JASA ARTICLE

No evidence for a benefit from masker harmonicity in the perception of speech in noise

Kurt Steinmetzger[1] and Stuart Rosen[2,a)]

[1]*Section of Biomagnetism, Department of Neurology, Heidelberg University Hospital, Im Neuenheimer Feld 400, 69120 Heidelberg, Germany*

[2]*Speech, Hearing and Phonetic Sciences, University College London (UCL), Chandler House, 2 Wakefield Street, London, WC1N 1PF, United Kingdom*

ABSTRACT:

When assessing the intelligibility of speech embedded in background noise, maskers with a harmonic spectral structure have been found to be much less detrimental to performance than noise-based interferers. While spectral "glimpsing" in between the resolved masker harmonics and reduced envelope modulations of harmonic maskers have been shown to contribute, this effect has primarily been attributed to the proposed ability of the auditory system to cancel harmonic maskers from the signal mixture. Here, speech intelligibility in the presence of harmonic and inharmonic maskers with similar spectral glimpsing opportunities and envelope modulation spectra was assessed to test the theory of harmonic cancellation. Speech reception thresholds obtained from normal-hearing listeners revealed no effect of masker harmonicity, neither for maskers with static nor dynamic pitch contours. The results show that harmonicity, or time-domain periodicity, as such, does not aid the segregation of speech and masker. Contrary to what might be assumed, this also implies that the saliency of the masker pitch did not affect auditory grouping. Instead, the current data suggest that the reduced masking effectiveness of harmonic sounds is due to the regular spacing of their spectral components.

https://doi.org/10.1121/10.0017065

(Received 19 July 2022; revised 6 January 2023; accepted 10 January 2023; published online 10 February 2023)

[Editor: Christian Lorenzi] Pages: 1064–1072

I. INTRODUCTION

Recovering a speech signal from a mixture of sound sources is a ubiquitous task in everyday life. How well a given interferer masks the target speech depends on a complex interplay of various acoustic factors. In normal hearing, for example, periodic tone complexes with a harmonic spectral structure are far less effective maskers of speech than aperiodic noise, with speech reception thresholds (SRTs) differing by up to 10 dB (Steinmetzger and Rosen, 2015). This effect, termed the *masker-periodicity benefit* (MPB), was found to be substantially larger than the fluctuating-masker benefit obtained from slow masker amplitude fluctuations and, to a lesser extent, even observed in cochlear implant users (Steinmetzger and Rosen, 2018). As speech is mostly voiced, and thus periodic, the MPB is also one important reason for why a competing talker is a less effective masker of speech than noise (e.g., Bronkhorst and Plomp, 1992; Brungart *et al.*, 2001; Rosen *et al.*, 2013). Factors thought to contribute to the MPB include the possibility to spectrally "glimpse" portions of the target speech in between the resolved lower masker harmonics (Deroche *et al.*, 2014a, 2014b; Guest and Oxenham, 2019), the absence of random envelope modulations in periodic maskers (Steinmetzger *et al.*, 2019; Stone *et al.*, 2011; Stone *et al.*, 2012), and, most importantly, periodicity-related pitch cues that aid stream segregation (e.g., Moore and Gockel, 2012; Oxenham, 2008; Rosen *et al.*, 2013). Attempts to predict the MPB have shown that speech intelligibility models including both auditory and modulation filterbanks, which were assumed to account for the respective contributions of spectral glimpsing and modulation masking, can only explain about half of its magnitude (Steinmetzger *et al.*, 2019). This suggests that the contribution of pitch-based stream segregation makes up for the other half of the MPB.

Regarding the contribution of auditory stream segregation, the harmonic relation of the component tones in periodic sounds has been claimed to be of particular importance, reflecting the fact that component frequencies that are integer multiples of the fundamental frequency ($F0$) result in sounds with a salient pitch. The theory of harmonic cancellation (de Cheveigné, 2021) posits that harmonicity, or periodicity in the time domain, enables the auditory system to segregate a harmonic masker from any kind of speech or nonspeech target signal by cancelling it from the signal mixture. The results from initial studies with competing artificial vowels (de Cheveigné *et al.*, 1995; de Cheveigné *et al.*, 1997), as well as subsequent masked-speech experiments (Deroche and Culling, 2011; Deroche *et al.*, 2014b; Prud'homme *et al.*, 2022b; Steinmetzger and Rosen, 2015),

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

https://doi.org/10.1121/10.0017065

may be taken to provide direct evidence for this theory as harmonic interferers were consistently less effective maskers of speech than inharmonic or aperiodic ones. The maskers in the aforementioned studies were rendered inharmonic by reverberating them and/or adding sinusoidal *F*0 modulations (Deroche and Culling, 2011), by jittering the masker harmonics (de Cheveigné *et al.*, 1995; de Cheveigné *et al.*, 1997; Deroche *et al.*, 2014b), or speech intelligibility was compared in the presence of harmonic complexes and aperiodic noise maskers (Prud'homme *et al.*, 2022b; Steinmetzger and Rosen, 2015). It should be noted that recent data by Prud'homme *et al.* (2022b) suggest that this pattern does not extend to speech-on-speech masking scenarios as they observed no difference between unprocessed and noise-vocoded speech maskers. However, none of these approaches allows the isolation of the contribution of harmonicity *per se* as the respective signal manipulations inevitably also affected spectral glimpsing opportunities and envelope modulations. An exception to this limitation is a study by Roberts *et al.* (2010) in which all of the components of harmonic complex maskers were shifted in frequency, thus preserving spectral regularity. However, as they applied this manipulation to the target speech too, the results are difficult to interpret. A similar situation occurs in considering the results of Popham *et al.* (2018), who concurrently jittered the harmonics of target and background speech, making it unclear if the claimed positive effect of harmonicity was due to its presence in the masker, target, or both.

In the current study, it was attempted to rule out these confounds by determining the intelligibility of unprocessed target speech in the presence of harmonic complex tone maskers as well as two types of inharmonic maskers with preserved spectral regularity. These were produced by either shifting all of the component tones by fixed amounts in frequency or rotating their spectra. To generate the frequency-shifted inharmonic maskers, all of the component tones were shifted by 25% of their *F*0 as this has been shown to result in the greatest degree of inharmonicity (Roberts *et al.*, 2010). Because spectral glimpsing opportunities increase with increasing masker *F*0 (Deroche *et al.*, 2014b), half of the stimuli were shifted upward and the other half downward. To obtain the spectrally rotated inharmonic maskers, the spectrograms of the harmonic maskers were reflected relative to a midpoint frequency of 2 kHz so that, for example, components near 200 Hz end up near 3.8 kHz and vice versa (Blesser, 1972; Scott *et al.*, 2000). All three of the masker types (*harmonic*, *shifted*, and *rotated*) were presented with static as well as dynamically varying *F*0 contours extracted from natural speech. Inharmonic complex tone maskers with dynamic *F*0 contours have so far not been used, although they provide the advantage of being inherently more speech-like than their static *F*0 equivalents. Additionally, all of the maskers were presented with low, medium, and high *F*0s to evaluate whether the effect of masker harmonicity generalises across masker *F*0 levels and *F*0 contour types. The SRTs for the different maskers were determined behaviourally using a sample of normal-hearing listeners. Supplemental acoustic analyses included an autocorrelation-based measure to compare the degree of periodicity of the different maskers as well as an analysis of the masker envelope modulations to verify whether the shifted and rotated inharmonic maskers did indeed not differ from their harmonic counterparts regarding their modulation spectra.

II. METHODS

A. Participants

Twelve normal-hearing listeners (ten females, two males) were tested. Their ages ranged from 18 to 35 years old (mean = 22.8 yr). All of the participants were native speakers of British English and had audiometric thresholds of less than 20 dB hearing level (HL) at octave frequencies between 125 and 8000 Hz. All of the subjects gave written consent prior to the experiment, and the study was approved by the University College London Research Ethics Committee.

B. Target sentence materials

Target sentences were drawn from the corpus created by Boyle *et al.* (2013) and spoken by an adult male Southern British English talker with a relatively high-pitched voice. This corpus is based on the well-known IEEE/Harvard sentences (Rothauser *et al.*, 1969), adapted to standard UK English, with some further lists added. It consists of 25 lists of 30 sentences each with each sentence containing 5 keywords for scoring. Based on 150 selected sentences sampled through the lists, the median *F*0 was ~156 Hz with the first and third quartiles ranging from 131 to 187 Hz (expressed as a proportion of time that each frequency occurs). To allow for a better equalisation of the spectra of the target materials and maskers, the target sentences were filtered between 180 Hz and 3.8 kHz (eighth-order Chebyshev type II high- and low-pass filters with forward/backward low-pass filtering for sufficient attenuation of the higher frequencies, respectively). All of the sentences were normalised to a common root mean square level.

C. Masker construction

In total, there were 19 distinct unintelligible maskers. The first of these was speech-shaped noise, which was drawn from a 23.8-s passage of noise and the only masker that had a continuous frequency spectrum. The 18 remaining maskers all contained discrete spectral components and were varied factorially on 3 attributes in a $3 \times 2 \times 3$ design: (1) whether their spectral components were harmonic, inharmonic through spectral shifting, or inharmonic through spectral rotation; (2) whether they had static or dynamic *F*0 contours; and (3) the median *F*0 of the contours, which were lower (at 100 Hz), approximately equal to (at 150 Hz) or higher (at 225 Hz) than the median *F*0 of the target materials, changes of 0.585 of an octave.

The maskers with discrete spectral components were based on recordings from the EUROM database (Chan *et al.*, 1995), consisting of 5- to 6-sentence passages read by

16 different male talkers. Using methods as previously described (Green and Rosen, 2013; Steinmetzger and Rosen, 2015), the *F*0 contours of the 16 passages were extracted and interpolated through unvoiced and silent periods to generate a continuous *F*0 contour. The 16 contours were then normalised separately to the 3 median *F*0s specified above. The harmonic maskers were then synthesised on a component-by-component basis with equal-amplitude components in sine phase up to 9 kHz, well beyond the range used in the actual experiment. Shifted inharmonic equivalents of these harmonic complex maskers were produced by shifting the frequencies of the component tones up or down by 25% of the *F*0 and then doing the same component-by-component synthesis. During the experiment, the upward- or downward-shifted version of a stimulus was picked randomly while ensuring that both shift directions occurred equally often overall. Rotated inharmonic stimuli were processed from the harmonic stimuli by spectral rotation around 2 kHz using the technique described by Blesser (1972).

The set of fixed *F*0s to be used for the static contours had two constraints. One arises from the fact that shifted and rotated stimuli are structurally identical for static contours (consisting of equally spaced, nonharmonically related discrete spectral components), and the rotated stimuli were required to have the same 25% degree of mistuning as the shifted stimuli. A possible set of static *F*0s that satisfied this criterion was therefore determined, resulting in a total of 90 values over the *F*0 range of 70–330 Hz. From this set, 20 *F*0s were chosen (1 for each test sentence in a single SRT measurement; see below) to be distributed in *F*0 in approximately the same way as the *F*0s in the dynamically varying contours. This was achieved by determining 20 evenly spaced quantiles for each of the 3 *F*0 distributions of the dynamically varying contours (low, mid, and high) and identifying the closest of the available *F*0s for each quantile.

Finally, all of the maskers were spectrally shaped to the long-term average spectrum of the target speech materials. Figure 1 shows examples of the maskers with discrete spectral components with a medium *F*0 level (median of 150 Hz).

D. Experimental design and procedure

There were 18 tone complex maskers that varied with respect to the factors *masker harmonicity* (harmonic, shifted, or rotated), *masker F0 contour* (static or dynamic), and *masker F0 level* (low, mid, or high), as well as speech-shaped noise. The intelligibility of unprocessed (but band-pass filtered) target speech was assessed in the presence of each of these maskers, resulting in 19 experimental conditions.

SRTs were determined for each condition using the first 20 sentences of the 30 in a single sentence list by tracking the signal-to-noise ratio (SNR) necessary to correctly repeat 50% of the key words in a sentence. The initial SNR was set to +9 dB and adjusted up or down by 9 dB before the first reversal, 7 dB before the second reversal, 5 dB before the third 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 a 3-dB step size. The masker level was kept constant, and the speech level was adjusted to achieve a specific SNR.

The verbal responses were scored by the experimenter before the next sentence was played. A so-called loose keyword scoring technique was applied in which the roots of the five keywords had to be correctly identified. No feedback was given following the responses. The presentation and logging of the responses was carried out using locally developed MATLAB software.

The order of the 19 conditions was fully randomised using a Latin square design as was the order of the sentence lists. For each trial, a random portion of the respective masker was picked and presented along with the target sentence. For the tone complex maskers, the order of the files from which these portions were extracted was permuted, ensuring that each talker file was picked once before any of them was repeated. The onset of all of the maskers was 600 ms before that of the target sentence, and they continued for another 100 ms after its

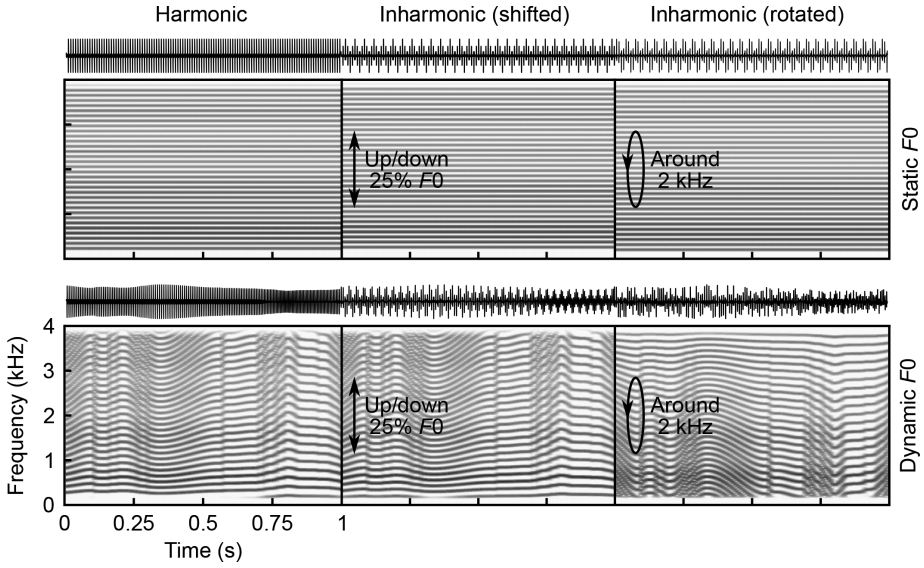

FIG. 1. Waveforms and narrowband spectrograms of examples of the harmonic and inharmonic maskers with static (upper) and dynamic *F*0 contours (lower) are shown. The shifted inharmonic maskers were produced by shifting the component tones of the harmonic maskers up or down by 25% of the median *F*0. The rotated inharmonic maskers were generated by rotating the spectra of the harmonic maskers around 2 kHz. All stimuli depicted have a median *F*0 level of 150 Hz.

https://doi.org/10.1121/10.0017065

end. The mixture of speech and masker was tapered on and off using 100 ms raised-cosine ramps.

Before the experiment, the participants were familiarised with the materials by presenting them with 3 trials in each of the 19 conditions using the same adaptive procedure as in the main experiment. Sentence lists 1–3 were reserved for the familiarisation procedure and lists 4–22 were used in the main experiment. The total duration of the experiment, including hearing screening and familiarisation procedure, was about 60 min long and the participants could take breaks whenever they wished to.

The experiment took place in a double-walled sound-attenuating booth. The stimuli were converted with 24-bit resolution at a sampling rate of 44.1 kHz using an RME Babyface soundcard (Haimhausen, Germany) and presented diotically over Sennheiser HD650 headphones (Wedemark, Germany). The level of the masker was set to about 70 dB sound pressure level (SPL) over a frequency range of 70 Hz–4 kHz as measured with an artificial ear (Brüel and Kjær, type 4153, Nærum, Denmark).

III. RESULTS

A. Behavioural results

The behavioural data are shown in Fig. 2 and were statistically analysed by fitting a general linear mixed-effects regression model in a top-down manner with *p*-values based on the Satterthwaite approximation of the degrees of freedom. The subjects and sentence lists were included as random effects. The detailed statistical results are provided in Table I. Speech-shaped noise was included in the experiment for comparison but left out of the statistical analysis for simplicity. The main effects of *masker harmonicity* (harmonic, shifted, or rotated), *masker F0 contour* (static or dynamic), and *masker F0 level* (low, mid, or high) were all highly significant. The final model also included the complete set of fixed-effects interactions, all of which were significant too.

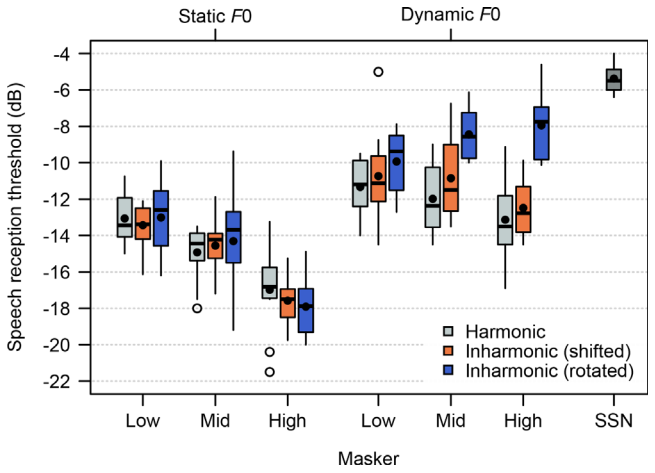

FIG. 2. (Color online) Behavioural results. The SRTs on the *y* axis indicate the SNRs required to correctly repeat 50% of the keywords. The black horizontal lines in the boxplots indicate the median, and the black dots indicate the mean. The boxes range from the first to the third quartiles, the whisker length is up to 1.5 times the interquartile range, and the black circles represent outliers. SSN, speech-shaped noise.

TABLE I. Statistical evaluation of the behavioural data, excluding the results for speech-shaped noise. The results are displayed for three different mixed-effects regression models, including all of the maskers or only maskers with static or dynamic *F*0 contours. Df, degrees of freedom; *F*, test statistic; *p*, probability value. Asterisks both here and in the main text indicate levels of statistical significance, with *** indicating $p < 0.001$, ** indicating $p < 0.01$ and * indicating $p < 0.05$.

Model 1: All maskers	df	*F*	*p*
Masker harmonicity	2, 176.46	31.78	<0.001***
Masker *F*0 contour	1, 174.09	545.19	<0.001***
Masker *F*0 level	2, 172.03	66.96	<0.001***
Masker harmonicity * masker *F*0 contour	2, 174.09	34.98	<0.001***
Masker harmonicity * masker *F*0 level	4, 173.00	3.23	0.014*
Masker *F*0 contour * masker *F*0 level	2, 171.54	37.91	<0.001***
Masker harmonicity * masker *F*0 contour * masker *F*0 level	4, 171.59	5.92	<0.001***
Model 2: Static *F*0 maskers			
Masker harmonicity	2, 81.49	0.20	0.821
Masker *F*0 level	2, 75.93	102.32	<0.001***
Masker harmonicity * masker *F*0 level	4, 76.69	1.07	0.376
Model 3: Dynamic *F*0 maskers			
Masker harmonicity	2, 80.28	71.91	<0.001***
Masker *F*0 level	2, 73.12	3.35	0.040*
Masker harmonicity * masker *F*0 level	4, 74.33	8.83	<0.001***

When only the static *F*0 maskers were included in the model, the SRTs were not affected by their harmonicity, as there was no significant main effect of *masker harmonicity* nor the *masker harmonicity** *masker F*0 *level* interaction. There was, however, a highly significant main effect of *masker F*0 *level*. Compared to the maskers with a high *F*0 level, SRTs for the mid and low *F*0 maskers were estimated to be 2.1 dB [$t_{(75.66)} = 3.87$, $p < 0.001$***] and 4.0 dB higher [$t_{(78.44)} = 7.37$, $p < 0.001$***], respectively.

A model that only included the dynamic *F*0 maskers returned a highly significant main effect of *masker harmonicity*, a significant main effect of *masker F0 level*, and also a highly significant interaction of the two factors. However, the significant main effect of *masker harmonicity*, as well as the significant interaction, can be attributed to the diverging pattern of results observed for the rotated maskers. In contrast to all of the other interferers, the SRTs for these maskers increased with increasing masker *F*0 level. Compared to the harmonic maskers, the estimated SRTs for the rotated maskers were, on average, 5.2 dB higher [$t_{(78.41)} = 10.13$, $p < 0.001$***]. Crucially, however, SRTs for the shifted maskers were not significantly higher than those for the harmonic interferers [0.7 dB, $t_{(73.49)} = 1.30$, $p = 0.196$]. Moreover, pairwise comparisons of the shifted and harmonic maskers at the different *F*0 levels also did not return

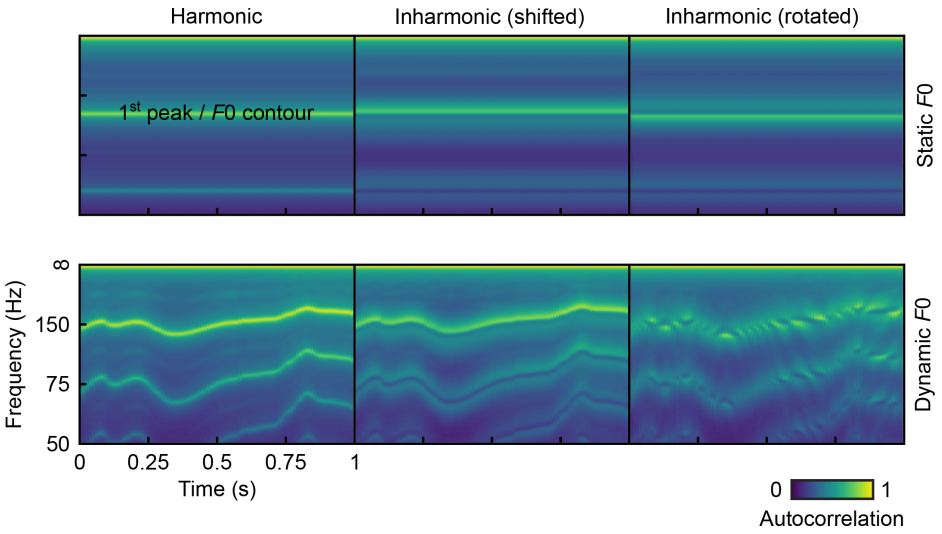

FIG. 3. (Color online) Summary autocorrelation function spectrograms of the masker examples displayed in Fig. 1. The first peak in these spectrograms represents the *F*0 contour and its intensity reflects the degree of periodicity.

any significant differences, even when not correcting for multiple comparisons [low $F0 = 0.6$ dB, $t_{(11)} = 1.09$, $p = 0.30$; mid $F0 = 1.1$ dB, $t_{(11)} = 2.01$, $p = 0.070$; high $F0 = 0.7$ dB, $t_{(11)} = 1.62$, $p = 0.133$].

B. Masker periodicity

To derive a measure of the periodicity of the different maskers, summary autocorrelation functions (SACFs) were computed (Meddis and Hewitt, 1991; Meddis and O'Mard, 1997). For each individual stimulus, autocorrelation functions were calculated for the low-pass filtered (second-order Butterworth, cutoff 1 kHz) outputs of 22 gammatone filters with equivalent rectangular bandwidths and centre frequencies ranging from 0.2 to 4 kHz and summed together into SACFs. This procedure was applied across the duration of each stimulus with a step size of 1 ms and a Hann-window size of 5 ms, resulting in the spectrographic representations appearing in Fig. 3.

After transforming lag times into frequencies, the first peak in these SACF spectrograms indicates the *F*0 contour of a given stimulus, whereas the height of this peak reflects the degree of periodicity and may also be interpreted as a measure of the pitch strength (Meddis and Hewitt, 1991; Yost *et al.*, 1996). As can be observed in Fig. 3, this peak is noticeably more pronounced for the harmonic stimuli. Furthermore, small pitch shifts are evident for both inharmonic maskers, which is in agreement with previous studies using frequency-shifted complex tones (de Boer, 1956; Patterson, 1973; Schouten *et al.*, 1962).

The time-averaged degree of periodicity was subsequently determined by computing the height of the first peak relative to the closest neighbouring trough for each consecutive SACF time frame and averaging these values across the duration of the respective stimulus. The resulting periodicity distributions for each combination of *masker harmonicity* (harmonic, shifted, or rotated) and *masker F0 contour* (static, dynamic) are shown in Fig. 4. For simplicity, the results were averaged across the three masker *F*0 levels. These data were then statistically analysed using an analysis of variance (ANOVA) with the factors *masker harmonicity* and *masker F0 contour*. Both main effects [*masker harmonicity*, $F_{(2,426)} = 1271.52$, $p < 0.001$***; *masker F0 contour*, $F_{(1,426)} = 150.59$, $p < 0.001$***], as well as their interaction [$F_{(2,426)} = 37.87$, $p < 0.001$***] were highly significant. Furthermore, a *post hoc* Tukey honestly significant difference (HSD) test showed that all of the pairwise comparisons were significant (absolute difference $\geq$ 0.02, adjusted $p \leq 0.003$**), apart from the combination of the shifted and rotated maskers with static *F*0 contours (difference $<$ –0.001, adjusted $p = 1$).

The general pattern emerging from this analysis is that the degree of periodicity of the inharmonic maskers falls in between those of the harmonic equivalents and speech-shaped noise (included in Fig. 4 for comparison). Additionally, maskers with dynamic *F*0 contours were overall found to be somewhat less periodic than their static *F*0 equivalents, and this difference was most pronounced for the rotated maskers.

C. Masker envelope modulations

To provide an explanation for the deviating pattern of SRTs observed for the rotated maskers with a dynamic *F*0, the envelope modulations of the maskers were analysed and

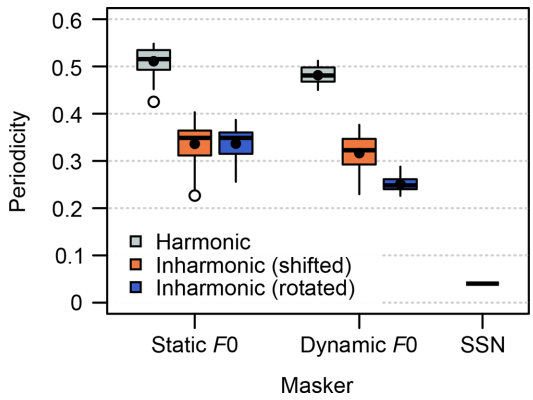

FIG. 4. (Color online) Masker periodicity. The values on the *y* axis indicate the degree of periodicity, or pitch strength, of the different maskers. The details of the boxplots are the same as those in Fig. 2. SSN, Speech-shaped noise.

https://doi.org/10.1121/10.0017065

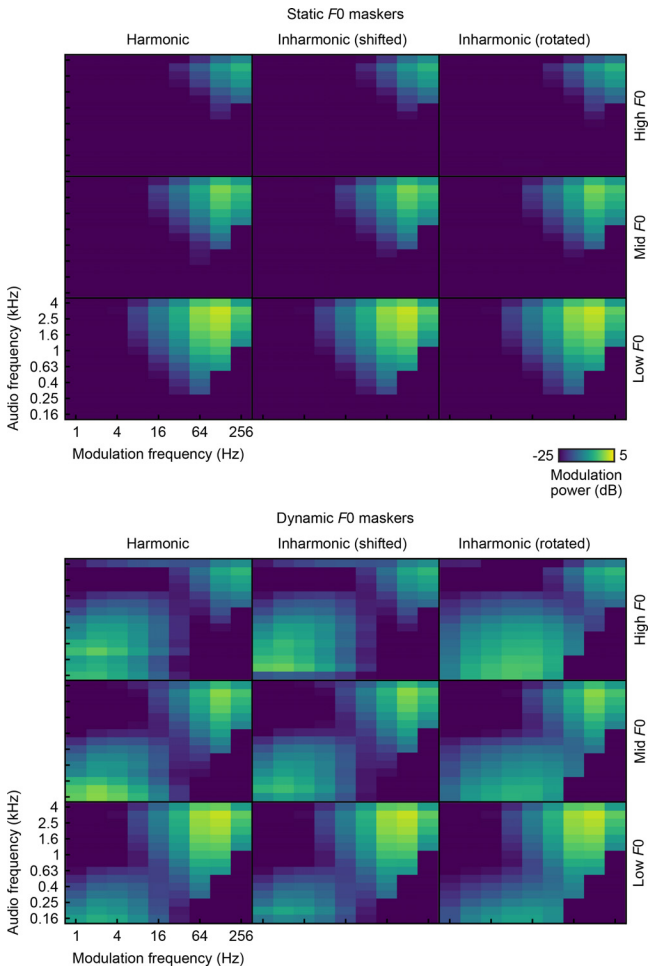

FIG. 5. (Color online) Masker envelope modulation spectrograms. (Upper) The average envelope modulation power for the maskers with static $F0$ contours. (Lower) The results for the maskers with dynamic $F0$ contours. The modulation power was computed for each combination of auditory (x axes) and modulation filter (y axes) using the front end of the mr-sEPSM speech intelligibility model (Jørgensen *et al.*, 2013).

compared. Using the front end of the mr-sEPSM speech intelligibility model (Jørgensen *et al.*, 2013), the modulation power for each combination of auditory (160–4000 Hz, 0.25-octave spacing) and modulation (1–256 Hz, one-octave spacing) filter channels was computed. The resulting envelope modulation spectrograms, averaged over all of the stimuli of the respective masker, are provided in Fig. 5. Briefly summarised, the static $F0$ maskers primarily have modulation power in the $F0$ region alone, whereas for the maskers with dynamic $F0$ contours, spectral components sweeping through low-frequency auditory filters create low-frequency modulations in addition to those in the $F0$ region. As expected, very similar results for all three masker types (harmonic, shifted, and rotated) were observed for the static $F0$ maskers. The remainder of this section thus focusses on the differences between the dynamic $F0$ maskers.

For simplicity, the modulation spectrograms of the dynamic $F0$ maskers were first averaged over all of the auditory filters, yielding the modulation spectra shown in Fig. 6. These envelope modulation spectra were then statistically analysed using an ANOVA with the factors *masker harmonicity*, *masker F0 level*, and *masker envelope modulations* (low, mid, or high). For the latter factor, the modulation power was averaged over filters tuned to low (1–8 Hz, $n = 4$), mid (16–64 Hz, $n = 3$), or high (128–256 Hz, $n = 2$) modulation frequencies. All of the main effects as well as all of the interactions were highly significant ($F \geq 14.75$, $p < 0.001$***). More interestingly, however, a *post hoc* Tukey HSD test confirmed that the modulation spectra of the harmonic and shifted maskers did not differ (difference = 0.03 dB, adjusted $p = 0.898$). This demonstrates that these two masker types only varied regarding their harmonicity, as intended. This test also showed that the modulation power at low frequencies increased with $F0$ level for all three masker types (low $F0$ to mid $F0$, difference = 2.6 dB; mid $F0$ to high $F0$, difference = 3.6 dB; adjusted $p < 0.001$*** in both cases).

Moreover, the modulation spectra of the rotated maskers differed markedly from the other two masker types. First, they had significantly less modulation power at low modulation rates than their harmonic (difference = −1.9 dB, adjusted $p < 0.001$***) and shifted (difference = −1.4 dB, adjusted $p < 0.001$***) equivalents. At intermediate rates, in contrast, the modulation power exceeded that of the harmonic (low $F0$ difference = 1.9 dB; mid, $F0$ difference = 3.5 dB; high $F0$ difference = 5.4 dB; adjusted $p < 0.001$*** in all of the cases) and shifted maskers (low $F0$ difference = 1.7 dB; mid $F0$ difference = 3.7 dB; high $F0$ difference = 5.1 dB; adjusted $p < 0.001$*** in all of the cases), and this effect increased with increasing $F0$ level. This difference appears to be the main reason for the unusual timbre and increased masking effectiveness of the rotated maskers. Last, neither the shifted (difference = 0.2 dB, adjusted $p = 0.601$) nor the rotated (difference = 0.03 dB, adjusted $p = 1$) maskers differed from the harmonic maskers regarding the envelope modulation power at high frequencies.

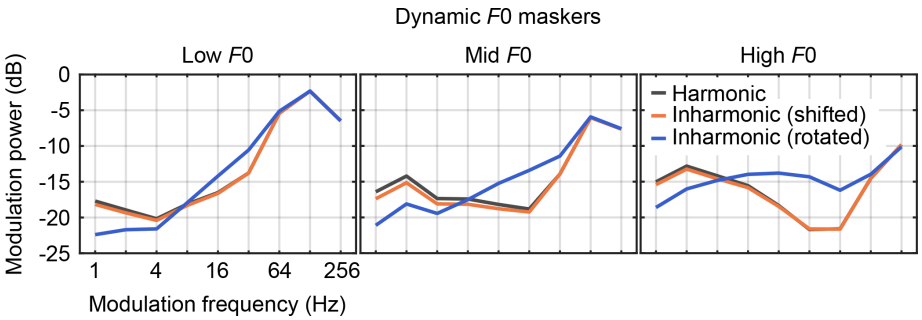

FIG. 6. (Color online) Masker envelope modulations. Envelope modulation spectra of the dynamic $F0$ maskers. For each masker, the average envelope modulation power for each modulation filter was computed across the entire set of stimuli using the front end of the mr-sEPSM speech intelligibility model (Jørgensen *et al.*, 2013).

IV. DISCUSSION

A. The rotated dynamic *F*0 maskers are outliers in terms of their acoustic properties

The acoustic analyses show that the envelope modulations and periodicity of the inharmonic maskers were very similar to those exhibited by their harmonic counterparts, with the exception of the rotated dynamic *F*0 maskers. In agreement with the primary aim of the present study, these materials thus allowed for an unbiased evaluation of the theory of harmonic cancellation.

Regarding the rotated dynamic *F*0 maskers, the modulation analysis showed increased modulation power at intermediate modulation rates (16–64 Hz) relative to the other maskers, an effect that increased with masker *F*0 level. This arises because the higher harmonics in the dynamic *F*0 maskers traverse a larger range of frequencies over a given time than the lower harmonics. Consider, for example, the *F*0 changing from 100 to 150 Hz over 200 ms, a change of 50 Hz. In a higher frequency region, say the 25th harmonic, this component will now be sweeping from 2.5 to 3.75 kHz, a change of 1250 Hz over the same time. After spectral rotation, this rapidly changing component will be at lower frequencies, sweeping through narrower auditory filters and thus resulting in increased envelope modulations (cf. Fig. 1). The increase in these modulations results in the rotated dynamic *F*0 maskers having a notably rougher timbre, which may be a limiting factor in stream segregation.

Modulations at these rates feature prominently in human screams (Arnal *et al.*, 2015), whereas the modulation spectrum of natural speech contains a dip in this region (see supplementary Fig. 1[1]). Moreover, these maskers were less periodic than the shifted equivalents, i.e., more inharmonic, and the increased spectral modulations of the lower harmonics may also have interfered with the ability to spectrally glimpse portions of the target speech. In summary, although spectral regularity was preserved in these maskers, as in all of the other inharmonic maskers used here, the spectral rotation not only affected their harmonicity but a range of other acoustic properties too.

B. No evidence for harmonic cancellation

Apart from the rotated dynamic *F*0 maskers, the behavioural data revealed no significant differences between the harmonic and inharmonic maskers, with SRTs differing by no more than about 1 dB throughout. Although these differences might have reached significance with a larger sample size, the implications of this finding would have remained marginal due to the small effect size. In stark contrast to the theory of harmonic cancellation (de Cheveigné, 2021), the current results thus imply that masker harmonicity *per se* does not affect the intelligibility of masked speech, irrespective of whether the maskers had static or dynamic *F*0 contours. The present findings therefore also render the previously used term of MPB inaccurate (Steinmetzger and Rosen, 2015, 2018).

Overall, SRTs for the maskers with dynamic *F*0 contours were higher than those for the static *F*0 equivalents, in line with previous findings (Deroche and Culling, 2011; Leclère *et al.*, 2017; Prud'homme *et al.*, 2022b), and performance also improved to a lesser extent with increasing *F*0 levels. As maskers with *F*0 modulations are slightly less harmonic than maskers with static *F*0 contours (cf. Fig. 4), the former effect has been taken as evidence for harmonic cancellation (Deroche and Culling, 2011). However, both findings can also be explained with increased modulation masking, as the dynamic *F*0 maskers have additional low-frequency modulations compared to the static *F*0 maskers (at 1–8 Hz, Fig. 5). This difference increased with the masker *F*0 level and thus counteracts the greater spectral glimpsing opportunities.

The current results are, furthermore, in contrast to those of Deroche *et al.* (2014b), who observed a small benefit from masker harmonicity when comparing speech intelligibility in the presence of harmonic and jittered inharmonic tone complexes. As demonstrated in that paper, jittering the masker harmonics resulted in greater spectral glimpsing opportunities compared to harmonic maskers. To rule out this confound, they used inharmonic complexes in which only the higher partials were jittered while the frequencies of the resolved lower partials were left unchanged. Nevertheless, this does not exclude the possibility that the presumed effect of masker harmonicity was due to the altered envelope modulations of the jittered maskers. Harmonic maskers have pronounced envelope modulations at the respective *F*0 rate, and these modulations are evident across auditory filters (cf. Fig. 5). Jittered maskers, in contrast, have far fewer *F*0-related modulations as the partials lack a common *F*0. Instead, the jittering introduces modulations at various rates that are not evenly distributed across auditory filters. This irregularity likely increases the effectiveness of the jittered maskers, an explanation that has been termed the "envelope modulation rate variability hypothesis" (Treurniet and Boucher, 2001). Rather than representing a genuine effect of masker harmonicity, the results of Deroche *et al.* (2014b) may therefore reflect the increased modulation masking caused by the jittered maskers.

Another study reporting a benefit from harmonicity using jittered stimuli is the study by Popham *et al.* (2018). However, as mentioned in the Introduction, the simultaneous jittering of target and background speech leaves the respective contributions of target and masker harmonicity unclear. Their results thus provide no explicit test of the theory of harmonic cancellation, whose central assumption is that masker harmonicity substantially aids the segregation of target and interferer while harmonicity of the target signal matters little. Unlike Deroche *et al.* (2014b), they furthermore used jittered speech stimuli in which spectral glimpsing opportunities and envelope modulations varied relative to unprocessed speech, further limiting the implications of the results.

https://doi.org/10.1121/10.0017065

C. Spectral regularity rather than harmonicity

Harmonic cancellation is an appealing, biologically plausible theory that is computationally simple and easily implementable in auditory models via comb filtering (de Cheveigné, 2021; Prud'homme *et al.*, 2020, 2022a). Nevertheless, the current data argue against the notion that the harmonicity of a masker aids its segregation from a target speech signal. It can furthermore be assumed that this finding will also apply to nonspeech target stimuli if the acoustic confounds described in Sec. IV B were precluded.

Instead of harmonicity, the regular spacing of discrete spectral components, which was maintained in the harmonic and inharmonic maskers used here, appears to be the crucial acoustic property explaining their reduced masking effectiveness relative to speech-shaped noise. This explanation was originally introduced by Roberts and Brunstrom (1998, 2001) to account for their finding that the ability to determine the pitch of a mistuned partial did not depend on the harmonicity of the tone complex. They observed that the spectral fusion of a sound into a single auditory object persisted in case of frequency shifted or stretched inharmonic complexes with preserved spectral regularity. Hence, the presence of a salient pitch was not required for auditory grouping, which is in line with the current results.

An attempt to reconcile the theory of harmonic cancellation with the results obtained with inharmonic maskers in which spectral regularity is preserved is the idea of *local* harmonic cancellation (de Cheveigné, 2021, p. 9). For any harmonic, a local $F0$ can be determined such that the neighbouring partials approach a harmonic series. However, even for the two immediately adjacent harmonics, this approach only works approximately, and the discrepancies increase for more distant partials. Given the increase in auditory filter bandwidths with increases in stimulus level, and the general broader tuning of higher frequency auditory filters that are most important for intelligibility, it appears unlikely that local harmonic cancellation can account for the current findings. There is also the question of the extent to which a masker with a dynamic $F0$ contour can have its $F0$ determined accurately enough to allow cancellation to operate. Answers to all of these questions require quantitative predictions from more computationally complete models.

V. CONCLUSION

Periodic sounds with a harmonic spectral structure, including voiced speech, are far less effective maskers of a target speech signal than noise. It has long been assumed that harmonicity, as such, is the crucial factor explaining this benefit, as put forward in the theory of harmonic cancellation. Using inharmonic complex tone maskers that were otherwise acoustically similar to their harmonic counterparts, the present study provided clear evidence against this theory as masker harmonicity had no effect on speech intelligibility. Because inharmonic sounds inevitably do not have a clear pitch, these results also demonstrate that the intuitive assumption that pitch information aids auditory grouping did not apply here. In previous studies claiming a benefit from masker harmonicity, the increased modulation masking caused by the inharmonic maskers was a key factor that was not controlled for. Consequently, it appears that none of the masked-speech studies arguing in favour of harmonic cancellation reported a genuine effect of harmonicity, thus explaining the discrepancies with the present results. Instead of harmonicity, it is suggested that spectral regularity is the acoustic property causing the reduced masking effectiveness of sounds with discrete spectral components.

ACKNOWLEDGMENTS

We would like to thank Roy Patterson for helpful advice regarding the analysis of the periodicity of the maskers, Trevor Agus for enlightening discussions about the modulation properties of the rotated dynamic $F0$ maskers, and Patrick Boyle of Advanced Bionics for providing the target sentence materials. This study has been funded with support from the European Commission under Contract No. FP7-PEOPLE-2011-290000 and the Dietmar Hopp Stiftung (Grant No. 2301 1239).

[1]See supplementary material at https://www.scitation.org/doi/suppl/10.1121/10.0017065 for the envelope modulation spectrogram and spectrum of the target speech materials.

Arnal, L. H., Flinker, A., Kleinschmidt, A., Giraud, A.-L., and Poeppel, D. (**2015**). "Human screams occupy a privileged niche in the communication soundscape," Curr. Biol. **25**, 2051–2056.

Blesser, B. (**1972**). "Speech perception under conditions of spectral transformation: I. Phonetic characteristics," J. Speech Hear. Res. **15**, 5–41.

Boyle, P. J., Nunn, T. B., O'Connor, A. F., and Moore, B. C. (**2013**). "STARR: A speech test for evaluation of the effectiveness of auditory prostheses under realistic conditions," Ear Hear. **34**, 203–212.

Bronkhorst, A. W., and Plomp, R. (**1992**). "Effect of multiple speechlike maskers on binaural speech recognition in normal and impaired hearing," J. Acoust. Soc. Am. **92**, 3132–3139.

Brungart, D., Simpson, B. D., Ericson, M. A., and Scott, K. R. (**2001**). "Informational and energetic masking effects in the perception of multiple simultaneous talkers," J. Acoust. Soc. Am. **110**, 2527–2538.

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.

de Boer, E. (**1956**). "Pitch of inharmonic signals," Nature **178**, 535–536.

de Cheveigné, A. (**2021**). "Harmonic cancellation—A fundamental of auditory scene analysis," Trends Hear. **25**, 233121652110414.

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. (**1997**). "Concurrent vowel identification. II. Effects of phase, harmonicity, and task," J. Acoust. Soc. Am. **101**, 2848–2856.

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.

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.
Guest, D. R., and Oxenham, A. J. (**2019**). "The role of pitch and harmonic cancellation when listening to speech in harmonic background sounds," J. Acoust. Soc. Am. **145**, 3011–3023.
Jørgensen, S., Ewert, S. D., and Dau, T. (**2013**). "A multi-resolution envelope-power based model for speech intelligibility," J. Acoust. Soc. Am. **134**, 436–446.
Leclère, T., Lavandier, M., and Deroche, M. L. (**2017**). "The intelligibility of speech in a harmonic masker varying in fundamental frequency contour, broadband temporal envelope, and spatial location," Hear. Res. **350**, 1–10.
Meddis, R., and Hewitt, M. J. (**1991**). "Virtual pitch and phase sensitivity of a computer model of the auditory periphery. I: Pitch identification," J. Acoust. Soc. Am. **89**, 2866–2882.
Meddis, R., and O'Mard, L. (**1997**). "A unitary model of pitch perception," J. Acoust. Soc. Am. **102**, 1811–1820.
Moore, B. C., and Gockel, H. E. (**2012**). "Properties of auditory stream formation," Philos. Trans. R. Soc. B **367**, 919–931.
Oxenham, A. J. (**2008**). "Pitch perception and auditory stream segregation: Implications for hearing loss and cochlear implants," Trends Amplif. **12**, 316–331.
Patterson, R. D. (**1973**). "The effects of relative phase and the number of components on residue pitch," J. Acoust. Soc. Am. **53**, 1565–1572.
Popham, S., Boebinger, D., Ellis, D. P., Kawahara, H., and McDermott, J. H. (**2018**). "Inharmonic speech reveals the role of harmonicity in the cocktail party problem," Nat. Commun. **9**, 2122.
Prud'homme, L., Lavandier, M., and Best, V. (**2020**). "A harmonic-cancellation-based model to predict speech intelligibility against a harmonic masker," J. Acoust. Soc. Am. **148**, 3246–3254.
Prud'homme, L., Lavandier, M., and Best, V. (**2022a**). "A dynamic binaural harmonic-cancellation model to predict speech intelligibility against a harmonic masker varying in intonation, temporal envelope, and location," Hear. Res **426**, 108535.
Prud'homme, L., Lavandier, M., and Best, V. (**2022b**). "Investigating the role of harmonic cancellation in speech-on-speech masking," Hear. Res **426**, 108562.
Roberts, B., and Brunstrom, J. M. (**1998**). "Perceptual segregation and pitch shifts of mistuned components in harmonic complexes and in regular inharmonic complexes," J. Acoust. Soc. Am. **104**, 2326–2338.
Roberts, B., and Brunstrom, J. M. (**2001**). "Perceptual fusion and fragmentation of complex tones made inharmonic by applying different degrees of frequency shift and spectral stretch," J. Acoust. Soc. Am. **110**, 2479–2490.
Roberts, B., Holmes, S. D., Darwin, C. J., and Brown, G. J. (**2010**). "Perception of concurrent sentences with harmonic or frequency-shifted voiced excitation: Performance of human listeners and of computational models based on autocorrelation," in *The Neurophysiological Bases of Auditory Perception*, edited by E. A. Lopez-Poveda, A. R. Palmer, and R. Meddis (Springer, New York), pp. 521–531.
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.
Schouten, J. F., Ritsma, R., and Cardozo, B. L. (**1962**). "Pitch of the residue," J. Acoust. Soc. Am. **34**, 1418–1424.
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.
Steinmetzger, K., and Rosen, S. (**2015**). "The role of periodicity in perceiving speech in quiet and in background noise," J. Acoust. Soc. Am. **138**, 3586–3599.
Steinmetzger, K., and Rosen, S. (**2018**). "The role of envelope periodicity in the perception of masked speech with simulated and real cochlear implants," J. Acoust. Soc. Am. **144**, 885–896.
Steinmetzger, K., Zaar, J., Relaño-Iborra, H., Rosen, S., and Dau, T. (**2019**). "Predicting the effects of periodicity on the intelligibility of masked speech: An evaluation of different modelling approaches and their limitations," J. Acoust. Soc. Am. **146**, 2562–2576.
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.
Treurniet, W. C., and Boucher, D. R. (**2001**). "A masking level difference due to harmonicity," J. Acoust. Soc. Am. **109**, 306–320.
Yost, W. A., Patterson, R., and Sheft, S. (**1996**). "A time domain description for the pitch strength of iterated rippled noise," J. Acoust. Soc. Am. **99**, 1066–1078.