Predicting the effects of periodicity on the intelligibility of masked speech: An evaluation of different modelling approaches and their limitations

Kurt Steinmetzger,[1,a)] Johannes Zaar,[2] Helia Relaño-Iborra,[2] Stuart Rosen,[1] and Torsten Dau[2]

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

[2]*Hearing Systems Section, Department of Health Technology, Technical University of Denmark, DK-2800 Kgs. Lyngby, Denmark*

(Received 27 February 2019; revised 27 August 2019; accepted 20 September 2019; published online 21 October 2019)

Four existing speech intelligibility models with different theoretical assumptions were used to predict previously published behavioural data. Those data showed that complex tones with pitch-related periodicity are far less effective maskers of speech than aperiodic noise. This so-called *masker-periodicity benefit* (MPB) far exceeded the *fluctuating-masker benefit* (FMB) obtained from slow masker envelope fluctuations. In contrast, the normal-hearing listeners hardly benefitted from periodicity in the target speech. All tested models consistently underestimated MPB and FMB, while most of them also overestimated the intelligibility of vocoded speech. To understand these shortcomings, the internal signal representations of the models were analysed in detail. The best-performing model, the correlation-based version of the speech-based envelope power spectrum model (sEPSM^{corr}), combined an auditory processing front end with a modulation filterbank and a correlation-based back end. This model was then modified to further improve the predictions. The resulting second version of the sEPSM^{corr} outperformed the original model with all tested maskers and accounted for about half the MPB, which can be attributed to reduced modulation masking caused by the periodic maskers. However, as the sEPSM^{corr}2 failed to account for the other half of the MPB, the results also indicate that future models should consider the contribution of pitch-related effects, such as enhanced stream segregation, to further improve their predictive power.
https://doi.org/10.1121/1.5129050

[ICB] Pages: 2562–2576

I. INTRODUCTION

Computational models that attempt to objectively predict the intelligibility of noise-corrupted or acoustically degraded speech signals have a long history (French and Steinberg, 1947; Kryter, 1962) and also received much attention in the recent past (e.g., Jørgensen *et al.*, 2013; Taal *et al.*, 2011). Although the common feature of such models is that they do not possess any linguistic knowledge but rely solely on analyses and transformations of the acoustic input, the processing steps vary widely across different models. For example, while the auditory processing stage (*front end*) of several models focusses on the envelope modulations of the stimulus materials, many others do not consider them explicitly. Part of the latter category are early speech intelligibility models, such as *articulation index* (AI; French and Steinberg, 1947) and *speech intelligibility index* (SII; ANSI S3.5, 1997), as well as their successors (e.g., *extended* SII; Rhebergen *et al.*, 2006). Modulation-based models, on the other hand, date back to the *speech transmission index* (STI; Steeneken and Houtgast, 1980) and also include the more recent modulation filterbank models, for example, the *multi-resolution speech-based envelope power spectrum model* (mr-sEPSM; Jørgensen *et al.*, 2013). In addition, it is useful to distinguish between models whose decision stage (*back end*) evaluates energetic differences between the input signals (e.g., ESII and mr-sEPSM) and those that are based on signal correlations, such as the *short-time objective intelligibility measure* (STOI; Taal *et al.*, 2011) and the *correlation-based version of the* mr-sEPSM (sEPSM^{corr}; Relaño-Iborra *et al.*, 2016).

One general difficulty in assessing speech intelligibility models is that they are usually devised to perform well in a specific set of conditions and, hence, the range of stimulus materials with which they are evaluated tends to be limited. Moreover, the materials often vary considerably, both across models, as well as regarding their acoustic properties across testing conditions, which makes it difficult to compare different models and understand their shortcomings. Consequently, testing a set of existing speech intelligibility models with a common data set obtained with materials that systematically vary with respect to certain relevant acoustic features appears to be a fruitful approach to compare, challenge, and further improve them (for a similar approach, see Schubotz *et al.*, 2016, and Van Kuyk *et al.*, 2018). At the same time, modelling experimental data also serves to

[a)]Current address: Section of Biomagnetism, Department of Neurology, Heidelberg University Hospital, Im Neuenheimer Feld 400, 69120 Heidelberg, Germany. Electronic mail: kurt.steinmetzger.12@ucl.ac.uk

gain a better understanding of them, particularly by examining the internal signal representations generated by models with different theoretical assumptions.

Following this line of thought, the current study is based on the stimuli and data described in Steinmetzger and Rosen (2015), where speech materials and maskers were introduced that vary regarding the amount of acoustic periodicity. Periodicity here denotes that a speech sound is voiced, as opposed to unvoiced (i.e., aperiodic). Vocoders that allowed the choice between a voiced or unvoiced source excitation were used to synthesise speech that is either (i) completely aperiodic (noise-vocoded), (ii) preserves the fundamental-frequency ($F0$) contours of the original recordings (Dudley-vocoding; Dudley, 1939), or (iii) is rendered completely periodic using interpolated versions of the original $F0$ contours ($F0$-vocoding).[1] Likewise, the maskers were either aperiodic (speech-shaped noise) or periodic (harmonic complexes with dynamically varying $F0$ contours derived from natural speech). These materials were constructed to test the hypothesis that periodicity helps to segregate a speech signal from a masker. The results showed that performance, as measured by tracking speech reception thresholds (SRTs; Plomp and Mimpen, 1979) at the 50%-correct level, was substantially better when the masker was periodic, while the normal-hearing listeners hardly benefitted from periodicity in the target speech. The former finding was termed the *masker-periodicity benefit* (MPB). Furthermore, both the periodic and the aperiodic maskers were presented in a steady-state version or were sinusoidally amplitude-modulated at a rate of 10 Hz to enable a *fluctuating-masker benefit* (FMB; Festen and Plomp, 1990). However, a substantial FMB required the target speech to have a very high intelligibility in quiet, and the effect was generally much smaller than the MPB.

Four existing speech intelligibility models with different theoretical assumptions were used to predict these behavioural data: *ESII* (Rhebergen *et al.*, 2006), *STOI* (Taal *et al.*, 2011), *mr-sEPSM* (Jørgensen *et al.*, 2013), and *sEPSM^{corr}* (Relaño-Iborra *et al.*, 2016). Briefly summarised, the ESII compares the energy of the envelopes of speech and the forward-masking corrected noise envelopes in each auditory filter using temporal windows whose durations decrease with increasing filter centre frequencies. These power estimates, which can be interpreted as indices of short-term envelope audibility, are then averaged over time and auditory filters, where the contribution of each filter is determined by a predefined band-importance function. STOI, in contrast, compares envelopes of unprocessed speech and the mixture of speech and noise in each auditory filter by computing the cross-correlation of segments of the two signals with a fixed length of 384 ms. Here, the speech signal in the mixture may be unprocessed or processed, depending on the experimental condition. Instead of considering power differences, it effectively measures how much the original speech envelope is distorted by adding background noise or processing the target speech, for example, by vocoding. The unweighted average of the individual correlation coefficients across time segments and auditory filters is assumed to vary along with the intelligibility of the mixture. The mr-sEPSM uses the speech-noise mixture and the noise alone as input signals and, additionally, differs from the former two models in that it employs a modulation filterbank after the initial auditory filtering and envelope extraction. For each combination of auditory and modulation filters, the envelope signal-to-noise ratio (SNR_{env}) is calculated by subtracting the modulation power of the noise from that of the mixture, and dividing it by the modulation power of the noise

$$\mathrm{SNR_{env}} = \frac{P_{\mathrm{env},S+N} - P_{\mathrm{env},N}}{P_{\mathrm{env},N}}. \quad (1)$$

These SNR_{env} values are computed using temporal windows that are inversely proportional to the respective modulation rate, and their unweighted mean across time, auditory, and modulation filters is hypothesised to be positively related to speech intelligibility. Finally, the sEPSM^{corr} was included, a hybrid model combining the modulation-based front end of the mr-sEPSM with a correlation-based back end inspired by STOI. Here, unprocessed speech and the speech-noise mixture are used as inputs, as in STOI, and the outputs from each individual combination of the auditory and modulation filters are correlated using the same multi-resolution time windowing approach as in the mr-sEPSM.

As shown in Fig. 1, each of these models is characterised by a specific combination of front end and back end. The reasoning behind this selection is that, first, a substantial portion of the MPB is thought to arise from the absence of random envelope modulations and the sparser modulation spectrum of the periodic maskers, leading to a reduced amount of modulation masking (Stone *et al.*, 2011; Stone *et al.*, 2012). By including models with and without a modulation filterbank in the auditory processing front end, the contribution of modulation masking can thus be tested and quantified. Second, it is thought that a correlation-based decision back end is required to account for the reduced intelligibility of vocoded speech. The latter differs from unprocessed speech less in terms of the envelope power, but primarily regarding the contours of the subband envelopes, which should be reflected in a correlation-based comparison.

As will be shown below, the sEPSM^{corr}, which combines a front end with a modulation filterbank and a correlation-based back end, indeed produced the most accurate predictions. However, a detailed examination of this model also revealed a crucial limitation in that the low resolution of the extracted subband envelopes diminishes the differences between the set of maskers included in the current study. A solution to overcome this limitation is therefore also presented. Furthermore, it will be shown that modulation filters

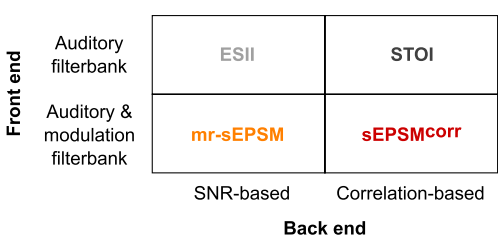

FIG. 1. (Color online) Modelling scheme. The four speech intelligibility models included in the current study were selected to enable a systematic evaluation of different auditory processing front ends and decision back ends.

in which unprocessed speech carries little relevant information, particularly those falling in between the slow envelope and voice pitch modulation ranges (i.e., ∼16–64 Hz), are not required to predict speech intelligibility. The resulting sEPSM^{corr}2 contains both of these modifications and further improved the predictions of the same set of data. It is described in more detail below.

II. sEPSM^{corr}2 MODEL DESCRIPTION

A. sEPSM^{corr} summary

First, we briefly summarise the processing steps of the original sEPSM^{corr}. A more detailed description, including all equations and the theoretical motivation for each step, is provided in Sec. II in Relaño-Iborra *et al.* (2016).

The model uses unprocessed speech as well as noisy or processed speech as input. In the first step of the preprocessing front end, both signals are filtered using 22 fourth-order gammatone filters with centre frequencies ranging from 63 Hz to 8 kHz and 1/3 octave spacing. Only filters for which the stimulus level of the unprocessed signal is above the hearing threshold are processed further. Next, a first-order low-pass filter with a cutoff at 150 Hz is applied to the Hilbert envelopes of the remaining filter outputs. The envelopes are then passed through a modulation filterbank comprising a third-order low-pass filter with a cutoff of 1 Hz and eight second-order band-pass filters with octave spacing, a constant quality factor of 1, and centre frequencies ranging from 2 to 256 Hz. For the outputs of modulation filters with centre frequencies above 10 Hz, another ("second-order") Hilbert envelope is then extracted and used for further processing, whereas the outputs of modulation filters with centre frequencies below 10 Hz are left unchanged. This processing step is meant to account for the diminished modulation-phase sensitivity of the human auditory system at higher modulation frequencies (Dau *et al.*, 1997a,b). Last, the resulting signals are logarithmically compressed.

In the decision back end of the model, the signals are processed in frequency-dependent time segments. The lengths of the non-overlapping rectangular windows are determined by the inverse of the respective modulation-filter centre frequency, ranging from 1 s at 1 Hz to 3.9 ms at 256 Hz. Modulation filter outputs with centre frequencies below one-fourth of a given auditory filter centre frequency are discarded. Next, the individual time segments of the two input signals are correlated, yielding a single correlation coefficient for each one. By replacing negative coefficients with 0, only positive correlations are considered for further processing. The time-integrated correlation for each combination of modulation filter and auditory filter is then obtained by taking the square root of the sum of the squared coefficients of the individual time segments. The final correlation metric χ is then derived by averaging the time-integrated correlations of all processed combinations of modulation filters and auditory filters. To relate the χ-values to speech intelligibility, a logistic function is used,

$$\Phi(\chi) = \frac{100}{1 + e^{(a \cdot \chi + b)}}, \tag{2}$$

where a and b are the free parameters, which have to be optimised during the calibration procedure.

B. Preserving the difference between aperiodic and periodic sounds

An analysis of the individual signal processing steps of the sEPSM^{corr} showed that the additional (second-order) Hilbert envelope extraction, which occurs immediately after the signals have been passed through the modulation filterbank, obliterates the difference between aperiodic and periodic sounds considerably. This processing step was, hence, omitted and replaced with full-wave rectification of the outputs of each modulation filter.

The differing effects of Hilbert envelope extraction and full-wave rectification of envelope-filtered signals are demonstrated in Fig. 2. The illustration is based on a portion of unprocessed speech, which is voiced during the first half but unvoiced during the second half [Fig. 2(A)], and the same portion of unprocessed speech mixed with a random segment of the steady periodic masker at an SNR of −5 dB. In the top panel of Fig. 2(B),[2] the outputs of the 2-kHz auditory filter

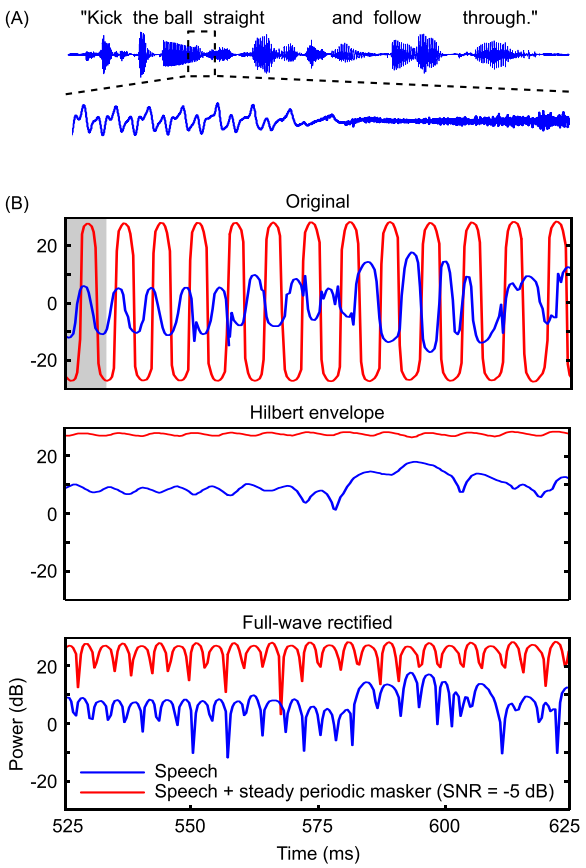

FIG. 2. (Color online) mr-sEPSM^{corr}2: Comparison of Hilbert envelope extraction and full-wave rectification of envelope-filtered signals. (A) shows the portion of unprocessed speech on which the example envelopes in (B) are based. These envelopes are the outputs of the 2-kHz auditory filter and the 128-Hz modulation filter for unprocessed speech (blue), and the same signal mixed with a random portion of the steady periodic masker (red, SNR = −5 dB). The grey bar indicates the length of the signal portions that are correlated, here, 7.8 ms (i.e., 1/128 Hz). In contrast to the Hilbert envelopes, full-wave rectification preserves the fine signal details present in the original envelopes.

and 128-Hz modulation filter are shown, the combination for which unprocessed speech has the strongest *F*0-related modulations [cf. Fig. 8(A)]. Crucially, the Hilbert envelopes of these two signals [Fig. 2(B), middle panel] are considerably smoother than the original signals shown above. Full-wave rectification [Fig. 2(B), bottom panel], on the other hand, preserves the fine signal details, particularly the difference between the periodic and aperiodic portions of the speech signal.

Furthermore, Fig. 2(B) shows that extracting the Hilbert envelope does not serve to discard phase information altogether, as intended. The first half of the example signals still shows some periodic fluctuations, albeit to a lesser degree. Full-wave rectification, in contrast, reduces the influence of phase differences not by attempting to flatten the signals, but by omitting the signal polarity.

C. Modulation filter selection algorithm

In a second step, it was examined whether modulation filters tuned to intermediate modulation rates (~16–64 Hz) are necessary to predict speech intelligibility, or if their exclusion even serves to improve predictions. This alteration is based on the theoretical consideration that speech carries no relevant linguistic information at these modulation frequencies (Arnal *et al.*, 2015; Joris *et al.*, 2004), neurophysiological evidence that there is little activity in response to these modulation rates in human auditory cortex (Giraud *et al.*, 2000), and the empirical observation that the speech modulation spectrum indeed shows a dip in this region (e.g., Fig. 7 in Steinmetzger and Rosen, 2017). Rather than discarding them *a priori*, an algorithm was developed that identifies and excludes any modulation filters from further processing for which the broadband modulation power of the unprocessed speech signal falls below a specified relative threshold.

Specifically, for each individual sentence, the Hilbert envelope of the unfiltered waveform was extracted and passed through the same modulation filterbank deployed in the mr-sEPSM and sEPSM^{corr} models. The power in each modulation filter was then calculated by taking the mean of the squared filter output and dividing it by the mean of the squared broadband Hilbert envelope divided by two, as in the mr-sEPSM. The latter operation only affected the scaling, not the actual results. After dB-conversion, these power estimates were compared to a relative exclusion criterion, defined as the median of the power across all nine modulation filters minus half its standard deviation. This criterion proved to be effective in selectively excluding modulation filters tuned to intermediate modulation rates. If the power at the output of a modulation filter fell below this exclusion criterion, the contribution of the corresponding modulation filter band to speech intelligibility was assumed to be negligible. As can be seen in Fig. 3, which shows the results of the algorithm for the first 100 IEEE sentences, the power in filters tuned to intermediate modulation frequencies (16, 32, and 64 Hz) mostly fell below the relative threshold.

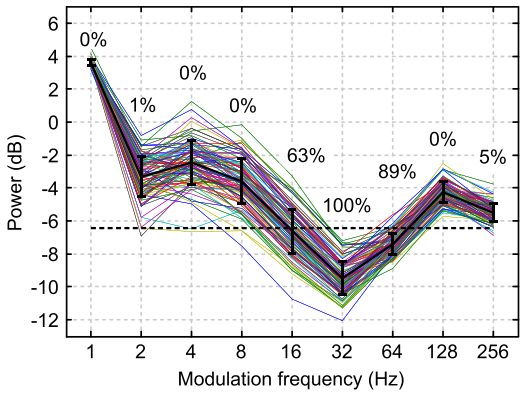

FIG. 3. (Color online) mr-sEPSM^{corr}2 modulation filter selection algorithm. The thin colored lines show the broadband modulation spectra of the first 100 IEEE sentences, and the overlaid black line shows the average and standard deviations. The percentages indicate how often a modulation filter was excluded because its power fell below the threshold value, defined as the median power across all nine modulation filters minus half its standard deviation. The black dotted line shows the average threshold value across all 100 sentences.

III. METHODS

A. Stimuli

Out of the 48 combinations of target speech and masker included in Steinmetzger and Rosen (2015), 8 were used in the current study. First, unprocessed speech combined with each of the four maskers (periodic and aperiodic, both steady and 10-Hz modulated) was included, as these conditions resulted in the largest MPBs and FMBs. Second, four types of vocoded speech mixed with steady noise were selected (noise-vocoded speech with 7 and 12 channels, Dudley-vocoded speech with 7 and 10 channels) to test whether a given model can account for the lowered intelligibility of vocoded speech and whether the predictions vary along with the number of vocoder channels. Examples of each individual target speech condition and masker are shown in Fig. 4.

All target speech conditions are based on recordings of the IEEE sentence corpus (Rothauser *et al.*, 1969) spoken by an adult male talker with a Southern British English accent and a mean *F*0 of 121.5 Hz. Using a channel vocoder implemented in MATLAB (MathWorks, Natick, MA), the original recordings were noise-vocoded by filtering them into the desired number of frequency bands (zero-phase shift, sixth-order Butterworth), based on equal basilar membrane distance (Greenwood, 1990) across a range of 0.1–11 kHz. Filter outputs were full-wave rectified and low-pass filtered at 30 Hz to extract the amplitude envelope (zero-phase shift, fourth-order Butterworth). The subband envelopes were then multiplied with a white noise and again band-pass filtered, as specified above. Before summing the signal together, the root-mean-square (RMS) level of each band was adjusted to that of the original band. The final waveforms were then low-pass filtered at 10 kHz (sixth-order elliptic). Dudley-vocoding was performed using the same routine, except that a pulse train following the original *F*0 contour was used as carrier signal instead of white noise when the original recording was voiced. *F*0 contours were extracted using the

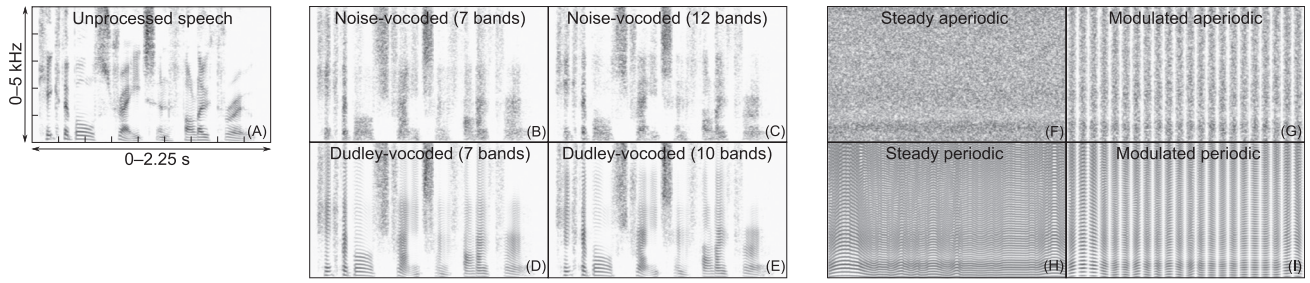

FIG. 4. Stimuli. Narrow-band spectrograms of examples of the five target speech conditions (A)–(E) and four maskers (F)–(I) used in the current study. The example sentence shown is IEEE 0204 ("Kick the ball straight and follow through.").

PRAAT (Boersma and Weenink, 2013) script ProsodyPro 4.3 (Xu, 2013) and hand-corrected where needed.

Aperiodic noise maskers were based on a 23.8-s passage of white noise, filtered to have the long-term average speech spectrum (LTASS; Byrne *et al.*, 1994). Periodic harmonic complex maskers were derived from recordings of the EUROM database (Chan *et al.*, 1995). Sixteen different male talkers reading five- to six-sentence passages with similar accent, speaking rate, and voice quality as the target talker were chosen. The *F*0 contours of these passages were interpolated through unvoiced and silent periods to synthesise the waveforms of the harmonic complexes on a period-by-period basis using the Liljencrants–Fant model (Fant *et al.*, 1985). The resulting complexes were then also filtered to have the LTASS and concatenated into a single file with a duration of 362.2 s. The aperiodic, as well as the periodic, maskers were either presented as steady-state versions or sinusoidally amplitude modulated at a rate of 10 Hz and with a modulation depth of 100%.

B. Procedure

In addition to the SRTs taken from Steinmetzger and Rosen (2015), psychometric functions (PFs) were fitted to the adaptive tracks that formed the basis of these results to enable a detailed comparison with the model predictions. Based on the logistic-regression procedure introduced by Wichmann and Hill (2001), PFs were fitted to the results of each individual listener before the function parameters were averaged together to form the group-level PF. For conditions including unprocessed speech, threshold and slope were free parameters, while the lapse and guess rates were set to 0. For conditions including vocoded speech, the lapse rate was also a free parameter, as human performance did not reach 100% in quiet. Subjects with threshold or slope estimates identified as outliers in boxplots with a whisker length of three times the interquartile range were excluded. This criterion applied to no more than 2 out of 12 subjects per stimulus condition.

For each model, percentage correct scores at seven SNRs, ranging from −20 to +10 dB in 5-dB steps, were obtained to yield estimated PFs that could be compared to the human PFs. This SNR range was chosen as it covers most of the human performance range across conditions, while the 5-dB step size was considered sufficiently accurate to test whether the models can reproduce the shape of the human PFs. For the models, SRTs were determined by finding the 50%-point on the graphs through the predicted values at the seven SNRs. The behavioural SRTs, on the other hand, were the original SRTs measured in the adaptive procedure, which is why there are some conditions in which SRTs and PFs differ somewhat.

Model predictions were based on the mean results of the first 100 IEEE sentences with a total duration of 224.8 s. Human data and model predictions in each stimulus condition were compared, first, by subtracting the estimated from the human SRTs such that positive SRT prediction errors indicate an overestimation of human performance. Second, to analyse the steepness of the modelled and human PFs in addition to their horizonal shifts, the slopes values of the human PFs were subtracted from the modelled ones, such that positive slope errors indicate that the model estimates were too steep. Slopes were calculated as the average change in percentage correct per dB SNR for the middle portion (i.e., 40%–60% correct) of the modelled and human PFs. To summarise these two types of condition-specific predictions errors, root-mean-square errors (RMSEs) across all conditions were also calculated, a measure that gives more weight to large individual prediction errors than the mean.

Each model was calibrated by minimising the RMSEs of the model predictions across the seven SNRs for the combination of unprocessed speech and steady noise, henceforth referred to as the *reference condition*. Consequently, this condition was omitted from all model evaluations. The fitting parameters were obtained by non-linear least squares optimisation and kept constant throughout. To transform the model coefficients into percentage correct scores, the logistic function employed in the sEPSM^{corr} [see Eq. (2)] was also used for ESII, STOI, and sEPSM^{corr}2. For the mr-sEPSM, the original ideal observer transformation was used [see Eqs. (7) and (8) in Jørgensen and Dau, 2011]. Here, the best fit was achieved by keeping the values for open-set materials used in Jørgensen and Dau (2011) and Jørgensen *et al.* (2013) for parameters q and m, increasing σ_s from 0.6 to 1, and manually optimising k. All fitting parameters are summarised in Table I.

IV. RESULTS AND DISCUSSION

For unprocessed speech mixed with the four different maskers, human data and the predictions of all five models are shown in Fig. 5. The corresponding results for the four types of vocoded speech mixed with steady noise are shown in Fig. 6. The prediction errors for all stimulus conditions

TABLE I. Model calibrations: Transformation procedures, fitting parameters, and RMSEs across all seven SNRs in the reference condition.

Model	Transformation	a	b	q	m	σ_s	k	RMSE
ESII	Logistic function	−16.19	5.88	—	—	—	—	0.17%
STOI	Logistic function	−23.96	16.50	—	—	—	—	0.50%
mr-sEPSM	Ideal observer	—	—	0.5	8000	0.9	0.32	0.58%
sEPSM^{corr}	Logistic function	−7.06	23.18	—	—	—	—	0.90%
sEPSM^{corr}2	Logistic function	−6.63	17.11	—	—	—	—	0.74%

and models are shown in Fig. 7, separately for SRTs [Fig. 7(A)] and PF slopes [Fig. 7(B)].

Irrespective of the masker, all five models consistently underestimated the SRTs of the human listeners with unprocessed target speech [Fig. 7(A)], albeit to varying degrees. Thus, the maskers were always predicted to be more effective than they actually were. For vocoded target speech, the opposite pattern was observed, apart from sEPSM^{corr} and sEPSM^{corr}2. Hence, human performance with vocoded speech was mostly overestimated. Overall, out of the four previously published models, SRT prediction errors were smallest for the sEPSM^{corr}, while there was little difference between the other three models (ESII, STOI, and mr-sEPSM). Compared to the original version of the model, SRT predictions for the sEPSM^{corr}2 were about 2–3 dB more accurate for all conditions including unprocessed speech, but substantial prediction errors persisted for the periodic maskers. With vocoded target speech, on the other hand, the estimated SRTs were reasonably accurate and hardly differed between the two model versions.

The slope errors shown in Fig. 7(B), in contrast, were generally smaller for models with SNR-based (ESII and mr-sEPSM) rather than correlation-based back ends (STOI, sEPSM^{corr}, and sEPSM^{corr}2), as reflected by the RMSEs across conditions. Moreover, the error patterns for the individual conditions resembled each other for models with the same back end type. While all models mostly predicted PFs that were

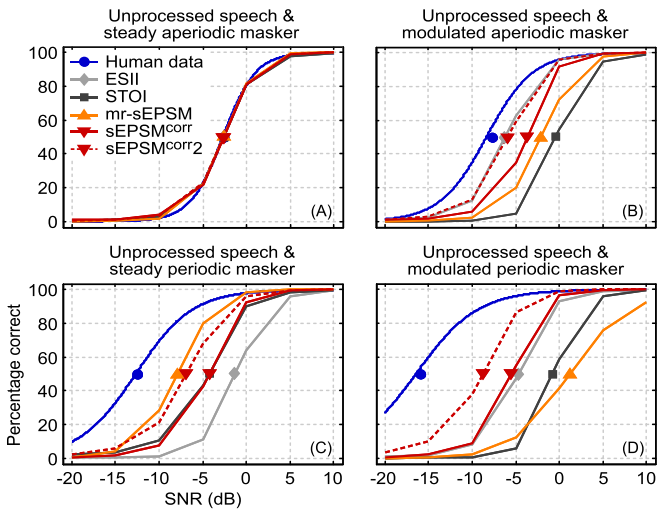

FIG. 5. (Color online) Human data and model predictions: unprocessed speech. SRTs and PFs for unprocessed speech mixed with four different maskers. SRTs are indicated by the symbols on top of the PFs. Unprocessed speech and steady noise served as reference condition for calibrating the models.

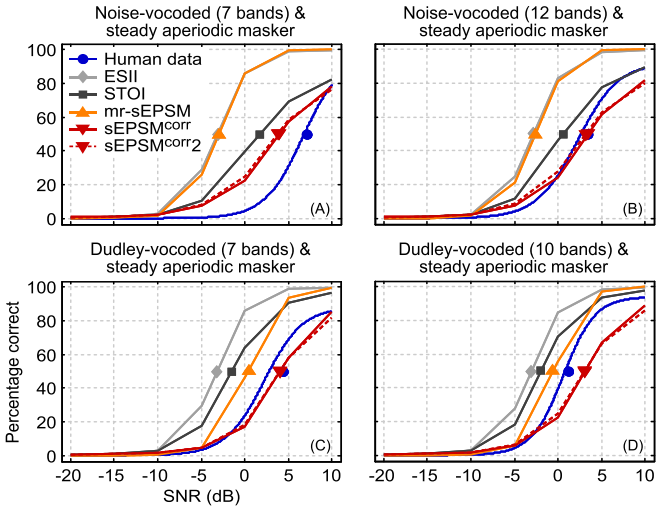

FIG. 6. (Color online) Human data and model predictions: vocoded speech. SRTs and PFs for four types of vocoded speech mixed with steady noise. As in Fig. 5, unprocessed speech and steady noise served as reference condition for calibrating the models.

steeper than the human equivalents for conditions including unprocessed speech, the slopes of the correlation-based models were much too shallow with vocoded target speech.

Below, the results of each model will be analysed and discussed in detail with a focus on the modulation-based models.

A. ESII

The ESII failed entirely to predict the MPB [Figs. 5(C) and 5(D)]. The predicted SRTs in the two conditions including periodic maskers were even about 1.5 dB higher than for

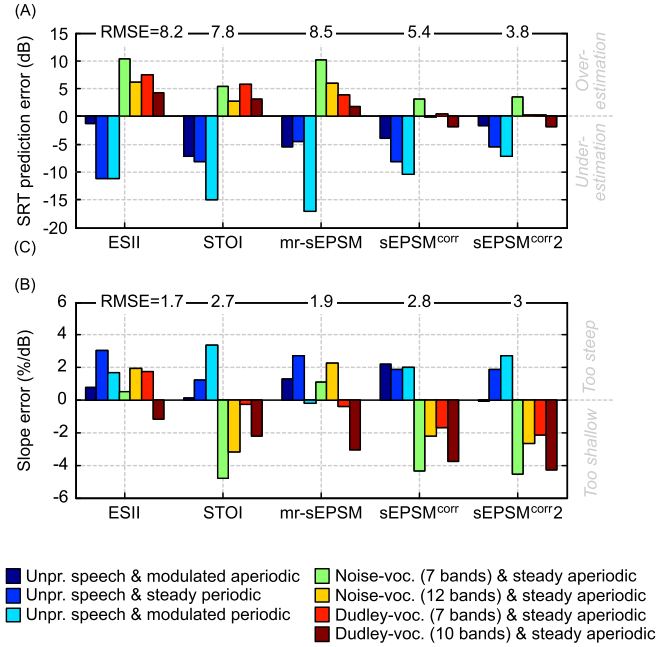

FIG. 7. (Color online) Model prediction errors. Deviations of the predicted SRTs (A) and the slopes of the predicted PFs (B) from the human data. For the SRTs, positive values indicate an overestimation of human performance. For the slopes, positive values indicate that predicted PFs were steeper than those of the human listeners. At the top of each panel, the RMSEs, averaged across all seven stimulus conditions, give an estimate of each model's overall performance.

the conditions including aperiodic maskers, making it the worst of the four tested models in this regard. The reason for this was found to be the envelope power estimation routine that precedes the forward-masking function, where the values of the intensity envelopes of each critical band are squared and averaged in 1-ms steps, resulting in slightly higher power estimates for the periodic maskers. In contrast, for the combination of unprocessed speech and modulated noise [Fig. 5(B)], the ESII correctly predicted a FMB and outperformed all models except the sEPSM^{corr}2 with a SRT estimate that was only 1.3 dB too high.

The ESII also failed to predict the decrease of intelligibility when the target speech was vocoded (Fig. 6). In all four conditions including vocoded speech, the predicted SRTs differed by no more than about 0.5 dB from the predicted SRT in the reference condition. As expected, envelope power differences alone, hence, do not account for the lower intelligibility of vocoded speech, irrespective of source periodicity and number of channels.

In contrast to the original studies (Rhebergen and Versfeld, 2005; Rhebergen *et al.*, 2006), the actual speech signals in each condition were used as input signals for the ESII, not speech-shaped noise. For conditions with unprocessed target speech, both model configurations were compared, and the differences were found to be marginal. The predicted SRTs never differed by more than 0.05 dB.

In summary, due to their similar power spectra, the ESII failed to differentiate between unprocessed speech, vocoded speech, and speech-shaped noise, although the measured speech intelligibility varied widely across these conditions.

B. STOI

STOI strongly underestimated the MPB with the effect amounting to only about 1.6 dB SRT for steady maskers and about 0.4 dB for modulated maskers [Figs. 5(C) and 5(D)]. This small correct trend can be explained by the absence of random modulations in the periodic maskers, which distort the mixture envelopes. The underestimation of this effect, in turn, appears to originate from the normalisation and clipping procedure that is applied to the envelopes of the mixture before calculating the correlation. Specifically, the clipping algorithm, which is intended to limit the influence of periods during which the speech envelope is completely masked, discards a substantial portion of the subtle envelope differences between the aperiodic and periodic maskers. Consequently, some of the acoustic properties of the masker are not represented in STOI.

Instead of a FMB, STOI predicted a decreased speech intelligibility in conditions including 10-Hz modulated maskers with SRTs that were on average about 2.8 dB higher relative to the corresponding conditions including steady-state maskers [Figs. 5(B) and 5(D)]. This result was expected, as the duration of the envelope segments that are analysed is relatively long (384 ms). Compared to a steady masker, the correlation of speech and mixture envelopes will generally be higher during the troughs of a modulated masker. However, if the analysis windows are longer than the masker troughs, this gain is outweighed by the fact that the envelopes of the speech and the speech plus the modulated masker barely resemble each other overall.

When the target speech was vocoded (Fig. 6), the model predictions also showed a correct trend, but the diminished intelligibility was on average underestimated by about 4.3 dB SRT. Moreover, the predicted SRTs increased by less than 1 dB when the number of bands in the vocoder was lowered, compared to about 3.5 dB for the listeners. However, STOI was the only model which correctly predicted that SRTs were on average about 2.5 dB better for Dudley- compared to noise-vocoded speech. This finding can again be explained by the random modulations, which are more pronounced in noise-vocoded speech and thus distort the envelopes.

Finally, it should be mentioned that an updated version of the STOI, the *extended* STOI (ESTOI; Jensen and Taal, 2016), has been published recently with the explicit aim to also predict speech intelligibility with modulated maskers. Rather than comparing short envelope segments, the ESTOI considers the spectral correlation of the unprocessed speech signal and the mixture of target speech and noise, which appears to be an effective approach, too.

C. mr-sEPSM

The mr-sEPSM outperformed the other models by accounting for a large portion of the MPB when the maskers were steady [∼5.2 dB SRT; Figs. 5(A) and 5(C)]. On the other hand, it failed to account for the FMB by predicting almost no such effect for aperiodic maskers [∼0.5 dB SRT; Figs. 5(A) and 5(B)] as well as a substantial effect in the opposite direction for periodic maskers [∼–9.2 dB SRT; Figs. 5(C) and 5(D)].

For noise-vocoded target speech [Figs. 6(A) and 6(B)], the estimated SRTs differed by no more than about 0.3 dB from the reference condition, but for Dudley-vocoded speech [Figs. 6(C) and 6(D)] the model showed a correct trend toward lower speech intelligibility by predicting SRTs that were about 3.1 dB higher with seven channels and about 2.1 dB with ten channels.

As a starting point for discussing the results of the modulation-based models, modulation spectrograms of all target speech conditions and maskers used in the current study are shown in Fig. 8. These representations were generated by computing the envelope power, as implemented in the front end of the mr-sEPSM, for each combination of auditory and modulation filters, time-averaged over the entire set of stimulus materials in each condition (see also Steinmetzger and Rosen, 2018). Figure 9, in contrast, shows modulation spectra of the eight different *combinations* of target speech and masker at each of the seven SNRs used for model prediction. These were obtained by averaging the dB-scaled modulation power estimates in the modulation spectrograms across all auditory filters.

For the stimulus materials used in this study, the following types of modulations can be distinguished (cf. Joris *et al.*, 2004; Rosen, 1992): Unprocessed and Dudley-vocoded speech [Figs. 8(A), 8(D), and 8(E)] contain modulations (a) that result from the low frequency harmonics during

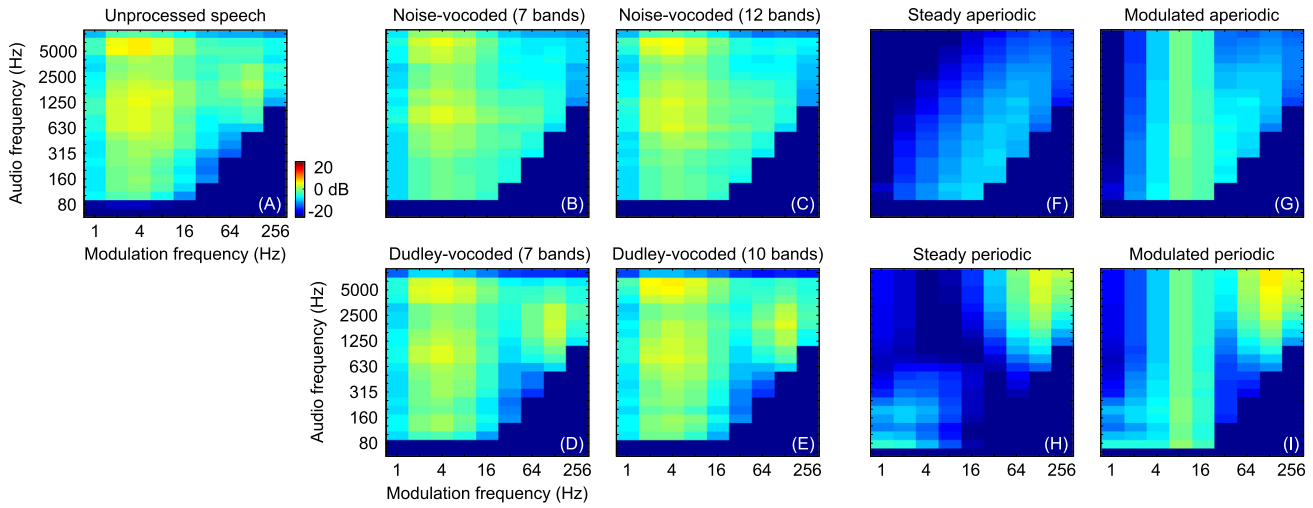

FIG. 8. (Color online) mr-sEPSM modulation spectrograms. Envelope modulation power of the five target speech conditions (A)–(E) and four maskers (F)–(I) used in the current study. For each combination of auditory (*y*-axis) and modulation filter (*x*-axis), the average modulation power across the entire set of stimulus materials was computed using the front end of the mr-sEPSM speech intelligibility model.

voiced speech sweeping through low-frequency, sharply tuned auditory filters due to their dynamic *F*0 contours (mainly in modulation filters with centre frequencies from 1 to 4 Hz), (b) at the word, syllable, and phoneme rate (∼2–8 Hz), and (c) at *F*0 frequencies (∼64–256 Hz). Noise-vocoded speech [Figs. 8(B) and 8(C)] only contains modulations of type (b). The periodic maskers [Figs. 8(H) and 8(I)] have modulations of types (a) and (c), whereas the aperiodic maskers [Figs. 8(F) and 8(G)] modulate randomly with the dominant modulation rates related to auditory filter bandwidths. Last, if they are amplitude-modulated [Figs. 8(G) and 8(I)], both maskers also show prominent 10-Hz modulations in addition to their original modulation profiles.

In the back end of the mr-sEPSM it is assumed that the smaller the modulation power of the noise, relative to the target speech (and the interaction component of speech and noise), the higher the predicted speech intelligibility [cf. Eq. (1)]. As can be seen in Figs. 8 and 9, while the steady versions of the maskers indeed have markedly less modulation energy than unprocessed speech, this is not the case for the modulated maskers. The superimposed 10-Hz modulations result in a modulation pattern that, to some extent, resembles that of speech, particularly in the case of the modulated periodic masker. These observations match the model predictions shown in Fig. 5, which show a correct trend toward better speech intelligibility with the steady periodic masker but are poor for both modulated maskers.

For vocoded target speech, it was found that lowering the number of channels slightly altered the distribution of modulation power across auditory filters [Figs. 8(B)–8(E)], but the overall modulation power across auditory filters remained unchanged [Fig. 9(A)]. Furthermore, Fig. 9(A) shows that all four types of vocoded speech had the same average amount of modulation power as unprocessed speech

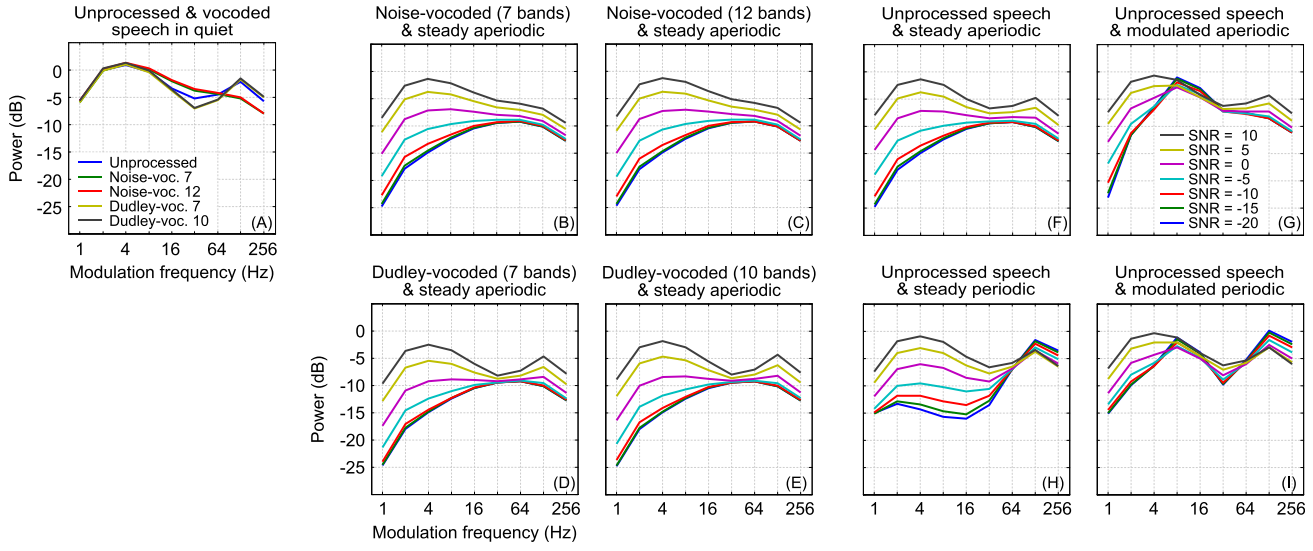

FIG. 9. (Color online) mr-sEPSM modulation spectra. The modulation spectra were generated by averaging modulation spectrograms of the kind shown in Fig. 8 over auditory filters, ignoring values smaller than −30 dB. For each of the eight combinations of target speech and masker (B)–(I), modulation spectra at the seven different SNRs are plotted. For comparison, (A) shows the modulation spectra of the five target speech conditions in quiet.

at modulation rates up to 8 Hz. In the presence of steady noise, however, there was a trend for a greater reduction of the modulation power at these low frequencies for Dudley-vocoded speech [Figs. 9(B)–9(F)]. The low-frequency modulations of unprocessed and noise-vocoded speech thus appear to be somewhat more robust in the presence of background noise.

In line with this observation, the predictions of the mr-sEPSM indeed did not change when noise-vocoded speech was used instead of unprocessed speech, irrespective of the number of channels [Figs. 5(A), 6(A), and 6(B)]. For Dudley-vocoded speech, in turn, speech intelligibility was correctly predicted to decrease, the more so with fewer channels [Figs. 6(C) and 6(D)]. In the latter case, the advantages of a modulation-based front end become apparent, as the ESII wrongly predicted that speech intelligibility with all four types of vocoded speech would be the same as with unprocessed speech.

D. sEPSM^{corr}

Of the four existing models, the sEPSM^{corr} was the only one for which the SRT predictions showed correct trends in all conditions. While both MPB [∼1.7 dB; Figs. 5(C) and 5(D)] and FMB [∼1.1 dB; Figs. 5(B) and 5(D)] were, on average, substantially underestimated, the model performed well when the target speech was vocoded (Fig. 6). For conditions including vocoded speech, SRTs were on average about 6.3 dB higher than for unprocessed speech and the SRT prediction errors in these conditions amounted to only about 1.4 dB on average [Fig. 7(A)].

As for the mr-sEPSM, the predictions were further analysed by visualising the internal signal representations of the model. In the case of the sEPSM^{corr}, this is complicated by the fact that the coefficients obtained by correlating the individual signal segments are time-integrated by taking the root of their sum of squares. Hence, the resulting χ-values [see Eq. (3) in Relaño-Iborra *et al.*, 2016] are exponentially larger at higher modulation frequencies, where the shorter window lengths of the individual segments lead to a much higher number of segments as compared to low modulation frequencies. Thus, it is more informative to study the differences of the correlations *across* conditions rather than their absolute values, so that the dominating pattern induced from the window-size difference is already considered. In the resulting "correlation difference spectrograms" (Fig. 10, upper row), the correlations in the reference condition were subtracted from those of each condition considered. Thus, positive values indicate higher correlations compared to the reference condition, and vice versa. As for the mr-sEPSM (cf. Fig. 9), the corresponding "correlation difference spectra" (Fig. 10, lower row) were computed by averaging the spectrograms over auditory filters, which allows the plotting together of results at different SNRs.

For the three comparisons including unprocessed target speech [Figs. 10(A)–10(C) and 10(F)–10(H)], it is apparent that higher correlations, relative to the reference condition, were offset by lower correlations elsewhere, particularly at lower SNRs. This is one reason why, despite the correct trends, the sEPSM^{corr} underestimated human performance in these conditions. Due to the strong 10-Hz modulations of the masker, the pattern observed for unprocessed speech and modulated noise [Figs. 10(A) and 10(F)] shows smaller correlations at lower modulation rates and higher correlations at higher modulation rates, caused by an unmasking of the speech modulations in the masker troughs. For the two periodic maskers [Figs. 10(B), 10(C), 10(G), and 10(H)], in contrast, the difference spectra show an unexpected (relative) decrease of the correlation coefficients at intermediate modulation frequencies and low SNRs, which counteracts the higher correlations at higher modulation rates. As speech carries no relevant information at these intermediate rates (cf. Sec. II), the question arises whether an inclusion of the respective modulation filters is useful at all.

For conditions including vocoded target speech, SRT prediction errors were markedly smaller than for the other

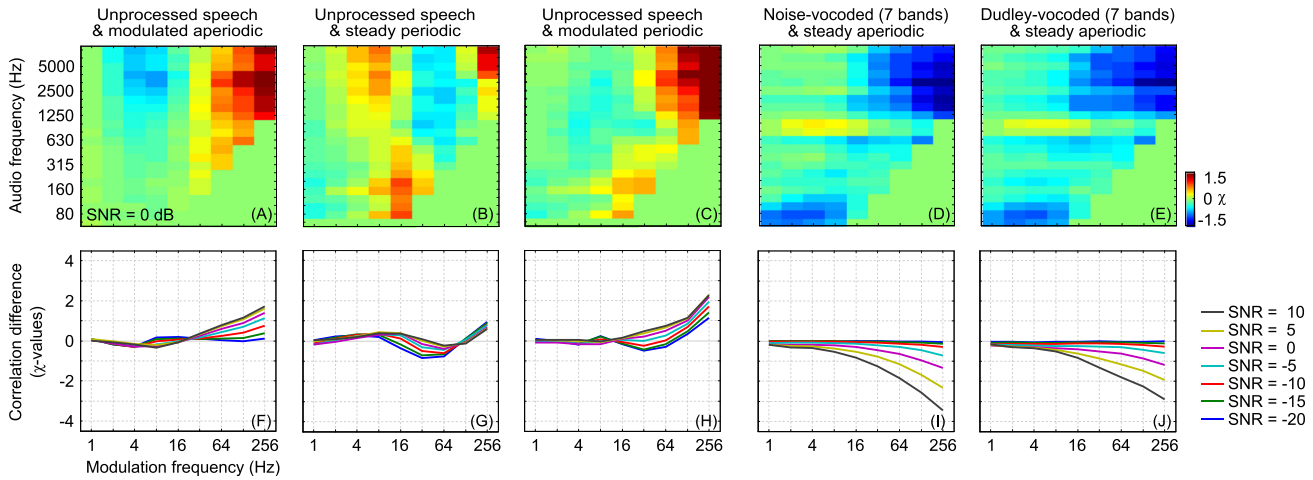

FIG. 10. (Color online) sEPSM^{corr} correlation difference spectrograms and spectra. Spectrograms (A)–(E): For each combination of auditory (y-axis) and modulation filters (x-axis), the time-integrated correlation-based decision metric (χ-value) in the reference condition (unprocessed speech and steady noise) was subtracted from that of the respective condition. In both conditions the SNR was 0 dB. Positive values correspond to a relative increase in the predicted speech intelligibility and vice versa. Each subplot is based on the entire set of materials in the respective condition. Spectra (F)–(J): correlation difference spectra of the same stimulus conditions at the seven different SNRs, obtained by averaging the spectrogram representations in the upper row over auditory filters.

models [cf. Fig. 7(A)]. Predictions were consistently more accurate than those of STOI and, particularly for noise-vocoded speech, also the mr-sEPSM, which demonstrates the benefit of combining a modulation-based front end and a correlation-based back end. However, a closer examination of Fig. 6 shows that the model predictions were very similar across the four conditions including vocoded speech, with the estimated SRTs differing by only about 1 dB. Changing the number of channels in the vocoder hence affected human performance to a larger degree than the model estimates. Furthermore, the sEPSM^{corr} could not reproduce the trend for slightly better behavioural performance with Dudley-vocoded speech. The latter finding can be explained with the correlation differences shown in Figs. 10(D)–10(J), which indeed are not consistently lower for Dudley-vocoded speech. While the differences were somewhat smaller for Dudley- than noise-vocoded speech at higher modulation rates, due to the preserved periodicity information, the opposite was true at lower modulation frequencies. As for the mr-sEPSM, the results thus suggest that the slower envelope modulations of noise-vocoded speech are more robust in the presence of background noise than those of Dudley-vocoded speech.

E. sEPSM^{corr}2

The sEPSM^{corr}2 improved the SRT predictions of the original sEPSM^{corr} in conditions including unprocessed speech by an average of about 2.7 dB [Figs. 5 and 7(A)], while the SRTs for conditions including vocoded speech hardly differed between the two model versions [∼0.05 dB on average; Figs. 6 and 7(A)].

As for the sEPSM^{corr}, correlation difference spectrograms and spectra were computed to facilitate a detailed examination of the model's predictions (Fig. 11). Conditions including vocoded speech have been omitted from Fig. 11, as the corresponding predictions did not substantially differ from the original model. Since the modulation filter selection algorithm, except for one single sentence (cf. Fig. 3), only excluded filters with centre frequencies above 8 Hz and full-wave rectification was by definition only applied to the outputs of modulation filters above 10 Hz, the correlation differences are the same for both models at low modulation frequencies.

First, for all three conditions shown in Fig. 11, the correlations relative to the reference condition were markedly higher at faster modulation rates, compared to the original model. Hence, full-wave rectifying the modulation filter outputs instead of extracting the Hilbert envelope served to emphasise the differences between steady noise and the other three maskers, as intended. The fact that the updated model also performed better with the modulated noise masker, indicates that the main reason for the improved predictions of the sEPSM^{corr}2 are the preserved random envelope modulations of the steady noise masker in the reference condition. Second, compared to the original sEPSM^{corr}, the negative correlation differences observed for the two periodic maskers at intermediate modulation frequencies and low SNRs were diminished. For the original model, this was one important reason for the poor predictions with these two maskers.

Even though the sEPSM^{corr}2 mostly discarded the intermediate modulation filters due to the modulation filter selection algorithm, the reduced negative correlation differences at these modulation frequencies indicate that the predictions might also improve when omitting the modulation filter selection stage. In fact, predictions of a version of the sEPSM^{corr}2 without the modulation filter selection algorithm were only slightly worse [RMSE of SRT prediction errors = 4; cf. Fig. 7(A)] than those of the full model. Last, a version of the sEPSM^{corr}2 that only included the modulation filter selection algorithm was tested (i.e., it still used the original second-order Hilbert transformation to extract the subband envelopes). Here, results did also improve compared to the sEPSM^{corr}, but by a much smaller margin (RMSE of SRT prediction errors = 4.9).

V. GENERAL DISCUSSION

A. Predicting the MPB

For steady maskers, the MPB of the human listeners amounted to about 10 dB in SRTs. The current section focusses on how well the models could account for this particular finding, as it is unaffected by the ability of the models

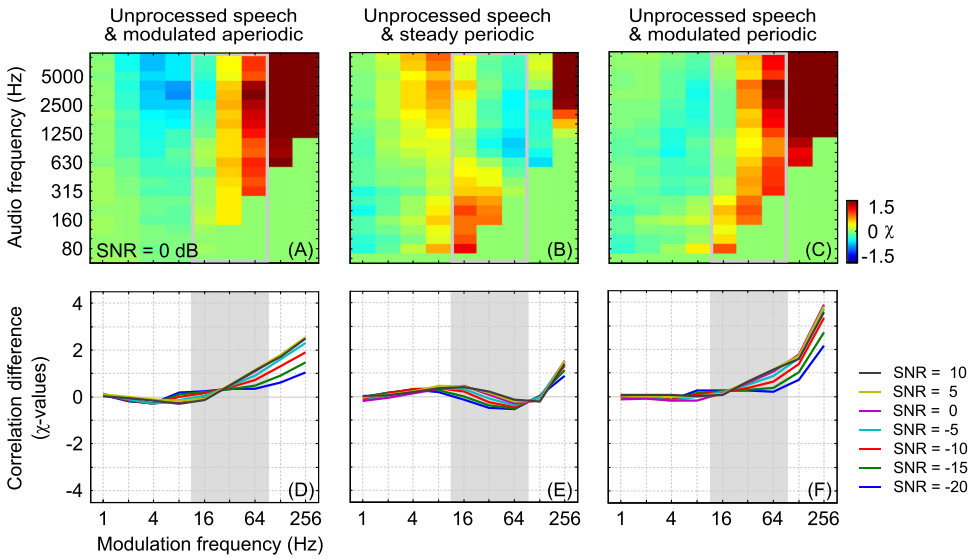

FIG. 11. (Color online) sEPSM^{corr}2 correlation difference spectrograms and spectra. As for the original sEPSM^{corr}, the time-integrated correlation-based decision metric (χ-value) in the reference condition (unprocessed speech and steady aperiodic masker) was subtracted from that in the respective condition to compute correlation difference spectrograms (A)–(C) and spectra ((D)–F). Modulation filters with intermediate frequencies, which were mostly discarded by the modified model, are marked in grey. All computational details, including the scaling, are as in Fig. 10.

to also predict the FMB. The two models that best predicted the MPB were mr-sEPSM and sEPSM^{corr}2, for which the SRT prediction errors were about 5 dB. Although far from accurate, these results are nevertheless informative as they suggest that about half the MPB is due to increased modulation masking caused by the aperiodic masker (Stone *et al.*, 2011; Stone *et al.*, 2012).

The other half of the MPB is hypothesised to be due to enhanced stream segregation associated with the pitch of periodic maskers (Deroche and Culling, 2011). Thus, the models of the sEPSM family could potentially be further improved by incorporating this mechanism, and the same also applies to the other models included in this study. As natural speech is mostly voiced, the benefits of these additions can also be expected to extend to other stimulus conditions such as interfering talkers.

Due to the use of two separate input signals, all models included in the current study have an unrealistically good ability to segregate target speech and masker, as they have access to either the unprocessed reference speech signal or the masker in isolation. Hence, they can reliably distinguish the two signals, irrespective of the processing condition. In a strict sense, for models including an unprocessed reference speech signal (i.e., STOI, sEPSM^{corr}, and sEPSM^{corr}2), this argument only applies if the target speech signal is unprocessed too, but this case is far more common than processed (e.g., vocoded) target signals. Thus, to account for the effect of streaming, the predictions of future models could be modified up or down depending on an additional processing step that quantifies the ease of stream segregation. In line with this idea, all models in the current study had PF slopes that were too steep for conditions including periodic maskers. Especially at very low SNRs, where pitch cues should be particularly helpful, the model predictions strongly underestimated human performance.

Rather than determining the actual pitch contours, this could be achieved by taking the strength of the *F*0-related envelope modulations in the mixture of speech and noise as a measure for the presence of pitch information. A related approach has recently been used by Josupeit and Hohmann (2017), who have modelled speech recognition in a multi-talker setting by finding the segments of the signal mixture with the highest amount of *F*0-related periodicity to identify the target talker. As can be seen in the modulation spectra shown in Figs. 9(F)–9(I), mixtures including periodic maskers generally had more energy in the highest two modulation filters. Moreover, the *F*0-related modulations were strongest at the most negative SNRs, where the masker pitch is assumed to be crucial for segregating speech and noise.

B. Predicting the FMB

For aperiodic maskers, the human listeners showed a FMB of about 5 dB in SRTs. Akin to the approach in Sec. V A, this part of the discussion focusses on this effect only, as it is unaffected by the ability of the models to also account for the MPB. The two models that could predict the FMB with reasonable accuracy (SRT prediction errors <2 dB) were ESII and sEPSM^{corr}2. These two models neither have the front nor back end in common (cf. Fig. 1). Consequently, from a theoretical point of view, different conceptual approaches can be used to successfully predict the FMB. However, the common feature of both models is that the length of the signal segments that are compared is frequency dependent and becomes very short (<10 ms) at the highest audio or modulation frequencies, respectively. The window length of STOI, in contrast, is fixed and comparably long (384 ms), resulting in the poorest FMB predictions of all tested models. As has been pointed out by, for example, Jørgensen *et al.* (2013), in the mr-sEPSM framework the successful prediction of the FMB can be attributed to the contribution of segments with window lengths shorter than the masker troughs, which enable the models to "listen in the dips." This pattern is also apparent in the correlation difference plots of the two sEPSM^{corr} models (Figs. 10 and 11), where the gains with fluctuating maskers are increasing at higher modulation frequencies. However, neither the mr-sEPSM nor the sEPSM^{corr} could fully account for the FMB in the present study as their ability to listen in the dips was counterbalanced by other effects (cf. Secs. IV C and IV D). As the mr-EPSM has been reported to account very well for the FMB when open-set sentence materials were used (Jørgensen *et al.*, 2013), one explanation for this discrepancy may be that the steady noise used in their study had different acoustical properties than the one used in the current study.

C. Predicting the intelligibility of vocoded speech

Overall, the SRT prediction errors of sEPSM^{corr} and sEPSM^{corr}2 were the smallest in conditions including vocoded target speech and there was virtually no difference between the two models (Fig. 6). It can thus be concluded that the combination of a modulation-based front end and a correlation back end is best suited to predict the lowered intelligibility of vocoded speech. However, the results of the two sEPSM^{corr} models were neither substantially affected by the number of channels in the vocoder nor by the periodicity of the target speech, in contrast to the human data, which showed moderate differences between these vocoding strategies. Instead, both models excelled at accounting for the generally lower intelligibility of vocoded speech. Moreover, the estimated PFs with vocoded speech were consistently steeper than the human PFs, a finding that also applies to STOI, pointing to a general limitation of correlation-based back ends. For STOI, in contrast, the reverse pattern was observed. While the intelligibility was overestimated throughout, the model predicted a slightly lower intelligibility with fewer vocoder channels, and the two conditions including Dudley-vocoded speech were correctly predicted to have a higher intelligibility then those with noise-vocoded speech.

As the front ends of all the models included in this study are technically able to represent the reduced spectral resolution of vocoded speech, the main reason for this seems to be that spectral information is not explicitly incorporated in the decision metrics of any of the tested models. One option to account for the diminished intelligibility of vocoded speech appears to be the inclusion of an across-frequency process, which quantifies the similarity across audio filters (Kates and

Arehart, 2014; Van Kuyk *et al.*, 2017). The HASPI model (*Hearing-Aid Speech Perception Index*), for example, was shown to perform well when the higher frequencies of speech were noise-vocoded (Kates and Arehart, 2014).

D. Modulation filter exclusion

The algorithm-based exclusion of modulation filters with intermediate centre frequencies (~16–64 Hz) was found to slightly improve the predictions of the sEPSM^{corr}2. Besides making the model slightly more parsimonious and computationally less intensive, this has several theoretical implications: First, these results suggest, as hypothesised, that amplitude modulations at intermediate rates do not carry information that is relevant for speech intelligibility. Second, under the assumption that a modulation filterbank of the kind implemented in mr-sEPSM and sEPSM^{corr} exists, this raises the question how the human auditory system integrates information across modulation filters. This is a crucial issue, as the contributions of the individual auditory and modulation filters are not determined by *a priori* weights in the sEPSM models, in contrast to other models such as ESII and the speech-based speech transmission index (Payton and Braida, 1999). One possibility is that the contribution of irrelevant filters is minimised. Alternatively, the concept of a modulation filterbank with a fixed set of filter centre frequencies may be challenged and substituted with a more flexible approach, where the filter tuning is dependent on the input signal, akin to the proposed exclusion algorithm.

E. Comparison with ASR-based modelling approaches

All models included in this study are so-called intrusive models that require either the unprocessed speech or the noise as a reference signal. Although it may be argued that these template signals mimic the speech-specific knowledge of the listeners, practical problems arise when the reference is unavailable. In contrast, models with automatic speech recognition (ASR) back ends usually do not require a reference signal. However, for example, in the case of the *Framework for Auditory Discrimination Experiments* (FADE; Schädler *et al.*, 2016; Schädler *et al.*, 2015), the same noisy speech materials are used in the training and test phases. In addition to making this model not strictly reference free, this procedure could potentially make the back end rely on acoustic features that do not generalise across different materials or are even irrelevant for human auditory perception. Although based on a very different conceptual approach, in which time-frequency units above a certain SNR are used as input for a missing-data ASR back end, the glimpsing model (Cooke, 2006) similarly relies on an implicit reference signal. However, the set of materials used for training is considerably larger than the test set in this case, which should preclude overtraining effects. A reference-free model that can also predict the intelligibility of speech materials recorded from unknown talkers has recently been proposed by Spille *et al.* (2018). Consisting of a front end including a deep neural network (DNN) in conjunction with an ASR-based back end, their model outperformed ESII, STOI, and mr-sEPSM in a range of conditions. Although the complexity of the DNN makes it difficult to know which acoustic cues were used, the model was shown to exploit the dips of a modulated masker. Same as the FADE (Schädler *et al.*, 2016), it could thus successfully account for the FMB.

However, since the prediction errors for the FMB were also small for ESII and sEPSM^{corr}2, the more relevant question is whether FADE and the model of Spille *et al.* (2018) could potentially account for the MPB too. As mentioned in Sec. V A, for models that require a reference signal, the segregation of speech and masker is unrealistically good, which may be one reason for the underestimated MPB. As this limitation applies to both sEPSM^{corr}2 and FADE, it appears unlikely that their prediction errors for the MPB would differ substantially. The model of Spille *et al.* (2018), on the other hand, would have to perform the stream segregation itself. Potentially, this could result in more accurate predictions of the MPB or, on the contrary, the model could confuse the periodic maskers with the target speech. However, since this model only includes envelope modulation filters with centre frequencies up to 27 Hz, the *F*0-related modulations are missing as potential cue, which makes a successful prediction of the MPB less likely.

F. sEPSM^{corr}2 backward compatibility

Finally, it was tested whether the introduced modifications are backward compatible with the original model, i.e., if the predictions of the sEPSM^{corr} reported in Relaño-Iborra *et al.* (2016) can be re-produced with the sEPSM^{corr}2. The predictive power of the sEPSM^{corr} was tested for a broad range of data sets, including speech mixed with fluctuating interferers, reverberant noisy speech, speech distorted with phase jitter and two noise-reduction algorithms, spectral subtraction, and ideal time frequency segregation (ITFS). As described in more detail in the Appendix, the sEPSM^{corr}2 was found to perform as well as or even better than the sEPSM^{corr}. Predictions improved markedly for speech mixed with additive noise and ITFS.

VI. SUMMARY AND CONCLUSIONS

To gain a better understanding of the behavioural data and evaluate different approaches to modelling the intelligibility of speech, the results reported in Steinmetzger and Rosen (2015) were predicted using four existing models (ESII, STOI, mr-sEPSM, and sEPSM^{corr}) as well as a modified version of one of them (sEPSM^{corr}2). The original data were obtained from normal-hearing listeners presented with various combinations of speech and background noise. The main finding was that subjects performed substantially better when the masker was periodic, while they only benefitted slightly from periodicity in the vocoded target speech. The listeners also showed a FMB when the maskers were amplitude modulated at a rate of 10 Hz, but this effect was markedly smaller than the MPB and was only observed for target speech that was very intelligible in quiet listening conditions.

In summary, the four previously published models consistently underestimated MPB, as well as FMB, albeit to varying degrees. For vocoded target speech, the opposite pattern

was observed, and all models overestimated the intelligibility of the materials, apart from the sEPSM^{corr}. To understand these shortcomings, the internal signal representation of each model was analysed in detail. Overall, the sEPSM^{corr}, characterised by a combination of modulation-based front end and correlation back end, produced the best predictions, whereas there was little difference among the other three models.

As the sEPSM^{corr} also underestimated MPB and FMB, a modified version of this model was developed to further improve the predictions. For the resulting sEPSM^{corr}2, the simulation of the diminished phase sensitivity of the auditory system at higher modulation frequencies was altered to better preserve the fine signal details, which led to reduced predictions errors, by 2–3 dB in SRT for both FMB and MPB. Additionally, an algorithm that excluded modulation filters in which speech has little energy, resulted in a small further improvement of the predictions. Discarded modulation filters were almost exclusively tuned to intermediate modulation rates (~16–64 Hz), suggesting that these do not contribute to speech intelligibility in the conditions considered here. In summary, the SRTs predicted by the sEPSM^{corr}2 showed that the model could account for the FMB, as well as the reduced intelligibility of vocoded speech, but still failed to explain about half of the MPB. While this finding helps to quantify the contribution of modulation masking to the MPB, it also shows that pitch-related effects, such as enhanced stream segregation, should be incorporated into future models.

ACKNOWLEDGMENTS

We thank Gaston Hilkhuysen and Koenraad Rhebergen for helpful comments and assistance with the ESII code. This project has been funded with support from the European Commission under Contract No. FP7-PEOPLE-2011-290000, the Dietmar Hopp Stiftung (Grant No. 2301 1239), and the Oticon Centre of Excellence for Hearing and Speech Sciences (CHeSS). A MATLAB implementation of the sEPSM^{corr}2 is available at https://bitbucket.org/heliaib/sepsm-corr/downloads.

TABLE II. Accuracy metrics for the sEPSM^{corr} and sEPSM^{corr}2 models.

	sEPSM^{corr}		sEPSM^{corr}2	
	ρ	Mean average error (MAE)	ρ	MAE
Additive noise	0.97	1.85 dB	0.99	0.66 dB
Reverberation	—	—	—	—
Spectral subtraction	0.82	0.59 dB	0.84	0.51 dB
Phase jitter	0.97	19.0%	0.97	19.0%
ITFS	0.79	12.1%	0.93	7.3%

APPENDIX: sEPSMcorr2 BACKWARD COMPATIBILITY

To assess the validity of the modifications made to the sEPSM^{corr}, an analysis under the same conditions as in Relaño-Iborra *et al.* (2016) was carried out. Table II summarises the accuracy of the sEPSM^{corr} and sEPSM^{corr}2 by means of the Pearson's correlation and the mean average error (MAE) between the model predictions and the data.

The sEPSM^{corr}2 led to substantially improved predictions in the conditions of speech with additive noise (with increased ρ and reduced MAE values) and for ITFS, where the Pearson's correlation increased (from $\rho = 0.79$ to $\rho = 0.93$) and the MAE decreased (from 12.1% to 7.3%). The results of the additive noise condition were to be expected, since it is a similar condition to that investigated in the present study (unprocessed noise with stationary and fluctuating interferers), and the improvements discussed in Sec. IV E still hold.

The results in conditions with ITFS (Fig. 12), on the other hand, were obtained for conditions not previously tested. The modelled dataset was obtained from Kjems *et al.* (2009). In their study, the effects of the configuration of the ideal binary mask (IBM; Brungart *et al.*, 2006) on speech intelligibility were investigated using the Dantale II sentence corpus. Four different interferers were considered: speech-shaped noise, car-cabin noise, noise produced by bottles on a conveyor belt, and two people speaking in a cafeteria. Furthermore, two different SNRs were used to generate the noisy mixture, corresponding to the 20%- and 50%-correct

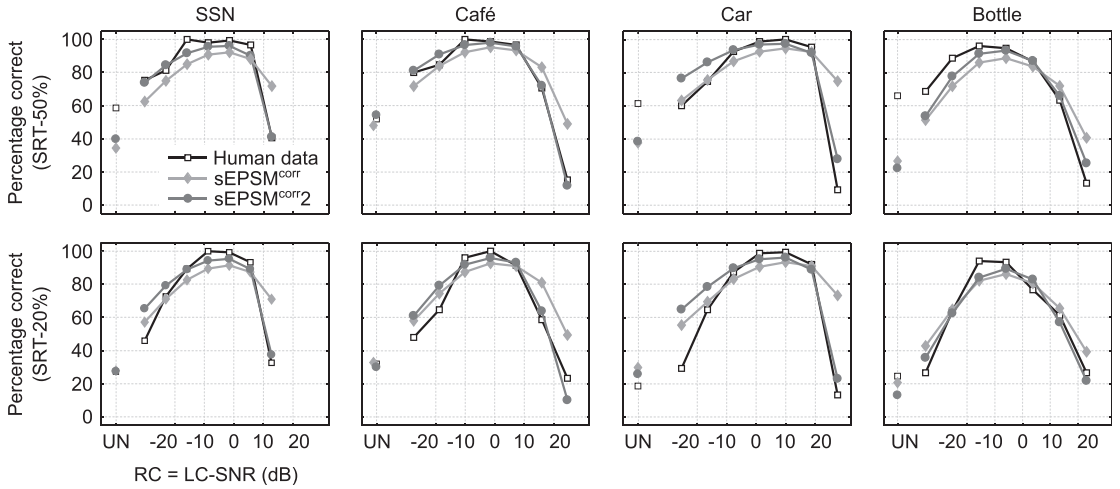

FIG. 12. sEPSM^{corr}2 backward compatibility. Intelligibility scores for ideal time frequency segregation (ITFS) processed speech for four different interferers (columns) and two SNRs (rows). The human data are taken from Kjems *et al.* (2009) and the figure is an adaptation of Fig. 6 in Relaño-Iborra *et al.* (2016).

points on the respective PF. Finally, eight different relative criteria (RC) for the IBM were considered for each noisy mixture.

The original sEPSM^{corr} was shown to accurately predict the effects of the different interferers, SNRs, and RCs. However, the sEPSM^{corr} to some extent overestimated human performance in conditions with strict (i.e., high) RC, i.e., IBMs with a low density (less than 1% of non-zero elements in the mask). In these conditions, the sEPSM^{corr}2 clearly outperformed the sEPSM^{corr}, yielding better correlations with the data.

To further investigate the source of the model improvements in this condition, the relative contributions of the modulation channel exclusion algorithm and the envelope extraction scheme were tested in isolation. It was found that both model modifications improved the performance of the original sEPSM^{corr} substantially when applied independently. However, the addition of the channel selection algorithm to the full-wave rectification did not yield any additional improvements of the model's performance. When considering the channel selection only ($\rho = 0.87$, MAE = 10.3%), the improvements resulting from the removal of the intermediate modulation bands from further processing are likely due to spurious correlations dominating those channels. Thus, when removed, the overall predicted scores are lower, in line with the human data. On the other hand, the full-wave rectification alone provides similar accuracy to that obtained in combination with the channel selection algorithm ($\rho = 0.92$, MAE = 7.21%). This means that the increase in resolution of the envelope signals, provided by full-wave rectification might suffice to predict the breakdown of intelligibility for those sparse masks with high RC. Despite maintaining the intermediate modulation bands, the increased acuity is likely to reduce the spurious correlations in them, thus reducing the averaged correlation values.

Like the sEPSM^{corr}, the sEPSM^{corr}2 cannot account for effects of reverberation on the intelligibility of noisy speech. Relaño-Iborra *et al*. (2016) showed that a long-term version of the model (without multi-resolution processing) could account for these effects, while compromising the model's predictive power in other conditions. This was not tested with the sEPSM^{corr}2, but it is hypothesised that a similar performance improvement for this condition could be obtained by applying long-term analysis.

[1]*F*0-vocoded speech was omitted in the current study, since its particularly low intelligibility requires a linguistic explanation (see Sec. II C in Steinmetzger and Rosen, 2015), which is beyond the scope of the computational speech intelligibility models considered here.

[2]All signals in Fig. 2(B) are shown after the subsequent logarithmic compression, just before they are segmented into time frames and correlated with each other. As the previously non-negative envelope signals were passed through a modulation filter, they can contain both positive and negative values.

ANSI S3.5 (**1997**). *Methods for the Calculation of the Speech Intelligibility Index* (American National Standards Institute, New York).

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.

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

Brungart, D. S., Chang, P. S., Simpson, B. D., and Wang, D. (**2006**). "Isolating the energetic component of speech-on-speech masking with ideal time-frequency segregation," J. Acoust. Soc. Am. **120**, 4007–4018.

Byrne, D., Dillon, H., Tran, K., Arlinger, S., Wilbraham, K., Cox, R., Hagerman, B., Hetu, R., Kei, J., and Lui, C. (**1994**). "An international comparison of long-term average speech spectra," J. Acoust. Soc. Am. **96**, 2108–2120.

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.

Cooke, M. "A glimpsing model of speech perception in noise," J. Acoust. Soc. Am. **119**, 1562–1573 (**2006**).

Dau, T., Kollmeier, B., and Kohlrausch, A. (**1997a**). "Modeling auditory processing of amplitude modulation. I. Detection and masking with narrow-band carriers," J. Acoust. Soc. Am. **102**, 2892–2905.

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

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.

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

Fant, G., Liljencrants, J., and Lin, Q.-G. (**1985**). "A four-parameter model of glottal flow," Speech Trans. Lab.: Q. Progress Status Rep. **4**, 1–13.

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.

French, N. R., and Steinberg, J. C. (**1947**). "Factors governing the intelligibility of speech sounds," J. Acoust. Soc. Am. **19**, 90–119.

Giraud, A.-L., Lorenzi, C., Ashburner, J., Wable, J., Johnsrude, I., Frackowiak, R., and Kleinschmidt, A. (**2000**). "Representation of the temporal envelope of sounds in the human brain," J. Neurophysiol. **84**, 1588–1598.

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

Jensen, J., and Taal, C. H. (**2016**). "An algorithm for predicting the intelligibility of speech masked by modulated noise maskers," IEEE/ACM Trans. Audio, Speech Lang. Process. **24**, 2009–2022.

Jørgensen, S., and Dau, T. (**2011**). "Predicting speech intelligibility based on the signal-to-noise envelope power ratio after modulation-frequency selective processing," J. Acoust. Soc. Am. **130**, 1475–1487.

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.

Joris, P., Schreiner, C., and Rees, A. (**2004**). "Neural processing of amplitude-modulated sounds," Physiol. Rev. **84**, 541–577.

Josupeit, A., and Hohmann, V. (**2017**). "Modeling speech localization, talker identification, and word recognition in a multi-talker setting," J. Acoust. Soc. Am. **142**, 35–54.

Kates, J. M., and Arehart, K. H. (**2014**). "The hearing-aid speech perception index (HASPI)," Speech Commun. **65**, 75–93.

Kjems, U., Boldt, J. B., Pedersen, M. S., Lunner, T., and Wang, D. (**2009**). "Role of mask pattern in intelligibility of ideal binary-masked noisy speech," J. Acoust. Soc. Am. **126**, 1415–1426.

Kryter, K. D. (**1962**). "Methods for the calculation and use of the articulation index," J. Acoust. Soc. Am. **34**, 1689–1697.

Payton, K. L., and Braida, L. D. (**1999**). "A method to determine the speech transmission index from speech waveforms," J. Acoust. Soc. Am. **106**, 3637–3648.

Plomp, R., and Mimpen, A. (**1979**). "Improving the reliability of testing the speech reception threshold for sentences," Int. J. Audiol. **18**, 43–52.

Relaño-Iborra, H., May, T., Zaar, J., Scheidiger, C., and Dau, T. (**2016**). "Predicting speech intelligibility based on a correlation metric in the envelope power spectrum domain," J. Acoust. Soc. Am. **140**, 2670–2679.

Rhebergen, K. S., and Versfeld, N. J. (**2005**). "A speech intelligibility index-based approach to predict the speech reception threshold for sentences in fluctuating noise for normal-hearing listeners," J. Acoust. Soc. Am. **117**, 2181–2192.

Rhebergen, K. S., Versfeld, N. J., and Dreschler, W. A. (**2006**). "Extended speech intelligibility index for the prediction of the speech reception threshold in fluctuating noise," J. Acoust. Soc. Am. **120**, 3988–3997.
Rosen, S. (**1992**). "Temporal information in speech: Acoustic, auditory and linguistic aspects," Philos. Trans. R. Soc. Lond. B Biol. Sci. **336**, 367–373.
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.
Schädler, M. R., Warzybok, A., Ewert, S. D., and Kollmeier, B. (**2016**). "A simulation framework for auditory discrimination experiments: Revealing the importance of across-frequency processing in speech perception," J. Acoust. Soc. Am. **139**, 2708–2722.
Schädler, M. R., Warzybok, A., Hochmuth, S., and Kollmeier, B. (**2015**). "Matrix sentence intelligibility prediction using an automatic speech recognition system," Int. J. Audiol. **54**, 100–107.
Schubotz, W., Brand, T., Kollmeier, B., and Ewert, S. D. (**2016**). "Monaural speech intelligibility and detection in maskers with varying amounts of spectro-temporal speech features," J. Acoust. Soc. Am. **140**, 524–540.
Spille, C., Ewert, S. D., Kollmeier, B., and Meyer, B. T. (**2018**). "Predicting speech intelligibility with deep neural networks," Comput. Speech Lang. **48**, 51–66.
Steeneken, H. J. M., and Houtgast, T. (**1980**). "A physical method for measuring speech-transmission quality," J. Acoust. Soc. Am. **67**, 318–326.
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. (**2017**). "Effects of acoustic periodicity and intelligibility on the neural oscillations in response to speech," Neuropsychologia **95**, 173–181.
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.
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.
Taal, C. H., Hendriks, R. C., Heusdens, R., and Jensen, J. (**2011**). "An algorithm for intelligibility prediction of time-frequency weighted noisy speech," IEEE Trans. Audio, Speech, Lang. Process. **9**, 2125–2136.
Van Kuyk, S., Kleijn, W. B., and Hendriks, R. C. (**2017**). "An instrumental intelligibility metric based on information theory," IEEE Signal Process. Lett. **25**, 115–119.
Van Kuyk, S., Kleijn, W. B., and Hendriks, R. C. (**2018**). "An evaluation of intrusive instrumental intelligibility metrics," IEEE/ACM Trans. Audio, Speech Lang. Process. **26**, 2153–2166.
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 in Tools and Resources for the Analysis of Speech Prosody*, pp. 7–10.