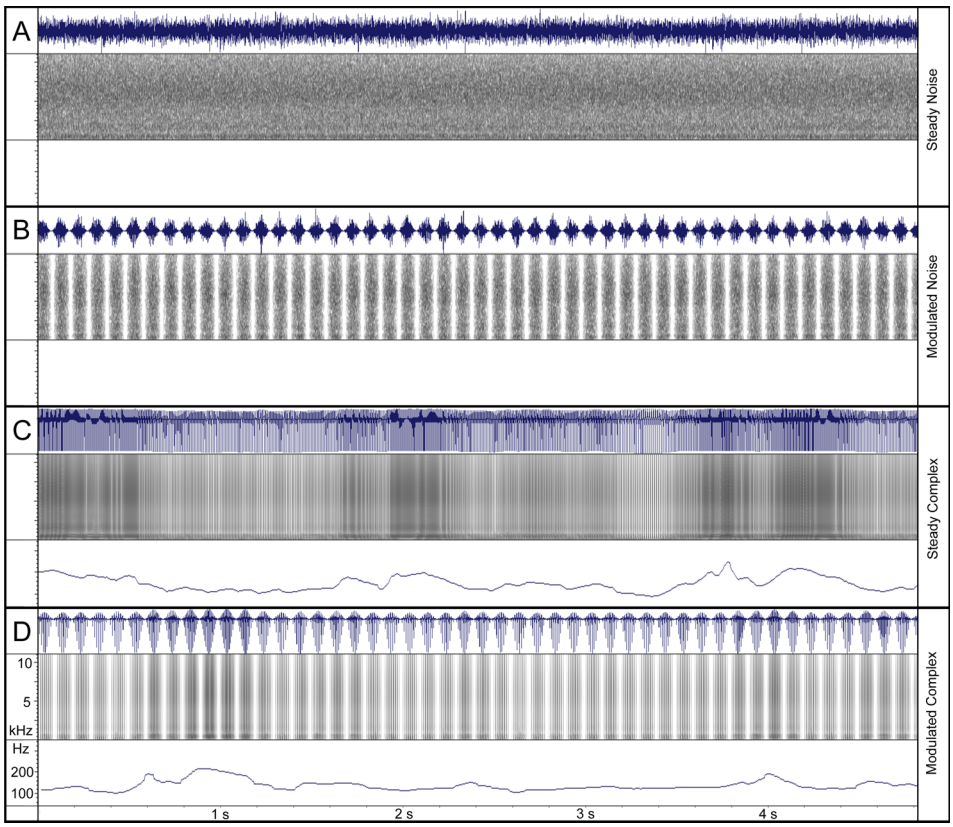

FIG. 3. (Color online) Waveforms, wideband spectrograms, and *F*0 contours of examples of the four maskers used in experiments 2 and 3. (A) An aperiodic steady-state speech-shaped noise, (B) an aperiodic speech-shaped noise with a 10 Hz sinusoidal amplitude modulation, (C) a periodic steady-state harmonic complex with a dynamically varying *F*0 contour, and (D) a periodic harmonic complex with a dynamically varying *F*0 contour and a 10 Hz sinusoidal amplitude modulation.

lapse rate (which sets an upper limit to the performance) was estimated with the constraint to be the same within the set of target speech conditions with a lower intelligibility (Nx7, FxNx7, and Fx12), as well as that with a higher intelligibility (Nx12, FxNx10, and Fx24). The guessing rate was set to zero throughout, since the low semantic predictability and high complexity of the open-set IEEE sentences precludes successful guessing.

C. Results and discussion

Figure 4 shows the SRTs obtained in experiment 2, together with those of experiment 3. For the three target speech conditions with lower intelligibility (Nx7, FxNx7, and Fx12) SRTs on a group level were positive throughout. The targets with higher intelligibility (Nx12, FxNx10, and Fx24) led to substantially lower SRTs and there was a trend for lower SRTs with more periodicity in the targets.

The data were analyzed using a mixed effects model with target intelligibility, target periodicity, masker fluctuations, and masker periodicity as fixed factors, and subjects as a random factor. The main effects of target intelligibility [$F(1,266) = 275.2$, $p < 0.001$] and masker periodicity [$F(1,266) = 110.4$, $p < 0.001$] were highly significant. The main effect of target periodicity [$F(2,264) = 3.1$, $p = 0.047$] was just significant, but there was no significant main effect of masker fluctuations [$F(1,264) = 3.0$, $p = 0.09$]. Furthermore, the interactions of target intelligibility and masker fluctuations [$F(1,266) = 11.1$, $p = 0.001$], target intelligibility and masker periodicity [$F(2,266) = 8.2$, $p < 0.01$], and target periodicity and masker periodicity [$F(2,266) = 6.0$, $p < 0.01$] were significant.

As can be seen in Fig. 4, the SRTs for the four maskers in the FxNx7 condition are closer together than in the other target speech conditions. *Post hoc* pairwise comparisons using Bonferroni-corrected *t*-tests confirmed this observation and showed no significant differences between these four conditions, indicating that neither masker fluctuations nor masker periodicity substantially affected the SRTs in this condition. This result is likely to be one of the main reasons for the significant interactions of target intelligibility and masker periodicity as well as target periodicity and masker periodicity.

In order to enable a more fine-grained examination of the effects of amplitude fluctuations in the masker, Fig. 5 plots the FMB, which is the difference in SRT of a steady compared to a fluctuating masker for each target and masker type. The FMBs of experiment 2 are again plotted together with those of experiment 3. Positive FMBs indicate that listeners were able to benefit from masker fluctuations. *Post hoc t*-tests showed that there were no significant differences between the steady and amplitude-modulated versions of the noise and complex maskers in any of the six target speech conditions. It can, however, be seen in Fig. 5 that there is a trend for more FMBs with the more intelligible targets. While we observed a small but consistent fluctuating-masker *interference* of up to 3 dB for the targets with lower intelligibility (Nx7, FxNx7, and Fx12), this effect disappears when the intelligibility of the targets is higher (Nx12, FxNx10, and Fx24), which also explains the significant interaction of target intelligibility and masker fluctuations.

Figure 6 plots the difference between aperiodic and periodic maskers, termed the *masker-periodicity benefit* (MPB), in Experiments 2 and 3. In stark contrast to the FMB, subjects did benefit from periodicity in the masker across all target speech conditions, with effects of up to about 7 dB. As for the FMB, the MPB increased with the intelligibility of the targets, explaining the significant interaction of target intelligibility and masker periodicity.

As the SRT results show, performance with the FxNx targets was least affected by the differences between the four maskers. This observation is also evident in the pattern of the MPB results, where the smallest benefits were found with the FxNx targets. *Post hoc t*-tests comparing the periodic and aperiodic maskers in all 6 target speech conditions showed that only in the FxNx7 condition was there no significant difference between these, no matter if they were steady [$t(11) = 0.16$, $p = 0.88$] or fluctuating [$t(11) = 0.60$, $p = 0.56$].

The FMB is known to be strongly influenced by the SNR at which a test is carried out (Freyman *et al.*, 2012; Smits and Festen, 2013). Our results suggest that the same is

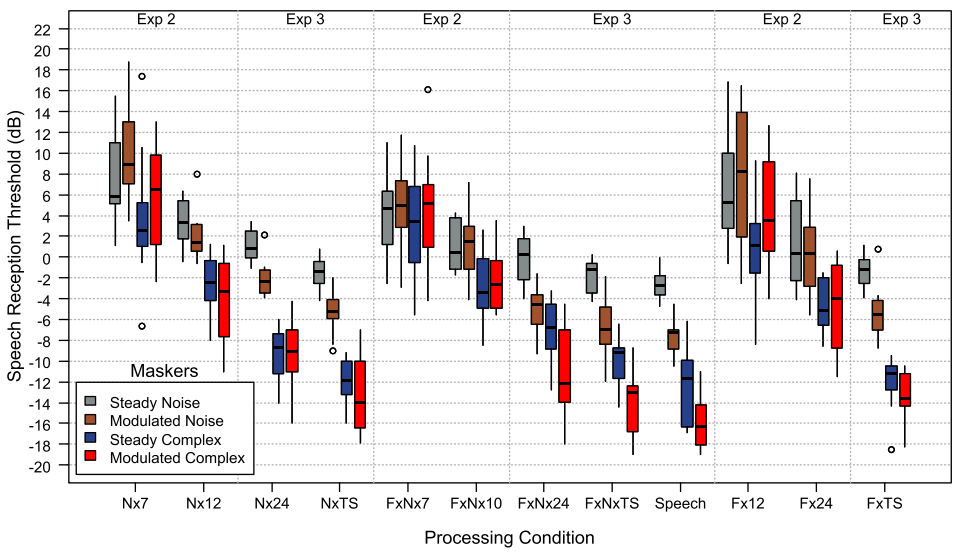

FIG. 4. (Color online) Boxplots of the SRTs obtained in Experiments 2 and 3. Each of the 12 target speech conditions on the *x* axis was tested in combination with the 4 different maskers shown in the legend. Nx stands for noise-vocoding, FxNx for Dudley-vocoding, and Fx for *F*0-vocoding. The numbers affixed to the processing conditions indicate the number of frequency bands in the vocoder. Conditions with the appendix "TS" were produced using TANDEM-STRAIGHT and Speech stands for unprocessed speech. The black horizontal lines in the boxplots indicate the median value.

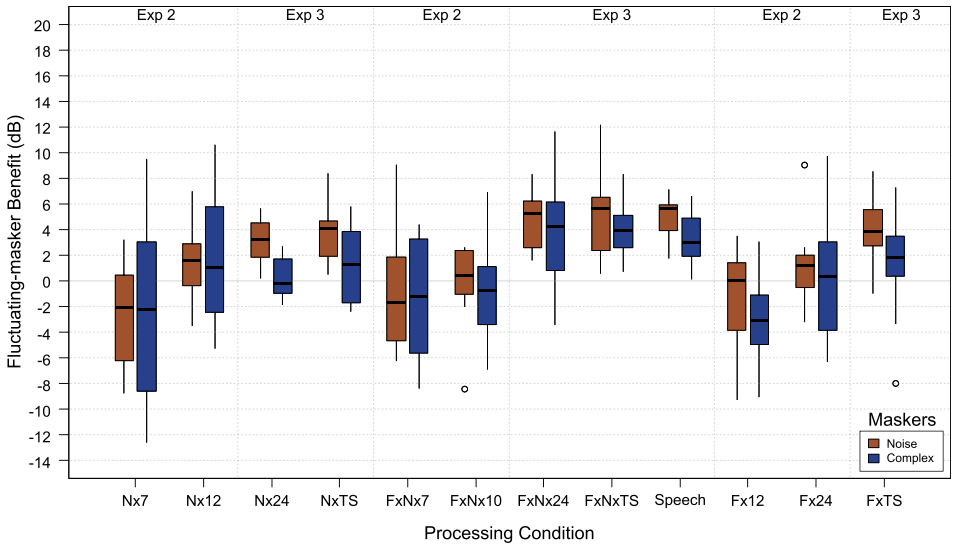

FIG. 5. (Color online) Boxplots of the FMBs obtained in Experiments 2 and 3. For each of the 12 target speech conditions on the x axis, the difference between the steady and amplitude-modulated version of the noise and harmonic complex maskers is plotted. Positive numbers on the y axis indicate a benefit. Target speech conditions are the same as in Fig. 4. The black horizontal lines in the boxplots indicate the median value.

true for the MPB, but the exact relation of the factors involved is difficult to grasp from the snapshot-like SRT data. In order to obtain a broader picture of the results we fitted PFs to the pooled data of each of the 24 target-masker combinations (Fig. 7). On average, the measured SRTs and the estimated 50%-correct values extracted from the PFs were about 0.9 dB apart, indicating a reasonably good fit. As reported previously (Freyman *et al.*, 2012; Smits and Festen, 2013) we found that steady maskers generally led to steeper slopes, as indicated by a significant t-test comparing the slopes of all conditions with steady maskers to all conditions with modulated maskers [$t(11) = 4.8$, $p < 0.001$]. A significant t-test also showed that slopes were steeper for noise maskers when compared to harmonic complex maskers [$t(11) = 3.3$, $p < 0.01$].

These data are also consistent with the idea that the size of the FMB depends on the SNR, with glimpsing observed almost exclusively at negative SNRs. This effect is particularly strong for the two Fx conditions where the slopes of the functions for steady and fluctuating maskers differ a lot, resulting in large fluctuating-masker interference at positive SNRs and similarly large FMBs at negative SNRs. Increasing the intelligibility of the target speech independently enhanced the likelihood of glimpsing, but only the combination with a negative SNR proved to be both necessary and sufficient to enable some degree of FMB.

Importantly, PFs were found to show three distinct patterns depending on the amount of periodicity in the target speech. These patterns are observable for the targets with lower as well as those with higher intelligibility, pointing to common underlying mechanisms involving aspects of periodicity. In both the Nx7 and Nx12 conditions, for example, the functions for steady and modulated maskers are aligned fairly close, while the distance between the noise and harmonic complex maskers is much larger, confirming the finding that the MPB is greater than the FMB. Similarly the close alignment of the boxplots in the FxNx conditions is reflected in the shapes of the respective PFs, which remain relatively close together across the whole range of SNRs. Finally, in the Fx conditions, as already mentioned, the effect of masker fluctuations, but not masker periodicity, depended heavily on the SNR.

Another observation worth mentioning is that the upper performance limits of the targets with lower intelligibility

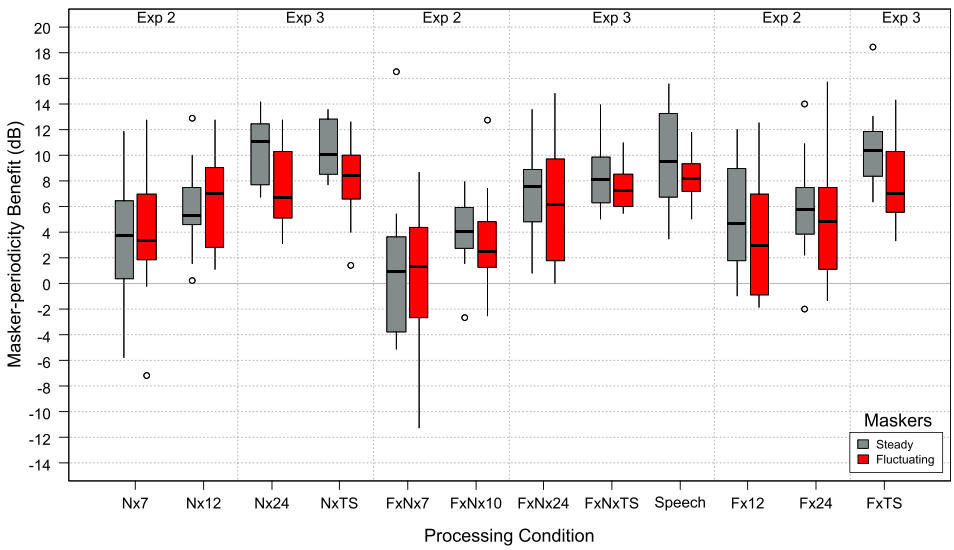

FIG. 6. (Color online) Boxplots of the MPBs obtained in Experiments 2 and 3. For each of the 12 target speech conditions on the x axis, the difference between the noise and harmonic complex version of the steady and amplitude-modulated maskers is plotted. Positive numbers on the y axis indicate a benefit. Target speech conditions are the same as in Fig. 4. The black horizontal lines in the boxplots indicate the median value.

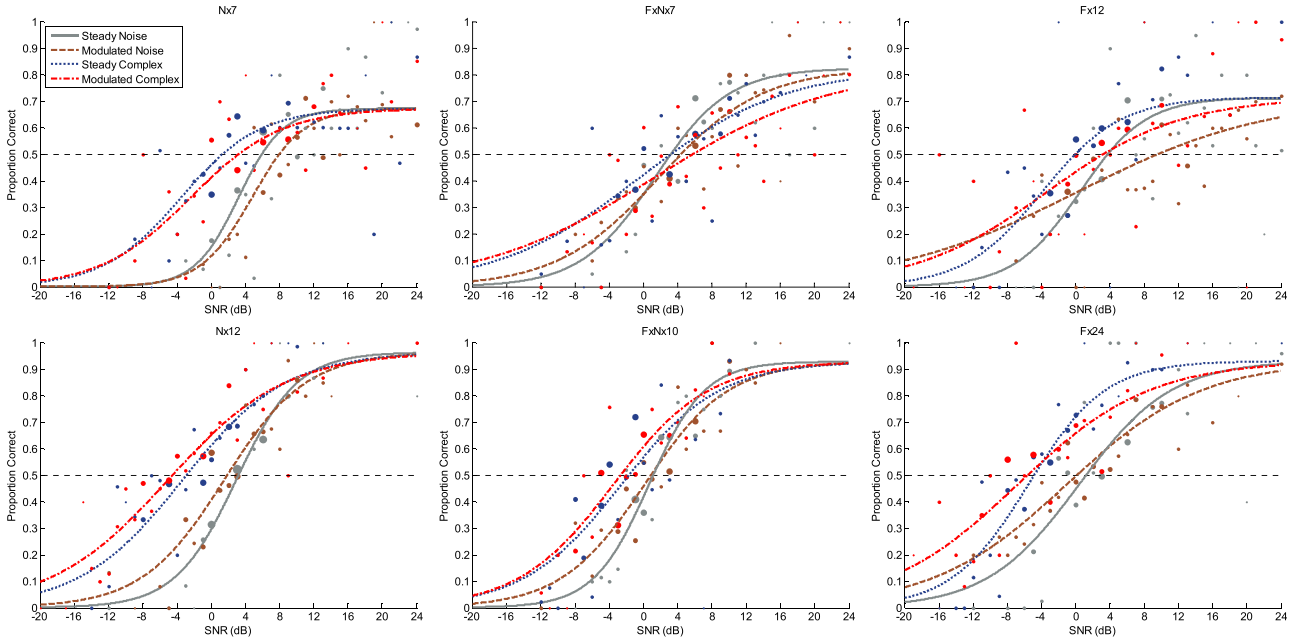

FIG. 7. (Color online) Psychometric functions fitted to the aggregated results of each of the 24 processing conditions (6 targets × 4 maskers) in experiment 2. The target speech condition is indicated above each of the six panels, and labels are the same as in Fig. 4. The horizontal line in each panel indicates the 50%-level that was tracked in the adaptive SRT procedure. The size of the points corresponds to the number of trials at a particular SNR.

(Nx7, FxNx7, and Fx12) differ considerably, with the FxNx7 condition leading to much better performance rates at higher SNRs. As masker levels were very low at these SNRs, the unnatural acoustic properties of the Nx7 and Fx12 targets would have been quite evident. Since the listeners were only presented with a few example sentences before the main experiment, their unfamiliarity with these materials may have affected performance.

IV. EXPERIMENT 3

A. Short introduction and rationale

A key finding of experiment 2 was that, on average, listeners always benefited from periodicity in the masker, but not from masker fluctuations, even when the intelligibility of the target speech was as high as about 90% in quiet. Additionally, there was a clear trend for more MPBs and FMBs (or less fluctuating-masker interference) when the intelligibility of the targets was higher and the resulting SRTs lower. In order to further investigate this relation, we kept the general design of experiment 2, but used target speech with intelligibility rates approaching ceiling level in order to enable testing at lower SRTs.

An initial obstacle of experiment 3 was that the band-vocoder software used in Experiments 1 and 2 cannot be employed to produce noise-vocoded stimuli with a very high number of bands. With more than 24 bands the individual harmonics begin to be resolved, which leads to a clear percept of the *F*0 and an overall less noise-like sound quality, thereby undermining the idea central to noise-vocoding.

An alternative vocoder that does not filter the input speech into separate frequency bands but instead separates the periodic and aperiodic components of the source from the spectral filter is TANDEM-STRAIGHT (Kawahara *et al.*, 2008). By default TANDEM-STRAIGHT produces very natural-sounding speech with a mixed source excitation, but the source estimation procedure can be adapted to produce fully aperiodic or fully periodic speech as well.

Apart from 24-band noise- and *F*0-vocoded speech (Nx24, Fx24), experiment 3 thus also included noise-vocoded, Dudley-vocoded, and *F*0-vocoded speech produced with TANDEM-STRAIGHT (henceforth referred to as NxTS, FxNxTS, and FxTS). Extending the idea of maximizing the spectral detail in the targets, we also used unprocessed speech (referred to as "Speech"). All six target speech conditions in experiment 3 should lead to near perfect intelligibility in quiet. As the results of experiment 1 show (see Fig. 2), the Nx16 and FxNx16 conditions already led to over 95% of correctly repeated keywords. Adding another eight frequency bands was therefore hypothesized to raise the performance levels in quiet to those of unprocessed speech. The even higher spectral resolution of the stimuli produced with TANDEM-STRAIGHT is assumed to result in similarly high scores.

B. Methods

1. Subjects

Twelve normal-hearing listeners (seven females) were tested. Their ages ranged from 18 to 30 yrs with a mean of 22.3 yrs. All participants were native speakers of British English, had audiometric thresholds of less than 20 dB HL at frequencies between 125 and 8000 Hz, and did not participate in Experiments 1 or 2.

2. Stimuli

The target materials were the same recordings of the IEEE sentence corpus as in Experiments 1 and 2, and the maskers were the same as in experiment 2.

3. Signal processing

For the Nx24 and FxNx24 conditions the same channel vocoder software as in Experiments 1 and 2 was used. TANDEM-STRAIGHT was used to produce noise-vocoded speech (NxTS) by keeping the default settings, but fixing the $F0$ to 0 Hz throughout. In order to synthesize Dudley-vocoded speech with TANDEM-STRAIGHT (FxNxTS), the default settings were used, but the values of the sigmoid parameter in the source estimation routine were fixed to 1 and −40, in order to minimize the level of the aperiodic component. This avoids the possibility that higher harmonics are noisier than lower ones, as is the case in natural speech, and ensures comparability with the Dudley-vocoded speech produced with a channel vocoder. The same technique was used to produce $F0$-vocoded speech with TANDEM-STRAIGHT (FxTS), but here the same interpolated $F0$ contours as for the channel vocoder were used as input for the source extraction routine. Additionally, the unprocessed IEEE recordings were used as a sixth target speech condition (Speech).

4. Procedure

The experimental setting and procedure was generally the same as in experiment 2. Before being tested, the participants were familiarized with the materials by listening to five example sentences of each of the three target conditions with an unnatural source (Nx24, NxTS, and FxTS) in quiet, followed by two unprocessed example sentences combined with each of the four maskers at an SNR of 0 dB. For the analyses of the PFs, the lapse rate was set to 0.

C. Results and discussion

The SRTs are shown in Fig. 4, along with the SRTs of experiment 2. As expected, unprocessed speech led to the lowest SRTs with all four maskers. Most importantly, the SRTs in experiment 3 show a stepwise descending pattern for each of the six target speech conditions, indicating that listeners benefited from amplitude fluctuations in the masker, but even more so from periodicity in the masker.

The data were analyzed using a mixed effects model with the fixed effects target condition, masker periodicity, and masker fluctuations, and subjects as a random factor. The main effects of target condition [$F(5,264) = 26.6$, $p < 0.001$], masker periodicity [$F(1,264) = 978.4$, $p < 0.001$], and masker fluctuations [$F(1,264) = 144.4$, $p < 0.001$] were all highly significant. There were also significant interactions of target condition and masker periodicity [$F(5,264) = 2.6$, $p < 0.05$], target condition and masker fluctuations [$F(5,264) = 3.6$, $p < 0.01$], and masker periodicity and masker fluctuations [$F(1,264) = 16.4$, $p < 0.001$].

The SRTs of the three conditions produced with TANDEM-STRAIGHT were almost as low as those of unprocessed speech as indicated by non-significant fixed coefficients [NxTS (1.1, $p = 0.23$), FxNxTS (0.9, $p = 0.34$), and FxTS (1.3, $p = 0.16$)]. The fixed coefficients of the 24-channel vocoded targets, on the other hand, indicate that they led to significantly higher SRTs than unprocessed speech [Nx24 (3.7, $p < 0.001$) and FxNx24 (2.4, $p < 0.01$)].

Furthermore, a separate mixed model that was similar to the previous one but included only the three TANDEM-STRAIGHT conditions showed no significant main effect of target condition [$F(2,132) = 0.48$, $p = 0.62$], indicating that the target periodicity in these conditions did not affect the SRTs.

The FMBs of experiment 3 (Fig. 5) show that the largest benefits were obtained for target speech conditions with a natural mixed source (FxNx24, FxNxTS, and Speech). Additionally, the FMB was consistently found to be lower for harmonic complex maskers. These two findings are likely to have caused the significant interactions of target condition and masker periodicity as well as masker periodicity and masker fluctuations, respectively. Furthermore, *post hoc* Bonferroni-corrected t-tests showed that for the completely voiced or unvoiced target speech (Nx24, NxTS, and FxTS), the FMB for complex maskers was not significantly different from zero. Thus, only target speech with a natural mixed source seems to enable substantial glimpsing in the presence of harmonic complex maskers.

Figure 6 shows the MPBs obtained in experiment 3, added to those of experiment 2. Listeners again strongly benefited from periodicity in the masker across all six target speech conditions. Importantly, with a maximum of about 11 dB in the Nx24 condition, the MPB was almost twice as large as the maximum FMB (about 6 dB, see Fig. 5). The MPB was also consistently larger for steady maskers, which is another reason for the significant interaction of masker periodicity and masker fluctuations. Additionally, *post hoc* t-tests showed that for steady maskers, the FxNx24 condition showed significantly less MPB than the Nx24 condition [$t(11) = 5.1$, $p < 0.001$], and that the same was true for the FxNxTS condition when compared to NxTS [$t(11) = 3.1$, $p < 0.05$] and FxTS [$t(11) = 2.6$, $p < 0.05$]. When the masker was steady, targets with a natural mixed source thus led to smaller MPBs than aperiodic or periodic target speech. This result also explains the significant interaction of target condition and masker periodicity.

As in experiment 2, we again fitted PFs to the pooled data of each of the 24 target-masker combinations (see Fig. 8). The measured SRTs and the estimated 50%-correct values extracted from the PFs were this time about 0.25 dB apart, indicating a good fit. T-tests again showed that steady maskers had steeper slopes than modulated maskers [$t(11) = 3.5$, $p < 0.01$] and that noise maskers had steeper slopes than harmonic complex maskers [$t(11) = 5.0$, $p < 0.001$]. The PFs in the current experiment are mostly located in the negative SNR region, but it is again evident that FMBs and MPBs diminish, or in the case of the FMBs even turn into an interference effect, once they approach positive SNRs. Additionally, the three target conditions with a mixed source (FxNx24, FxNxTS, and Speech) all show a more even spacing of the PFs across the four maskers. The latter observation corresponds well with the FMBs of experiment 3 (Fig. 5), which show that only targets with a mixed source enabled the listeners to substantially benefit from fluctuations in both the noise and the harmonic complex maskers.

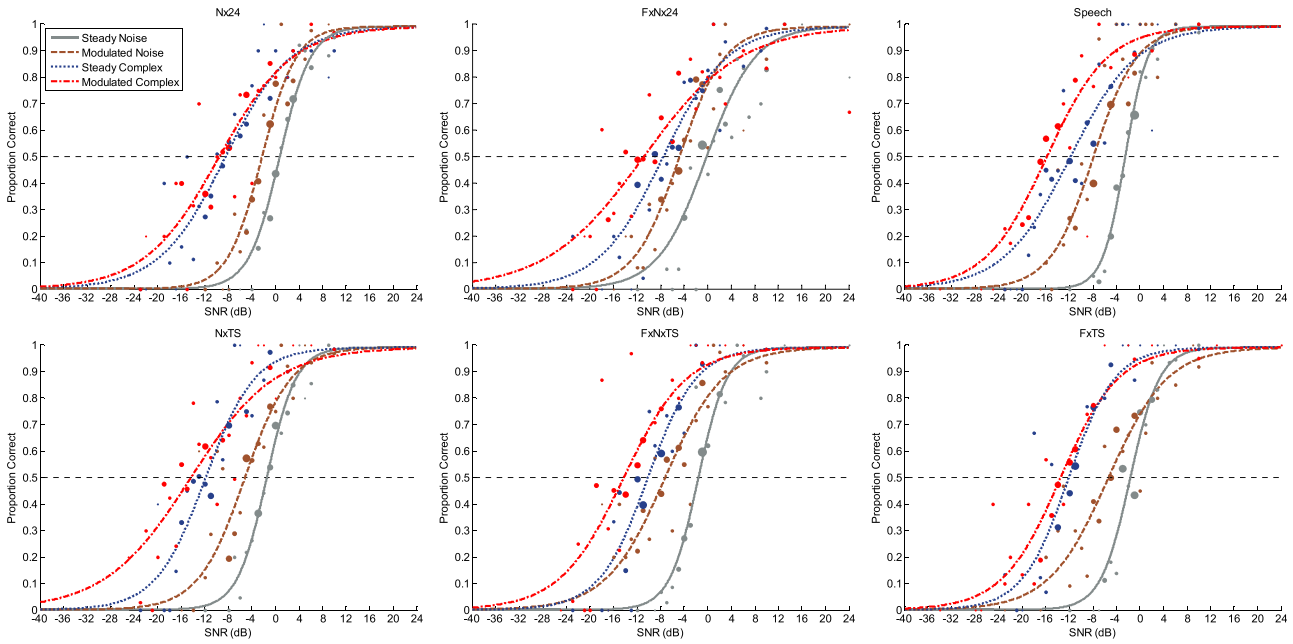

FIG. 8. (Color online) Psychometric functions fitted to the aggregated results of each of the 24 processing conditions (6 targets × 4 maskers) in experiment 3. The target speech condition is indicated above each of the six panels, and the labels are the same as in Fig. 4. The horizontal line in each panel indicates the 50%-level that was tracked in the adaptive SRT procedure. The size of the points corresponds to the number of trials at a particular SNR.

V. GENERAL DISCUSSION

A. Target periodicity in background noise

Generally speaking, the amount of periodicity in the target speech affected the SRTs in Experiments 2 and 3 relatively little. The main effect of target periodicity was just significant in experiment 2, but the direct comparison of three conditions produced with TANDEM-STRAIGHT in experiment 3 revealed no effect of target periodicity. This is somewhat surprising, since one might expect that, for instance, the combination of an aperiodic target with an aperiodic masker would be particularly difficult due to a lack of cues that aid stream segregation. Yet, as the SRTs in Fig. 4 show, performance with the fully voiced Fx targets was in no case more than about 2 dB better than with the aperiodic Nx targets for the two aperiodic noise maskers. The patterns of the PFs for the Nx and Fx targets in experiment 2 (see Fig. 7) in particular, however, reveal that while at the 50%-correct level differences between these target conditions are relatively small, the performance with the Nx targets at lower SNRs is indeed much poorer when the masker is aperiodic. The shapes of the PFs thus confirm that periodicity is important in segregating competing auditory streams, making it clear that SRTs alone are not sufficient in obtaining a complete picture of the patterns in the data. In contrast, this issue does not arise when evaluating the performance in the FxNx conditions. Here the results vary much less between the different maskers across SNRs, suggesting that speech with a natural mix of periodicity and aperiodicity leads to a much more robust percept.

B. Masker fluctuations

The effect of masker fluctuations was found to strongly depend on the intelligibility of the target speech, with interference effects of about 2 dB observed in experiment 2 and maximum benefit of almost 6 dB in experiment 3. This trend is in line with previous studies reporting a strongly reduced ability to glimpse for hearing-impaired listeners and CI users. A recent attempt to model SRTs in fluctuating noise by Smits and Festen (2013) also supports these results by predicting reduced or even negative FMBs at very high SNRs.

Based on the findings of Stone and colleagues (Stone *et al.*, 2011; Stone *et al.*, 2012) this trend could also be explained with reference to the concept of modulation masking. While the 10 Hz sinusoidal amplitude-modulations of the maskers potentially enabled the glimpsing of sections of target speech, they also introduced additional amplitude modulations to the masker envelope that could interfere with informative modulations in the targets. The benefits of glimpsing seem to outweigh the modulation masking at lower SNRs, but not at higher SNRs, where the target speech is already audible when the masker is steady.

The PFs of experiment 2, however, again show that examining the results only through SRTs can be deceptive. While the small effects of masker fluctuations in the Nx and especially the FxNx conditions are fairly stable across different SNRs, much larger and more variable effects were found in the Fx condition. Here masker fluctuations led to considerable benefits at low SNRs, but also particularly large interference effects at high SNRs.

A less well-established result of the current study is that, apart from the targets with lower intelligibility in experiment 2, there appears to be more glimpsing when the masker is aperiodic. This difference is particularly pronounced for the Nx and Fx targets, and might be due to the fact that complex maskers are inherently more coherent and thus easier to segregate from the target speech, no matter if steady or fluctuating.

The largest FMBs of about 6 dB have been found for target speech with a mixed source and a high intelligibility (FxNx24, FxNxTS, and Speech). In conjunction with the small differences in FMB between the noise and complex maskers for these targets, this suggests that a natural mix of periodicity and aperiodicity in the target speech aids glimpsing. Although the maximum FMBs obtained with the Nx targets are only about 2 dB smaller, this finding hence does support the notion that TFS information in the target speech is important in order to benefit from masker fluctuations (Gnansia *et al.*, 2009; Lorenzi *et al.*, 2006).

C. Masker periodicity

The large and consistent MPBs of up to about 11 dB (see Fig. 6) suggest that periodicity in the masker is even more important than masker fluctuations in attempting to segregate target speech from background noise. This finding is in close agreement with the harmonic cancellation theory (de Cheveigné *et al.*, 1995; de Cheveigné *et al.*, 1997b) which states that harmonicity in the masker enables the auditory system to effectively subtract the masking sound from the signal mixture.

There is, however, an additional explanation of the MPB that does not rely on harmonicity but instead the glimpsing opportunities that arise in between the individual harmonics of the complex maskers. A recent study by Deroche *et al.* (2014b) refers to this mechanism as "spectral glimpsing" and provides evidence that spectral glimpsing and harmonic cancellation contribute independently in explaining the MPB. First, they showed that due to the increasing size of spectral dips, both harmonic and inharmonic complexes were less effective in masking the target speech as their *F*0 frequencies increased. In addition, they report that even after controlling for the generally greater spectral glimpsing opportunities in inharmonic maskers, the harmonic complexes still led to consistently lower SRTs and that this effect is independent of the *F*0 frequencies of the complexes.

Another factor explaining the reduced effectiveness of periodic maskers is that, apart from fluctuations at the rate of the *F*0, the envelopes of harmonic complexes with a stationary *F*0 hardly fluctuate, particularly not at the low modulation rates essential for speech intelligibility (Deroche *et al.*, 2014b). As Stone and colleagues (Stone *et al.*, 2011; Stone *et al.*, 2012) have shown, envelope fluctuations, rather than envelope energy, are the primary reason for the effectiveness of aperiodic noise maskers. Contrary to the maskers used by Deroche *et al.* (2014b), the harmonic complexes in the current study had varying *F*0-contours in order to make them more speech-like and thus more ecologically valid. These changes in *F*0, however, also introduce additional slow modulations to the envelopes of the lower auditory filters and it remains to be determined whether this has a substantial effect on performance.

The pattern in the SRTs as well as the PFs shows that the MPB is smallest for targets with a mixed source (FxNx). One possible explanation for this could be that the gaps in the *F*0 contours of these targets made it slightly more difficult to form two separate auditory streams. For the aperiodic and periodic targets in contrast this is likely to be easier since in the former case the harmonic background can be canceled out (de Cheveigné, 1998), while in the latter case, two *F*0 contours are present throughout. Furthermore, the MPB tended to be larger for steady than for fluctuating maskers, which seems intuitive given the fact that in fluctuating maskers there are sections with little or no masker energy, while for steady maskers energy is present throughout.

Crucially, the harmonic complex maskers used in the current study were not only meant to provide a periodic counterpart to the more commonly used aperiodic noise maskers, but also designed in an attempt to better match the acoustic characteristics of speech. Connected stress-timed speech, such as English, is voiced about 50% of the time, while unvoiced sections and pauses only amount to about 25% each (Dellwo *et al.*, 2007; Fourcin, 2010). A harmonic complex masker is thus *per se* more speech-like than an aperiodic noise masker.

As mentioned before, the *F*0s of the IEEE targets and complex maskers differed by about a semitone. It has been shown that even *F*0 differences of this order can help to tell apart signal and noise, but these findings are restricted to artificial stationary vowels (Culling and Darwin, 1993; de Cheveigné *et al.*, 1997a). As described by Darwin (2008), natural speech is too variable for such small differences in *F*0 to matter much. The mechanism for segregating stationary vowels with similar *F*0 frequencies relies on beats caused by the close spacing of the harmonics, which oscillate at relatively slow rates. Studies using real speech as targets have consequently reported hardly any benefit for *F*0 differences of about one semitone and gradual changes as the difference was increased (Bird and Darwin, 1998; Brokx and Nooteboom, 1982).

VI. SUMMARY AND CONCLUSION

The present study found that in quiet testing conditions, aperiodic noise-vocoded speech and vocoded speech with a natural amount of source periodicity were equally intelligible, while fully periodic vocoded speech with an interpolated *F*0 contour is much harder to understand. In the presence of a masker, periodicity in the target speech had a surprisingly small effect. Performance was slightly better with more target periodicity, but only when SRTs were relatively high. Periodicity in the masker, on the other hand, was found to strongly aid speech intelligibility, and this effect was much larger than the FMBs observed. Generally, the higher the intelligibility of the target speech in quiet, the larger were the observed MPBs and FMBs, and a substantial FMB, in particular, required the target speech intelligibility in quiet to be close to ceiling.

In summary, our results show that periodicity in the masker, but surprisingly not the target speech, is an important factor in tracking a speech signal through a background noise. Factors that are thought to underlie the MPB include the presence of discrete spectral components, the relatively sparse modulation spectrum, and the harmonic relation of

the individual components. Further research is needed to identify the respective contributions of these factors.

ACKNOWLEDGMENTS

This project has been funded with support from the European Commission under Contract No. FP7-PEOPLE-2011-290000 and the Medical Research Council of the UK (Grant No. G1001255). We thank Alan O. Cinneide whose software was used to generate the Liljencrants-Fant glottal pulses, Hideki Kawahara and Jeanne Clarke for code and helpful advice concerning TANDEM-STRAIGHT, as well as Tim Green, Martin Cooke, and Natalie Berger for helpful comments.

Apoux, F., Youngdahl, C. L., Yoho, S. E., and Healy, E. W. (**2015**). "Dual-carrier processing to convey temporal fine structure cues: Implications for cochlear implants," J. Acoust. Soc. Am. **138**, 1469–1480.

Bacon, S. P., Opie, J. M., and Montoya, D. Y. (**1998**). "The effects of hearing loss and noise masking on the masking release for speech in temporally complex backgrounds," J. Speech Lang. Hear. Res. **41**, 549–563.

Bernstein, J. G. W., and Brungart, D. (**2011**). "Effects of spectral smearing and temporal fine-structure distortion on the fluctuating-masker benefit for speech at a fixed signal-to-noise ratio," J. Acoust. Soc. Am. **130**, 473–488.

Bernstein, J. G. W., and Grant, K. W. (**2009**). "Auditory and auditory-visual intelligibility of speech in fluctuating maskers for normal-hearing and hearing-impaired listeners," J. Acoust. Soc. Am. **125**, 3358–3372.

Bird, J., and Darwin, C. (**1998**). "Effects of a difference in fundamental frequency in separating two sentences," in *Psychophysical and Physiological Advances in Hearing*, edited by I. R. Palmer, A. Rees, A. Q. Summerfield, and R. Meddis (Whurr, London), pp. 263–269.

Boersma, P., and Weenink, D. (**2013**). "Praat: Doing phonetics by computer [Computer program]," version 5.3.49, http://www.praat.org/ (Last viewed May 13, 2015).

Brokx, J., and Nooteboom, S. (**1982**). "Intonation and the perceptual separation of simultaneous voices," J. Phonetics **10**, 23–36.

Chan, D., Fourcin, A., Gibbon, D., Granström, B., Huckvale, M., Kokkinas, G., Kvale, L., Lamel, L., Lindberg, L., and Moreno, A. (**1995**). "EUROM—A spoken language resource for the EU," in *Proceedings of Eurospeech*, pp. 867–880.

Culling, J. F., and Darwin, C. (**1993**). "Perceptual separation of simultaneous vowels: Within and across-formant grouping by F0," J. Acoust. Soc. Am. **93**, 3454–3467.

Cullington, H. E., and Zeng, F.-G. (**2008**). "Speech recognition with varying numbers and types of competing talkers by normal-hearing, cochlear-implant, and implant simulation subjects," J. Acoust. Soc. Am. **123**, 450–461.

Darwin, C. (**2008**). "Listening to speech in the presence of other sounds," Philos. Trans. R. Soc. London B **363**, 1011–1021.

de Cheveigné, A. (**1998**). "Cancellation model of pitch perception," J. Acoust. Soc. Am. **103**, 1261–1271.

de Cheveigné, A., Kawahara, H., Tsuzaki, M., and Aikawa, K. (**1997a**). "Concurrent vowel identification. I. Effects of relative amplitude and F difference," J. Acoust. Soc. Am. **101**, 2839–2847.

de Cheveigné, A., McAdams, S., Laroche, J., and Rosenberg, M. (**1995**). "Identification of concurrent harmonic and inharmonic vowels: A test of the theory of harmonic cancellation and enhancement," J. Acoust. Soc. Am. **97**, 3736–3748.

de Cheveigné, A., McAdams, S., and Marin, C. M. (**1997b**). "Concurrent vowel identification. II. Effects of phase, harmonicity, and task," J. Acoust. Soc. Am. **101**, 2848–2856.

Dellwo, V., Fourcin, A., and Abberton, E. (**2007**). "Rhythmical classification of languages based on voice parameters," in *Proceedings of the 16th International Congress of Phonetic Sciences*, pp. 1129-1132.

Deroche, M. L., and Culling, J. F. (**2011**). "Voice segregation by difference in fundamental frequency: Evidence for harmonic cancellation," J. Acoust. Soc. Am. **130**, 2855–2865.

Deroche, M. L., Culling, J. F., Chatterjee, M., and Limb, C. J. (**2014a**). "Roles of the target and masker fundamental frequencies in voice segregation," J. Acoust. Soc. Am. **136**, 1225–1236.

Deroche, M. L., Culling, J. F., Chatterjee, M., and Limb, C. J. (**2014b**). "Speech recognition against harmonic and inharmonic complexes: Spectral dips and periodicity," J. Acoust. Soc. Am. **135**, 2873–2884.

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

Fant, G. (**1960**). *Acoustic Theory of Speech Production* (Mouton, The Hague).

Fant, G., Liljencrants, J., and Lin, Q.-G. (**1985**). "A four-parameter model of glottal flow," STL-QPSR **4**, 1–13.

Fastl, H., and Zwicker, E. (**2007**). *Psychoacoustics: Facts and Models* (Springer, Berlin).

Faulkner, A., Rosen, S., and Smith, C. (**2000**). "Effects of the salience of pitch and periodicity information on the intelligibility of four-channel vocoded speech: Implications for cochlear implants," J. Acoust. Soc. Am. **108**, 1877–1887.

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

Fourcin, A. (**2010**). "A note on voice timing and the evolution of connected speech," Logoped. Phoniatr. Vocology **35**, 74–80.

Freyman, R. L., Griffin, A. M., and Oxenham, A. J. (**2012**). "Intelligibility of whispered speech in stationary and modulated noise maskers," J. Acoust. Soc. Am. **132**, 2514–2523.

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.

Gnansia, D., Pean, V., Meyer, B., and Lorenzi, C. (**2009**). "Effects of spectral smearing and temporal fine structure degradation on speech masking release," J. Acoust. Soc. Am. **125**, 4023–4033.

Green, T., and Rosen, S. (**2013**). "Phase effects on the masking of speech by harmonic complexes: Variations with level," J. Acoust. Soc. Am. **134**, 2876–2883.

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

Hopkins, K., and Moore, B. C. (**2009**). "The contribution of temporal fine structure to the intelligibility of speech in steady and modulated noise," J. Acoust. Soc. Am. **125**, 442–446.

Hopkins, K., Moore, B. C., and Stone, M. A. (**2008**). "Effects of moderate cochlear hearing loss on the ability to benefit from temporal fine structure information in speech," J. Acoust. Soc. Am. **123**, 1140–1153.

Kawahara, H., Morise, M., Takahashi, T., Nisimura, R., Irino, T., and Banno, H. (**2008**). "TANDEM-STRAIGHT: A temporally stable power spectral representation for periodic signals and applications to interference-free spectrum, F0, and aperiodicity estimation," in *Proceedings of International Conference on Acoustics, Speech and Signal Processing*, pp. 3933–3936.

Kwon, B. J., Perry, T. T., Wilhelm, C. L., and Healy, E. W. (**2012**). "Sentence recognition in noise promoting or suppressing masking release by normal-hearing and cochlear-implant listeners," J. Acoust. Soc. Am. **131**, 3111–3119.

Loizou, P. C. (**2013**). *Speech Enhancement: Theory and Practice* (CRC Press, Boca Raton, FL).

Lorenzi, C., Gilbert, G., Carn, H., Garnier, S., and Moore, B. C. (**2006**). "Speech perception problems of the hearing impaired reflect inability to use temporal fine structure," Proc. Natl. Acad. Sci. U.S.A. **103**, 18866–18869.

Miller, G. A., and Licklider, J. (**1950**). "The intelligibility of interrupted speech," J. Acoust. Soc. Am. **22**, 167–173.

Moore, B. C. J. (**2012**). "The importance of temporal fine structure for the intelligibility of speech in complex backgrounds," in *Speech Perception and Auditory Disorders*, edited by T. Dau, M. L. Jepsen, T. Poulsen, and J. C. Dalsgaard (The Danavox Jubilee Foundation, Ballerup, Denmark), pp. 21–32.

Nelson, P. B., Jin, S.-H., Carney, A. E., and Nelson, D. A. (**2003**). "Understanding speech in modulated interference: Cochlear implant users and normal-hearing listeners," J. Acoust. Soc. Am. **113**, 961–968.

Obleser, J., and Weisz, N. (**2012**). "Suppressed alpha oscillations predict intelligibility of speech and its acoustic details," Cereb. Cortex **22**, 2466–2477.

Oxenham, A. J. (**2008**). "Pitch perception and auditory stream segregation: Implications for hearing loss and cochlear implants," Trends Amplif. **12**, 316–331.

Oxenham, A. J., and Simonson, A. M. (**2009**). "Masking release for low-and high-pass-filtered speech in the presence of noise and single-talker interference," J. Acoust. Soc. Am. **125**, 457–468.

Peters, R. W., Moore, B. C. J., and Baer, T. (**1998**). "Speech reception thresholds in noise with and without spectral and temporal dips for hearing-impaired and normally hearing people," J. Acoust. Soc. Am. **103**, 577–587.
Plomp, R., and Mimpen, A. (**1979**). "Improving the reliability of testing the speech reception threshold for sentences," Int. J. Audiol. **18**, 43–52.
Rasch, R., and Plomp, R. (**1999**). "The perception of musical tones," in *The Psychology of Music*, edited by D. Deutsch (Academic Press, San Diego), pp. 89–112.
Rosen, S. (**1992**). "Temporal information in speech: Acoustic, auditory and linguistic aspects," Philos. Trans. R. Soc. London B **336**, 367–373.
Rosen, S., Souza, P., Ekelund, C., and Majeed, A. (**2013**). "Listening to speech in a background of other talkers: Effects of talker number and noise vocoding," J. Acoust. Soc. Am. **133**, 2431–2443.
Rothauser, E. H., Chapman, N. D., Guttman, N., Nordby, K. S., Silbiger, H. R., Urbanek, G. E., and Weinstock, M. (**1969**). "IEEE recommended practice for speech quality measurements," IEEE Trans. Audio Electroacoust. **17**, 225–246.
Scott, S. K., Blank, C. C., Rosen, S., and Wise, R. J. (**2000**). "Identification of a pathway for intelligible speech in the left temporal lobe," Brain **123**, 2400–2406.
Shannon, R. V., Zeng, F.-G., Kamath, V., Wygonski, J., and Ekelid, M. (**1995**). "Speech recognition with primarily temporal cues," Science **270**, 303–304.
Smits, C., and Festen, J. M. (**2013**). "The interpretation of speech reception threshold data in normal-hearing and hearing-impaired listeners: II. Fluctuating noise," J. Acoust. Soc. Am. **133**, 3004–3015.
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., Füllgrabe, C., Mackinnon, R. C., and Moore, B. C. (**2011**). "The importance for speech intelligibility of random fluctuations in 'steady' background noise," J. Acoust. Soc. Am. **130**, 2874–2881.
Stone, M. A., Füllgrabe, C., and Moore, B. C. (**2012**). "Notionally steady background noise acts primarily as a modulation masker of speech," J. Acoust. Soc. Am. **132**, 317–326.
Vestergaard, M. D., and Patterson, R. D. (**2009**). "Effects of voicing in the recognition of concurrent syllables," J. Acoust. Soc. Am. **126**, 2860–2863.
Whitmal, N. A. III, Poissant, S. F., Freyman, R. L., and Helfer, K. S. (**2007**). "Speech intelligibility in cochlear implant simulations: Effects of carrier type, interfering noise, and subject experience," J. Acoust. Soc. Am. **122**, 2376–2388.
Wichmann, F. A., and Hill, N. J. (**2001**). "The psychometric function: I. Fitting, sampling, and goodness of fit," Percept. Psychophys. **63**, 1293–1313.
Xu, Y. (**2013**). "ProsodyPro—A tool for large-scale systematic prosody analysis," in *Proceedings of Tools and Resources for the Analysis of Speech Prosody*, pp. 7–10.