Predicting speech intelligibility based on a correlation metric in the envelope power spectrum domain

Iborra, Helia Relano; May, Tobias; Zaar, Johannes; Scheidiger, Christoph; Dau, Torsten

Published in:
Journal of the Acoustical Society of America

Link to article, DOI:
10.1121/1.4964505

Publication date:
2016

Document Version
Publisher's PDF, also known as Version of record

Predicting speech intelligibility based on a correlation metric in the envelope power spectrum domain

Helia Relaño-Iborra,[a] Tobias May, Johannes Zaar, Christoph Scheidiger, and Torsten Dau
Hearing Systems Group, Department of Electrical Engineering, Technical University of Denmark, DK-2800 Kgs. Lyngby, Denmark

(Received 28 April 2016; revised 8 August 2016; accepted 22 September 2016; published online 17 October 2016)

A speech intelligibility prediction model is proposed that combines the auditory processing front end of the multi-resolution speech-based envelope power spectrum model [mr-sEPSM; Jørgensen, Ewert, and Dau (2013). J. Acoust. Soc. Am. **134**(1), 436–446] with a correlation back end inspired by the short-time objective intelligibility measure [STOI; Taal, Hendriks, Heusdens, and Jensen (2011). IEEE Trans. Audio Speech Lang. Process. **19**(7), 2125–2136]. This "hybrid" model, named sEPSM^{corr}, is shown to account for the effects of stationary and fluctuating additive interferers as well as for the effects of non-linear distortions, such as spectral subtraction, phase jitter, and ideal time frequency segregation (ITFS). The model shows a broader predictive range than both the original mr-sEPSM (which fails in the phase-jitter and ITFS conditions) and STOI (which fails to predict the influence of fluctuating interferers), albeit with lower accuracy than the source models in some individual conditions. Similar to other models that employ a short-term correlation-based back end, including STOI, the proposed model fails to account for the effects of room reverberation on speech intelligibility. Overall, the model might be valuable for evaluating the effects of a large range of interferers and distortions on speech intelligibility, including consequences of hearing impairment and hearing-instrument signal processing. [http://dx.doi.org/10.1121/1.4964505]

[JFL] Pages: 2670–2679

I. INTRODUCTION

Speech is the main tool used by humans to communicate with one another, making it a key factor in most social interactions. The way in which humans process and decode speech signals has been a focus of research for decades and various speech perception models have been presented that attempt to quantify the effects of the acoustic properties of the target speech and the interferers, the effects of the environment (e.g., a room) or transmission channel (e.g., a communication device or a hearing instrument), as well as effects of auditory processing (e.g., a hearing loss) on speech intelligibility. Such models have been useful for the development and evaluation of new telecommunication systems, hearing-aid algorithms, and speech synthesis systems.

The research on objective speech intelligibility measures started in the first half of the 20th century. The first intelligibility model was developed by Harvey Fletcher in the 1920s (see Allen, 1996), although it was first made public by French and Steinberg (1947). The model could account for intelligibility scores in quiet and in the presence of additive noise. The concepts underlying this model, called the articulation index (AI), were thoroughly described by Kryter (1962) and later standardized by ANSI (1969). The AI is based on the assumption that background noise affects speech intelligibility differently in different frequency bands. The AI was later extended and modified into the speech intelligibility index (SII; ANSI, 1997), which includes corrections for hearing sensitivity loss, speech level, and upward and downward spread of masking.

The predictions of the AI and SII are based on a weighted average of the long-term signal-to-noise-ratio (SNR) in different frequency bands, using the clean speech signal and the background noise as inputs. This long-term analysis implies that the models are insensitive to short-term effects, e.g., the ability of human listeners to utilize speech information in the dips of temporally fluctuating maskers, such as interfering speech, often referred to as "listening-in the-dips" (Festen and Plomp, 1990). Such a dip-listening strategy can lead to a reduced amount of masking, or interference, as compared to a steady-state condition (Festen and Plomp, 1990). As a modification of the standard SII, the extended speech intelligibility index (ESII; Rhebergen *et al.*, 2006), a short-term analysis was introduced to improve the model's performance in fluctuating noise. However, since the ESII assumes that the clean speech and the noise can be accessed separately, it cannot account for conditions where the speech and noise mixture have been subjected to non-linear processing, such as noise reduction algorithms or amplitude compression schemes (Rhebergen *et al.*, 2009).

Another approach to speech intelligibility modeling has been the analysis of the stimulus characteristics in the modulation domain. Houtgast *et al.* (1980) proposed the speech transmission index (STI), based on the concept of the modulation transfer function, which is obtained by

[a]Electronic mail: heliaib@elektro.dtu.dk

measuring the change in the modulation depth of a probe signal, a modulated noise, as a function of modulation frequency. The STI was demonstrated to be successful in conditions with reverberant speech and in conditions with speech presented in additive noise. However, as shown by Ludvigsen *et al.* (1993), the STI cannot account for effects of non-linear processing, such as spectral subtraction, on speech intelligibility and is not sensitive to the effects of masking release in conditions with fluctuating interferers. Several subsequent models were developed that are based on the concept of the STI. The speech-based STI (Payton and Braida, 1999) considers speech signals as an input to the model, instead of the fixed probe signal used in the original STI, and thus generalizes the model to various types of speech materials. Another modification of the STI, the coherence-based STI (Kates and Arehart, 2005) was shown to account for non-linear processing, such as peak-clipping. An extensive review of the STI-based approaches and other speech intelligibility models (Holube and Kollmeier, 1996; Drullman *et al.*, 1994; Ludvigsen *et al.*, 1990) was provided by Goldsworthy and Greenberg (2004), investigating their ability to account for different types of non-linear distortions. Their results showed that none of the tested models performed accurately in all experimental conditions considered in their study.

More recently, two models have been presented that account for speech intelligibility data in conditions where the STI- and SII-based approaches fail: The short-time objective intelligibility (STOI) measure (Taal *et al.*, 2011) and the speech-based envelope power spectrum model (sEPSM; Jørgensen and Dau, 2011). The STOI is based on the idea that the similarity between the clean speech and the processed (noisy) speech is related to speech intelligibility. The outputs of a front end processing based on a discrete Fourier transform decomposition are analyzed by means of a back end that performs a cross correlation between the clean speech and the processed speech. STOI accounts for effects of ideal time frequency segregation (ITFS), a noise reduction scheme that applies a binary mask onto the time-frequency (T-F) representation of the noisy speech (Wang, 2005; Brungart *et al.*, 2006), as well as for the effects of other noise reduction algorithms. However, as discussed in Taal *et al.* (2011), STOI may not be suitable for predicting the intelligibility of reverberant speech. Furthermore, the model can be expected to fail in conditions with fluctuating interferers since it applies relatively long integration time windows (of about 380 ms duration), whereas studies have suggested the need for shorter time constants to account for such conditions (e.g., Rhebergen *et al.*, 2006; Jørgensen *et al.*, 2013).

The sEPSM operates in the envelope-frequency domain and assumes that the SNR of the noisy speech in the envelope domain (SNR_{env}), after the processing through a peripheral bandpass filterbank and a subsequent modulation filterbank at the output of each peripheral filter, is related to speech intelligibility. The predictions are based on the analysis of the noisy speech and the noise alone in terms of their intrinsic envelope fluctuations, an analysis that was originally considered in the framework of the envelope power spectrum model (Dau *et al.*, 1999; Ewert and Dau, 2000) to account for (non-speech) modulation detection and masking data. The sEPSM was shown to account for effects of reverberation, additive noise, and spectral subtraction, a non-linear noise-reduction algorithm (Jørgensen and Dau, 2011). Furthermore, a "multi-resolution" version of the model [multi-resolution speech-based envelope power spectrum model (mr-sEPSM), Jørgensen *et al.*, 2013] was shown to account for the effects of masking release in fluctuating noise. However, Chabot-Leclerc *et al.* (2014) showed that the sEPSM fails in conditions of phase jitter distortion. Furthermore, since the model operates on the (processed) noisy speech and the (processed) noise alone, it might not be sensitive to the effects of ITFS processing, which is only applied to the noisy speech but not to the noise alone.

Thus, the two speech perception modeling approaches (STOI and sEPSM) appear to exhibit complementary strengths and limitations. The hypothesis of the present study was that a combination of the building blocks in the front end preprocessing of one of the models, the sEPSM, and the back end processing of the other model, STOI, may account for the data from a broader range of conditions. A "hybrid" model was developed here, referred to as $sEPSM^{corr}$, which combines the preprocessing of the mr-sEPSM with a cross-correlation back end similar to the one used in STOI. The results obtained with the proposed model were compared to the original models in the conditions of fluctuating-noise interferers, reverberation, and non-linear distortions (spectral subtraction, phase jitter, and ITFS).

II. MODEL DESCRIPTION

The overall structure of the proposed model, the $sEPSM^{corr}$, is shown in Fig. 1. The model consists of an auditory preprocessing front end and a decision back end. The clean speech and the degraded, or processed, speech signals are sampled at a rate of 22 kHz and processed by the auditory front end. The resulting signal representations are then compared in the decision back end.

A. Auditory preprocessing stages

The first stage of the auditory preprocessing simulates the frequency-selective processing on the basilar membrane and is represented by an auditory filterbank consisting of 22 fourth-order gammatone filters with center frequencies ranging from 63 Hz to 8 kHz with 1/3 octave spacing (Patterson *et al.*, 1987). The filterbank output is processed further only if the stimulus level in a given band is above the hearing threshold in quiet (ISO, 2005). The envelope is extracted in each frequency channel by calculating the analytic signal using the Hilbert transform, and taking its absolute value. The envelope in each channel is then filtered by a first-order low-pass filter with a cutoff frequency of $f_c = 150$ Hz, reflecting the sluggishness of the auditory system to follow fast envelope fluctuations (Ewert and Dau, 2000; Kohlrausch *et al.*, 2000). This is followed by a modulation filterbank consisting of a third-order low-pass filter with a cutoff frequency $f_c = 1$ Hz in parallel with eight second-order bandpass filters with octave spacing, a constant quality factor Q of 1, and center frequencies ranging from 2 to 256 Hz, as in

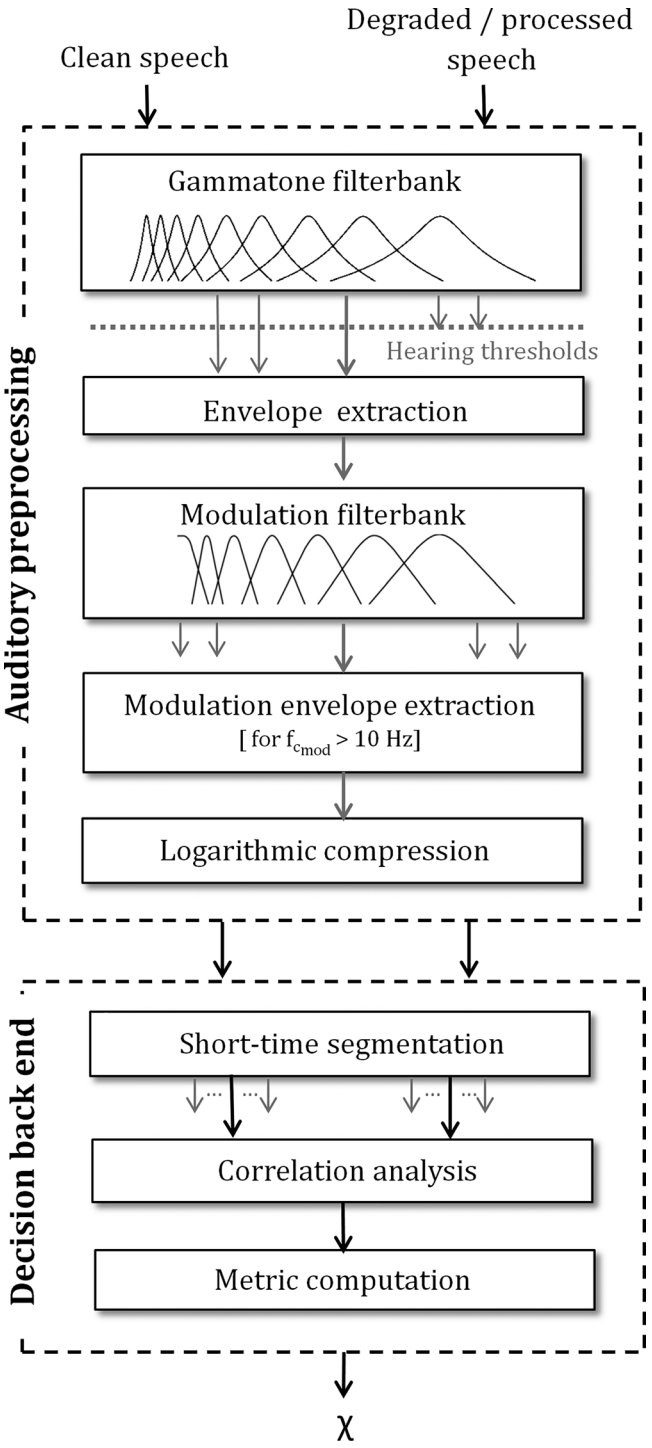

FIG. 1. Structure of the proposed model. The clean speech and the degraded or processed noisy mixture are processed through the auditory front end, including a gammatone filterbank, envelope extraction, a modulation filterbank, and a logarithmic amplitude compression. The outputs of the two signals are then analyzed in short time windows by means of their cross-correlation in the model's back end.

the mr-sEPSM (Jørgensen *et al.*, 2013). To model the modulation-phase sensitivity along the auditory pathway and its limitations (Langner and Schreiner, 1988), the time signals at the outputs of the modulation filters centered at frequencies below 10 Hz remain unchanged (exhibiting positive and negative amplitudes), whereas another (second-order) Hilbert envelope is calculated from the time signals at the outputs of the modulation filters centered at frequencies above 10 Hz. The modulation-phase sensitivity at low modulation frequencies in the proposed model was not included in the original sEPSM, but is inspired by the assumptions made in the auditory signal processing model of Dau *et al.* (1997a,b), which combines such a processing stage with a correlation-based (template-matching) back end. At the output, the stimulus representations are logarithmically compressed in amplitude to satisfy Weber's law in the modulation domain, motivated by data on modulation depth discrimination (Ewert and Dau, 2004).

B. Decision back end processing

Each modulation filtered output is processed in different time segments depending on its center frequency, as in the mr-sEPSM. Rectangular windows with no overlap and duration proportional to the inverse of the respective modulation-filter center frequency are applied, i.e., the segment durations range from 1 s for the 1-Hz modulation filter to 3.9 ms for the 256-Hz modulation filter. Thus, the number of considered segments is directly proportional to the modulation frequency, i.e., the higher the modulation filter's center frequency, the more segments are considered. Only the outputs of the modulation filters with a center frequency below one-fourth of the corresponding auditory filter's center frequency are included in the computation (Verhey *et al.*, 1999; Jørgensen *et al.*, 2013).

The outputs of each auditory filter and each modulation filter for the two inputs are cross-correlated with zero lag on a segment-by-segment basis. With $\boldsymbol{x}$ and $\boldsymbol{y}$ being the clean speech signal and the noisy speech signal vectors, respectively, and similarly to Eq. (5) in Taal *et al.* (2011), the correlation coefficient is defined as

$$\chi'_{k,j,i} = \frac{(\boldsymbol{x}_{k,j,i} - \bar{\boldsymbol{x}}_{k,j,i}) \cdot (\boldsymbol{y}_{k,j,i} - \bar{\boldsymbol{y}}_{k,j,i})}{\|\boldsymbol{x}_{k,j,i} - \bar{\boldsymbol{x}}_{k,j,i}\| \cdot \|\boldsymbol{y}_{k,j,i} - \bar{\boldsymbol{y}}_{k,j,i}\|}, \tag{1}$$

where the correlation, χ', between the clean speech signal and the noisy speech signal is calculated for each time segment (k), modulation filter (j), and auditory filter (i). The correlation coefficient in Eq. (1) ranges from -1 to 1. In the framework of the model, segments with negative correlations are assumed not to contribute to intelligibility. Thus, the following correction is applied:

$$\chi_{k,j,i} = \begin{cases} \chi'_{k,j,i} & \text{if } \chi'_{k,j,i} > 0 \\ 0 & \text{otherwise.} \end{cases} \tag{2}$$

Afterwards, the correlation values are integrated across time (i.e., across segments), using a "multiple looks" approach (Viemeister and Wakefield, 1991)

$$\chi_{j,i} = \sqrt{\sum_{k=1}^{K(j)} (\chi_{k,j,i})^2}, \tag{3}$$

with $K(j)$ indicating the number of segments obtained from the output of modulation filter j. Then, the values are

averaged across all modulation and gammatone filters resulting in the final correlation metric

$$\chi = \frac{1}{I \cdot J - J_{\text{exc}}} \cdot \sum_{i,j} \chi_{j,i}, \tag{4}$$

where I represents the total number of gammatone filters (excluding those where the stimulus energy is below the hearing threshold), J denotes the total number of modulation filters, and J_{exc} is the total number of modulation filters centered at frequencies above one-fourth of each gammatone filter center frequency and thus excluded from the computation.

The correlation-based output of the proposed model, χ, increases monotonically with SNR. To create a mapping between χ and intelligibility scores, a logistic function is applied to the model outcome

$$\Phi(\chi) = \frac{100}{1 + e^{(a \cdot \chi + b)}}, \tag{5}$$

where a and b represent the free parameters of the curve. To obtain the optimal values of a and b, a fitting condition has to be defined. In this study, the model was "calibrated" separately to two speech corpora, whereby all model parameters were then kept fixed for a given material throughout the different experimental conditions (see Sec. III C).

III. METHODS

A. Speech materials

Two speech corpora were used. The first one was the "conversational language understanding evaluation" (CLUE; Nielsen and Dau, 2009). The CLUE consists of Danish five-word sentences spoken by a male native Danish speaker. The sentences were constructed from an open word set, are grammatically correct, and represent daily-life communication. The other material was taken from the DANTALE II corpus (Wagener *et al.*, 2003), a Danish matrix sentence test recorded by a female native Danish speaker. DANTALE II consists of five words taken from a base of ten sentences (i.e., a closed set) that have the same structure (name + verb + numeral + adjective + object). The sentences are grammatically correct but have no meaning.

B. Experimental conditions

In the present study, the proposed model was evaluated in conditions with (i) speech mixed with stationary or non-stationary interferers, (ii) speech in the presence of reverberation, and (iii) speech subjected to different types of non-linear processing. In all conditions, the models were evaluated using 100 sentences. The accuracy of the models was studied in terms of their Pearson's correlation with the data and the mean average error (MAE).

1. Influence of additive noise

The model was evaluated with three types of interfering noise: A speech-shaped noise (SSN), which was also used to fit the model; an 8-Hz sinusoidally amplitude-modulated (SAM) SSN with a modulation depth of 1; and the speech-like, but non-semantic, international speech test signal (ISTS; Holube *et al.*, 2010). CLUE sentences were mixed with the noises and the simulated speech reception thresholds (SRTs) were compared to the corresponding measured data from Jørgensen *et al.* (2013). A range of SNRs from −27 to 3 dB, with a step size of 3 dB, was considered to generate the inputs to the model.

2. Effect of reverberation

The CLUE sentences were mixed with SSN at different SNRs in the range from −9 to +9 dB, in 3-dB steps. Each mixture was convolved with impulse responses corresponding to reverberation times of $T_{60} = 0$, 0.4, 0.7, 1.3, and 2.3 s. The impulse responses were the same as the ones used in the study by Jørgensen and Dau (2011). They were created with the room acoustics software ODEON (Christensen, 2001) using a rectangular room of 3200 m^3, with the absorption coefficient of the walls adjusted such that the room had constant reverberation times across the octave bands from 63 to 8000 Hz. As the convolution operation introduces a time shift and a reverberant tail, while the correlation metric assumes zero lag between the two signals, a correction was carried out such that the clean speech and the reverberant noisy mixture were time aligned and had the same duration (by shifting the convolved signal and cropping its reverberant tail). The simulations were compared to the data presented in Jørgensen and Dau (2011).

3. Non-linear processing

Three types of non-linear processing were considered: (i) Noise reduction via spectral subtraction, (ii) a phase jitter distortion, and (iii) ITFS. The spectral subtraction processing was applied to the noisy speech (consisting of CLUE sentences and SSN) using the approach proposed by Berouti *et al.* (1979) which follows the equation:

$$\widehat{S(f)} = \sqrt{P_Y(f) - \kappa \widehat{P}_N(f)}, \tag{6}$$

where $\widehat{S(f)}$ is the enhanced magnitude spectrum of the noisy mixture after spectral subtraction. $P_N(f)$ and $P_Y(f)$ are the averaged power spectra of the noise alone and the original speech-plus-noise mixture, respectively (assuming access to the noise alone signal). Here, the average power spectrum was calculated as the mean from their corresponding short-term power spectral densities obtained using a Hanning window of 1024 samples and 50% overlap. Values for the over-subtraction factor, κ, of 0, 0.5, 1, 2, 4, and 8 were considered, with $\kappa = 0$ representing the unprocessed condition. The model was tested at SNRs ranging from −9 to +9 dB, in 3-dB steps. SRTs were simulated and compared to the data of Jørgensen and Dau (2011).

In the case of the phase-jitter distortion, the effect of small phase changes applied to the SSN noise and the CLUE speech mixture was studied. The phase jittering had the form

$$r(t) = \mathrm{Re}\{s(t)e^{j\Theta(t)}\} = s(t)\cos(\Theta(t)), \quad (7)$$

where $s(t)$ represents the input signal, $r(t)$ is the distorted signal, and $\Theta(t)$ is a random process with a uniform probability distribution between $[0, 2\alpha\pi]$, with α ranging between 0 and 1 (Elhilali *et al.*, 2003). The amount of phase jitter applied to the signal was thus controlled by the parameter α. Phase distortions corresponding to severity values of $\alpha = 0$, 0.125, 0.25, 0.375, 0.5, 0.625, 0.75, 0.875, and 1 were applied to the mixture on a sample-by-sample basis. The inputs to the model were in this case the clean signal and the noisy speech presented at an SNR of 5 dB and distorted with phase jitter. The simulations were compared to the data obtained in Chabot-Leclerc *et al.* (2014).

In the case of ITFS, the noise reduction technique proposed by Brungart *et al.* (2006) was considered, where an ideal binary mask (IBM) is applied to the T-F representation of the noisy speech. The IBM (Wang, 2005) is a binary matrix constructed by comparing the *a priori* known SNR within each T-F-unit to a local criterion (LC) such that

$$\mathrm{IBM}(t,f) = \begin{cases} 1 & \text{if } \mathrm{SNR}(t,f) > \mathrm{LC} \\ 0 & \text{otherwise.} \end{cases} \quad (8)$$

As in previous studies, the relative criterion (RC), defined as RC = LC-SNR, was used here to present the results. Unlike the LC, the RC can directly be related to the density of the IBM, i.e., the percentage of ones in the mask, regardless of the SNR of the noisy speech. In the present study, as in the experimental study of Kjems *et al.* (2009), Dantale II sentences were mixed with four different interferers: SSN, car-cabin noise (denoted as "Car"), noise produced by bottles on a conveyor belt ("Bottle"), and two people speaking in a cafeteria ("Café"). Two different SNR values were considered for the noisy mixture, corresponding to the 50% and 20% correct points on the respective psychometric function (obtained with the unprocessed noisy signals). As the psychometric functions are specific for each interferer, the two selected SNR values are different for each noise condition. Finally, IBMs were applied for eight different RC values per interferer and SNR. In total, 64 datapoints were considered (8 RC × 2 SNR × 4 interferers). The simulations were compared to the data presented in Kjems *et al.* (2009).

C. Mapping to speech intelligibility scores

Each speech material has a specific psychometric function relating SNRs to speech intelligibility scores. The logistic function of the proposed model [Eq. (5)] was fitted separately to the two speech corpora to account for their respective psychometric functions. For the CLUE corpus, the parameters of the logistic function were fitted to the data obtained with SSN. The fitted parameters were then kept constant across all experimental conditions considered in the present study that used the CLUE corpus.

Regarding the ITFS processing, the parameters of the logistic function were fitted to the Dantale II corpus that was used in this condition. Specifically, the parameters of Eq. (5) were fitted to the data obtained with the SSN interferer, 2 SNR values, and 8 LC values (16 data points). The resulting parameter values were then used when evaluating the model with the remaining interferers (Car, Bottle, and Café). The simulated psychometric functions obtained for the two speech materials are shown in Fig. 2. The corresponding parameters (a, b) are listed in Table I.

IV. RESULTS

A. Stationary and non-stationary interferers

The open symbols in Fig. 3 represent the measured SRTs from Jørgensen *et al.* (2013) for the conditions with the SSN (left), SAM (middle), and ISTS (right) interferers. The data show a masking release for the SAM and the ISTS conditions, as reflected by the decreased SRT values in these conditions compared to the one in the SSN reference condition. The simulations obtained with the proposed model, $\mathrm{sEPSM}^{\mathrm{corr}}$, are indicated by the filled black circles and the simulations obtained with mr-sEPSM and STOI are represented by the gray squares and the dark gray diamonds, respectively. The proposed model ($\rho = 0.97$, MAE = 1.85 dB) and the mr-sEPSM ($\rho = 0.99$, MAE = 1.16 dB) account well for the measured data, with $\mathrm{sEPSM}^{\mathrm{corr}}$ slightly underestimating the SRT in the SAM condition, whereas STOI ($\rho = 0.54$, MAE = 7.08 dB) does not capture the effect of a release from masking in the conditions with SAM and ISTS.

B. Reverberation

Figure 4 shows SRTs as a function of the room reverberation time. The open symbols show the data from Jørgensen and Dau (2011), which indicate a decrease of speech intelligibility with increasing reverberation time. The mr-sEPSM

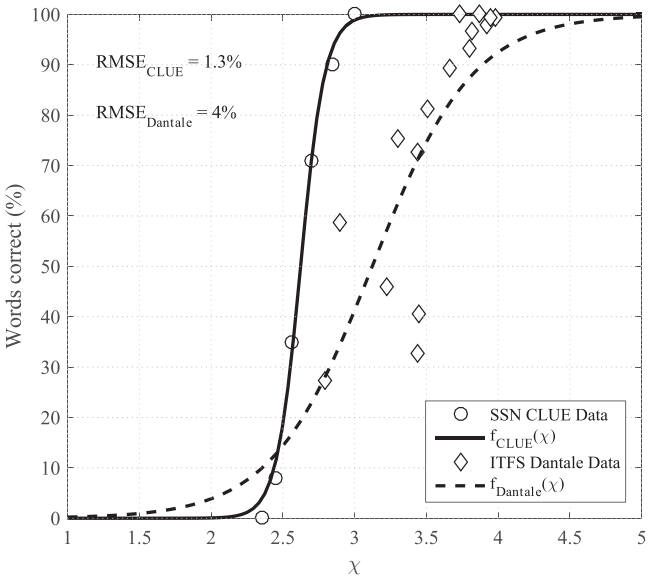

FIG. 2. Fitted psychometric functions for the two speech materials used in the present study: CLUE (circles; Nielsen and Dau, 2009) considering speech mixed with SSN, and Dantale II (diamonds; Wagener *et al.*, 2003) considering speech mixed with SSN and processed by ITFS. The solid line represents the resulting fitted psychometric function for the CLUE material and the dashed line indicates the corresponding fitted psychometric function for Dantale II.

TABLE I. Fitted values of the free parameters of the sigmoid function to map the sEPSM^{corr} predictions to human data. Two Danish speech materials were considered: CLUE (Nielsen and Dau, 2009) and Dantale II (Wagener *et al.*, 2003).

	a	*b*
CLUE	*−11.9*	*31.1*
Dantale II	*−2.9*	*9.0*

(gray squares) correctly describes the data ($\rho = 0.99$, MAE = 0.3 dB). However, both STOI (dark gray diamonds) and the proposed model sEPSM^{corr} (black circles) fail to account for the effect of reverberation. In fact, SRTs could only be calculated for the condition with $T_{60} = 0.4$ s, as the intelligibility scores obtained at different SNRs did not reach 50% for higher reverberation times. This implies that, for these models, the level of the noise has essentially no effect on the predicted intelligibility once reverberation is applied, resulting in very low intelligibility scores even for high SNRs. The light gray circles represent simulations obtained with a modified version of the model (sEPSM^{corr,LT}) which will be discussed further below (Sec. V C).

C. Non-linear processing

Figure 5 (top panel) shows the results obtained for noisy speech with applied spectral subtraction. It can be seen that all models can account for the decrease in intelligibility when increasing the over-subtraction factor, κ, as observed in the measured data (open symbols) from Jørgensen and Dau (2011). STOI (gray diamonds; $\rho = 0.94$, MAE = 0.3 dB) and mr-sEPSM (gray squares; $\rho = 0.95$, MAE = 0.4 dB) provide accurate predictions. The proposed model, sEPSM^{corr}, shows somewhat larger deviations from the data (black circles; $\rho = 0.82$, MAE = 0.6 dB), which are mainly due to

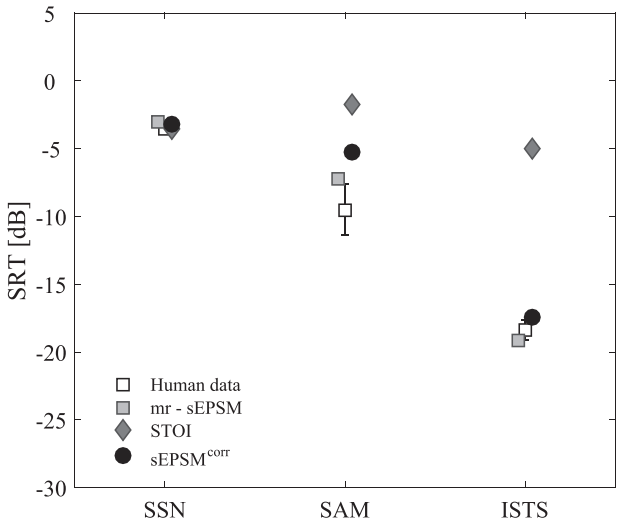

FIG. 3. SRT predictions for different additive noises: SSN, 8-Hz SAM-SSN and the ISTS. The gray squares correspond to mr-sEPSM predictions, whereas STOI and sEPSM^{corr} predictions are indicated by gray diamonds and black circles, respectively. The human data from Jørgensen *et al.* (2013) are shown as open squares, where the error bars represent plus/minus one standard deviation across listeners.

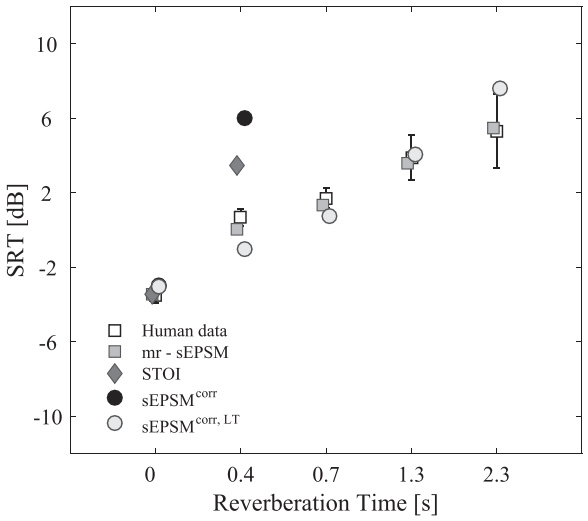

FIG. 4. SRT predictions for reverberation times of 0, 0.4, 0.7, 1.3, and 2.3 s. Gray squares correspond to mr-sEPSM predictions, whereas STOI and sEPSM^{corr} predictions are indicated by gray diamonds and black circles, respectively. Predictions obtained with an alternative version of the proposed model, using long-term integration (sEPSM^{corr,LT}), are indicated by the light gray circles. The human data from Jørgensen and Dau (2011) are shown as open squares, where the error bars represent plus/minus one standard deviation across listeners.

the fact that the model does not capture the initial increase in SRT from the unprocessed condition ($\kappa = 0$) to the processed condition ($\kappa = 0.5$). Nonetheless, sEPSM^{corr} does account for the decreasing speech intelligibility with increasing amount of noise reduction observed in the data.

The bottom panel of Fig. 5 shows the results for the phase-jitter condition. Intelligibility scores are shown, in percent, as a function of the phase jitter parameter α for a fixed SNR of 5 dB. The measured data (open symbols; Chabot-Leclerc *et al.*, 2014) show a non-monotonic pattern with minima of intelligibility at $\alpha = 0.5$ and $\alpha = 1$ and a local maximum at $\alpha = 0.75$. For the intelligibility minima at $\alpha = 0.5$ and $\alpha = 1$, the random phase values range between $[0, \pi]$ and $[0, 2\pi]$, respectively; after the cosine operation [cf. Eq. (7)], each sample of the original signal is thus multiplied by a random value between $[-1, 1]$, resulting in white noise modulated by the signal's envelope. The mr-sEPSM (MAE = 49.4%) fails in this condition. The model is essentially insensitive to this type of distortion. In contrast, both STOI (MAE 9%) and the proposed model sEPSM^{corr} (MAE 19%) account reasonably well[1] for the data, with the STOI model exhibiting more accurate predictions than the sEPSM^{corr} for $\alpha \geq 0.5$.

Figure 6 shows the effect of ITFS processing on speech intelligibility. The results are shown as intelligibility scores, in percent correct, as a function of the RC for the conditions with SSN (left panels), cafeteria noise (Café, second column), car noise (third column), and bottle noise (fourth column). The open symbols represent the measured data obtained by Kjems *et al.* (2009). In the first row, the results for noisy speech with an SNR corresponding to 50% intelligibility of the unprocessed speech are shown. The second row represents the results for an SNR corresponding to 20% speech intelligibility for each interferer.

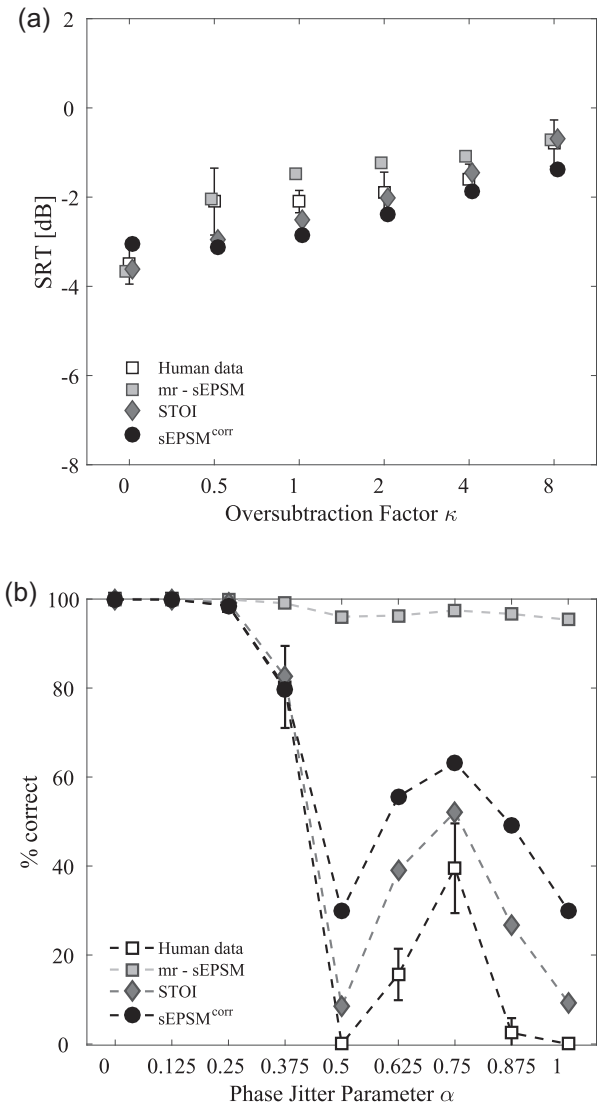

FIG. 5. Model predictions for two types of non-linear processing. (a) SRT predictions for noisy speech subjected to spectral subtraction (with over subtraction factors of $\kappa = 0$, 0.5, 1, 2, 4, and 8). (b) Intelligibility scores for noisy speech with phase jitter ($\alpha = 0$, 0.15, 0.25, 0.375, 0.5, 0.675, 0.75, 0.875, and 1). Gray squares correspond to mr-sEPSM predictions, whereas STOI and sEPSM^{corr} predictions are indicated by gray diamonds and black circles, respectively. The human data from Jørgensen and Dau (2011) and Chabot-Leclerc *et al.* (2014) are represented as open squares, where the error bars represent plus/minus one standard deviation across listeners.

STOI (gray diamonds) provides the most accurate predictions ($\rho = 0.95$, MAE 6.7%), followed by the proposed model sEPSM^{corr} (black circles; $\rho = 0.79$, MAE 12.1%) which has some limitations in the conditions with high RCs (i.e., low densities of the IBM) where intelligibility is overestimated, particularly in the conditions with the SSN and Car interferers. The mr-sEPSM fails in this condition (gray squares; $\rho = 0.39$, MAE 23.5%) and predicts very large intelligibility scores independent of the RC. The large deviation from the data for this model is due to the SNR_{env} metric not being monotonically related to the intelligibility scores for the different RCs.

Table II summarizes the simulation results obtained with all models in all conditions investigated here. The proposed model, sEPSM^{corr}, successfully describes most of the data. The model is able to account for the masking release obtained with fluctuating interferers where STOI fails. In addition, the model correctly describes the data obtained in the conditions with non-linear processing, as STOI, whereas the original mr-sEPSM fails in the phase jitter and ITFS conditions. However, as STOI, the proposed model fails to account for the effects of room reverberation whereas the original mr-sEPSM has been successful in this condition.

V. DISCUSSION

A. SNR vs correlation metrics

One of the biggest advantages of the proposed model, in comparison to previous versions of the sEPSM, is its ability to account for phase jitter distortions and the effects of ITFS. In contrast to the SNR-based metric, the correlation metric is able to capture the effects of non-linear distortions. Phase jitter is a distortion that affects the phase of the signal by adding random phase shifts. The fact that the envelope of the signal is mostly unaffected by such a distortion makes models based on the SNR in the envelope domain, like the mr-sEPSM or the classic STI, insensitive to changes in the intelligibility of phase jittered speech. The study by Chabot-Leclerc *et al.* (2014) showed that, in order to account for the data in such conditions, the sEPSM would require an additional stage that evaluates speech information across frequency bands. In contrast, the sEPSM^{corr} does not need an explicit across-frequency analysis (nor does STOI). By assessing the clean signal and the distorted mixture as inputs to the model, where the original phase information is preserved in the clean signal, the correlation analysis is able to quantify the signal degradation effectively, linking it to speech intelligibility.

In the case of ITFS, the mr-sEPSM largely overestimates the intelligibility of the processed speech. This is most likely due to the introduction of abrupt modulations (caused by imposing the binary masks on the speech mixture), which are interpreted as being beneficial to speech intelligibility by the model. The predicted intelligibility scores of the correlation-based models, STOI and sEPSM^{corr}, are much closer to those observed in the human data. The sEPSM^{corr} predictions deviate most from the human data in cases where the mask density is low, i.e., when RC > 20 dB which corresponds to 1% of ones in the mask. When using such a strict criterion, only very few T-F elements of the noisy mixture are retained after applying the mask which substantially reduces the intelligibility of the noisy speech. The model overestimates the intelligibility scores in this extreme case. STOI provided the best predictions in this condition. However, it should be noted that STOI was designed specifically to account for the set of data presented in Kjems *et al.* (2009), such that the window size and other model parameters were tailored to fit these data, as described in Taal *et al.* (2011).

B. Role of the temporal analysis and integration in conditions of fluctuating interferers

The proposed model can account for the reduced SRTs (i.e., better intelligibility) in the presence of fluctuating interferers compared to those obtained in stationary noise. In

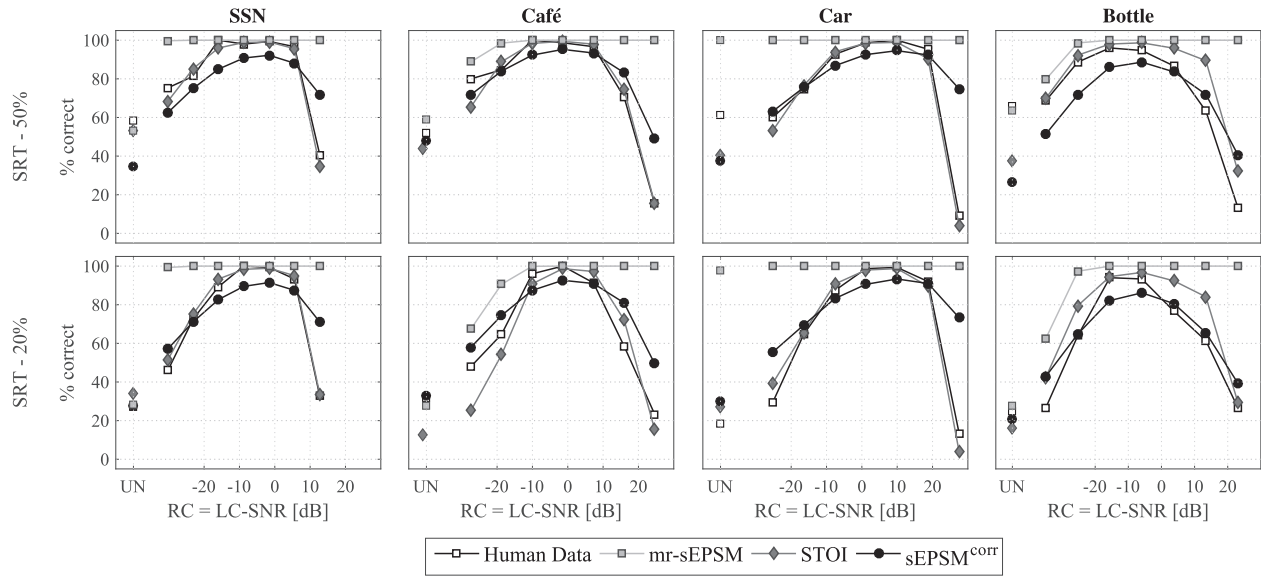

FIG. 6. Intelligibility scores for ITFS processed speech with four different interferers (columns) and two SNRs (rows). The gray squares show predictions obtained with mr-sEPSM, whereas STOI and sEPSM$^{\text{corr}}$ predictions are indicated by gray diamonds and black circles, respectively. The human data from Kjems *et al.* (2009) are shown as open squares.

contrast, STOI is not able to predict the influence of fluctuating interferers, despite the fact that both models employ a correlation-based back end. Jørgensen *et al.* (2013) demonstrated that a multi-resolution analysis is crucial in the mr-sEPSM model to account for a masking release. Since the sEPSM$^{\text{corr}}$ uses a similar approach, its ability to predict the effects of fluctuating interferers is likely also due to the temporal resolution in the analysis, which assumes window durations inversely proportional to the center frequency of the modulation filter.

To study the effect of the temporal resolution assumed in the sEPSM$^{\text{corr}}$, different versions of the model were considered, which used window sizes that were *constant* across modulation filters. Durations of 20, 50, 100, 300, 500, and 1000 ms were compared to the multi-resolution approach (where multiple time constants are applied in parallel), as well as to a long-term model which analyzes the full-duration input signals. The different model versions were tested in conditions of additive noise (as in Sec. III B 1). Figure 7 shows the results of the simulations, in terms of the root-mean-square error (RMSE) resulting from each model's predictions with respect to the measured data. The left-most filled circle in the figure indicates the result obtained with the current version of the model, i.e., assuming multiple time constants as reflected in the multi-resolution approach. The remaining filled circles show the results for the different fixed-duration windows. It can be seen that an increase of the window duration led to an increase of the RMSE, with a strong effect particularly at durations above 100 ms. The results are consistent with the observation that STOI (which uses an analysis window of 380 ms) fails in these conditions. Furthermore, the long-term model (right-most filled circle) showed the highest error value, consistent with the findings of Jørgensen *et al.* (2013).

The way in which the model's back end integrates the correlation values across time windows also has an impact on the simulation results. The proposed multiple-looks integration strategy [Eq. (3)] has the implicit effect of emphasizing high-frequency modulation filters ($f_{c,\text{mod}} > 32$ Hz). Since the time windows are shorter for high-frequency modulation filters, the model uses substantially more windows for the analysis of these modulation bands, compared to the low-frequency modulation bands. This implies that using Eq. (3) to accumulate the correlation values across time results in a stronger contribution of the high-frequency modulation channels to the model's final metric.

To further analyze the influence of the high-frequency modulations, an alternative model version was considered that linearly averages the correlation values across time windows, instead of using Eq. (3), thus giving equal weight to each modulation band. This alternative integration was again tested in conditions of additive noise. The open circles in Fig. 7 show the results obtained with the linear averaging. This metric leads to a large RMSE (of about 7 dB) when combined with the multi-resolution processing (left-most open circle in Fig. 7). The time averaging strategy was also combined with fixed-duration analysis windows yielding

TABLE II. Results of the statistical evaluation of mr-sEPSM, STOI, and sEPSM$^{\text{corr}}$. MAE and Pearson's correlation (ρ) values are provided. "—" indicates no value was obtained for that condition/model. "*" indicates that values were obtained with the sEPSM$^{\text{corr,LT}}$ model (see Sec. V C).

	mr-sEPSM		STOI		sEPSMcorr	
	ρ	MAE	ρ	MAE	ρ	MAE
Additive noise	0.99	1.16 dB	0.54	7.08 dB	0.97	1.85 dB
Reverberation	0.99	0.31 dB	—	—	0.94*	1.09 dB*
Spectral subtraction	0.95	0.36 dB	0.94	0.29 dB	0.82	0.60 dB
Phase jitter	—	49.4%	—	9.0%	—	19.0%
ITFS	0.39	23.5%	0.95	6.7%	0.79	12.1%

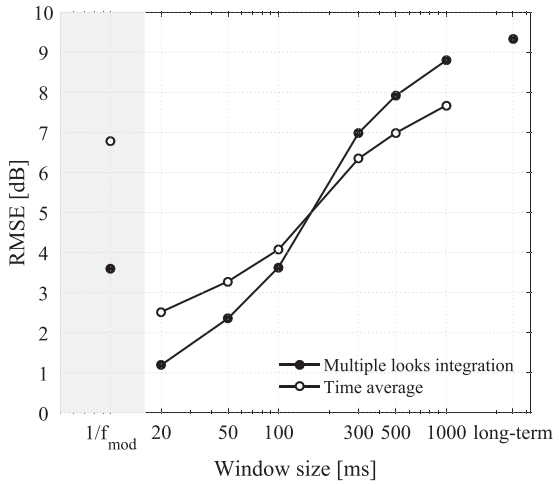

FIG. 7. RMSE as a function of the window size calculated for model predictions of speech with additive noise in relation to the human data from Jørgensen *et al.* (2013). The filled circles indicate the results using multiple-looks integration [Eq. (3)] of the correlation metric across time frames. The open circles show predictions obtained with a modified model where a linear averaging of the correlation metric across time frames was applied. On the left (gray area), the result for the proposed multi-resolution model is shown. On the right, the result for a long-term version of the model is shown.

similar results as the original model (i.e., high errors for window durations above 100 ms). This is consistent with the predictions obtained with STOI that uses linear averaging of the correlation values combined with a fixed window size.

The simulations shown in Fig. 7 suggest that short time windows (<50 ms) lead to better predictions than the multi-resolution processing. However, this is only the case for the condition of fluctuating interferers. In the case of non-linear processing, especially spectral subtraction, the short time windows strongly degraded the predictions (not shown). In addition, the computation time was substantially increased when shorter windows were used, which further motivated the choice of the multi-resolution approach.

C. Limitations in reverberant conditions

The proposed model cannot account for the effects of room reverberation. When reverberation is applied, the level of the noise has essentially no effect on the predicted intelligibility of the speech mixture, i.e., the model produces very low intelligibility scores, even for high SNRs. This is consistent with the results from previous studies that showed that correlation-based models are generally not adequate to predict the intelligibility of reverberant speech (Goldsworthy and Greenberg, 2004; Taal *et al.*, 2011). Furthermore, Taal *et al.* (2011) argued that the use of short windows (including their window choice of 380 ms in STOI) could have a negative impact on the performance of correlation-based models under reverberation, although they did not elaborate on this argument. Distortions produced by reverberation, namely, temporal smearing and self-masking due to reflections, cannot be captured by short windows. This also applies to the multi-resolution approach of the sEPSM^{corr}, in which the processing of the shorter windows of the high-frequency modulation bands is emphasized by the multiple-looks integration strategy. With the current modeling approach, it was not possible to find an implementation of the sEPSM^{corr} that could account both for the effects of room reverberation and for the effects of dip listening in fluctuating interferers. While the latter condition requires that the model uses short time windows and an emphasis of high-frequency modulations, longer time constants and low-frequency modulations seem to be more crucial in reverberant conditions.

To further analyze the limitations of the correlation metric when calculated in short time intervals, an alternative model that employs a long-term correlation of the internal signal representations across the full-duration input signals was considered. The resulting metric was not three-dimensional as in the proposed model (with a correlation value obtained for each time window, modulation filter, and auditory filter), but two-dimensional, producing only one correlation value per modulation channel and auditory channel. In this realization of the model, a time-integration strategy was not required. All the remaining model stages remained unchanged, with the compressive stage being specifically critical in this condition. The results obtained with the long-term model are indicated as light gray circles in Fig. 4. It can be seen that this long-term approach (denoted sEPSM^{corr,LT}) accurately predicts the human data for reverberant speech ($\rho = 0.94$, MAE = 1.1 dB). This demonstrates that a correlation-based analysis of the internal representations combined with the sEPSM front end can convey information about the intelligibility of reverberant speech, as long as it is not combined with short time windows. However, this version of the model would clearly fail in other conditions that require short time constants (as indicated by the right-most point in Fig. 7); thus, it is offered here as an alternative path to account for the intelligibility of reverberant speech but not as a general model to account for all conditions considered in the present study.

VI. CONCLUSION

A new speech intelligibility prediction model was presented. The model operates on the clean unprocessed speech and the noisy mixture and combines the front end of the mr-sEPSM model (Jørgensen *et al.*, 2013) with a correlation-based back end similar to the one employed in the STOI measure (Taal *et al.*, 2011). It was demonstrated that this "hybrid" model, named sEPSM^{corr}, accounts for the effects of stationary and fluctuating noise interferers as well as for various effects of non-linear distortions, such as spectral subtraction, phase jitter, and ITFS processing. The predictive power of the model was thus broader than that of the original mr-sEPSM, which failed in the phase-jitter and I conditions, and also broader than that of STOI, which failed to account for the effect of fluctuating interferers. However, the predictions of the proposed model were in some conditions slightly less accurate than those of one or both of the source models. Furthermore, similar to other models with a correlation-based back end (including STOI), the sEPSM^{corr} in its current form failed to account for the effects of room reverberation. An alternative model design was provided to account for such reverberant conditions. Overall, the proposed model might be useful for evaluating a large variety of interferences

and distortions on speech intelligibility, including effects of hearing impairment and hearing-instrument signal processing.

A MATLAB implementation of the model is available at: https://bitbucket.org/heliaib/sepsm-corr.

ACKNOWLEDGMENTS

The authors would like to thank Søren Jørgensen for comments on earlier versions of the model, Alexandre Chabot-Leclerc for providing the data in the phase-jitter experiment, and Claus Forup Corlin Jespersgaard for providing the ITFS data. We also thank two anonymous reviewers for their very helpful and supportive comments. This work was supported by the Oticon Centre of Excellence for Hearing and Speech Sciences (CHeSS), the EU FET Grant TWO!EARS, ICT-618075, and by the Centre for Applied Hearing Research (CAHR). The Auditory modeling toolbox (http://amtoolbox.sourceforge.net/) has made scripts for STOI and sEPSM models online available for free.

[1]Only the MAE was considered here since the Pearson's correlation does not provide an informative error metric in this condition.

Allen, J. B. (**1996**). "Harvey Fletcher's role in the creation of communication acoustics," J. Acoust. Soc. Am. **99**(4), 1825–1839.

ANSI (**1969**). ANSI S3.5-1969, *Methods for Calculation of the Articulation Index* (American National Standards Institute, New York).

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

Berouti, M., Schwartz, R., and Makhoul, J. (**1979**). "Enhancement of speech corrupted by acoustic noise," IEEE Int. Conf. Acoust., Speech, Signal Process. **4**, 208–211.

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**(6), 4007–4018.

Chabot-Leclerc, A., Jørgensen, S., and Dau, T. (**2014**). "The role of auditory spectro-temporal modulation filtering and the decision metric for speech intelligibility prediction," J. Acoust. Soc. Am. **135**(6), 3502–3512.

Christensen, C. L. (**2001**). "Odeon a design tool for auditorium acoustics, noise control and loudspeaker systems," Proc. Reprod. Sound **17**, 137–144.

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**(5), 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**(5), 2906–2919.

Dau, T., Verhey, J., and Kohlrausch, A. (**1999**). "Intrinsic envelope fluctuations and modulation-detection thresholds for narrow-band noise carriers," J. Acoust. Soc. Am. **106**(5), 2752–2760.

Drullman, R., Festen, J. M., and Plomp, R. (**1994**). "Effect of reducing slow temporal modulations on speech reception," J. Acoust. Soc. Am. **95**, 2670–2680.

Elhilali, M., Chi, T., and Shamma, S. A. (**2003**). "A spectro-temporal modulation index (STMI) for assessment of speech intelligibility," Speech Commun. **41**, 331–348.

Ewert, S. D., and Dau, T. (**2000**). "Characterizing frequency selectivity for envelope fluctuations," J. Acoust. Soc. Am. **108**(3), 1181–1196.

Ewert, S. D., and Dau, T. (**2004**). "External and internal limitations in amplitude-modulation processing," J. Acoust. Soc. Am. **116**(1), 478–490.

Festen, J., 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**(4), 1725–1736.

French, N. R., and Steinberg, J. C. (**1947**). "Factors governing the intelligibility of speech sounds," J. Acoust. Soc. Am. **19**(1), 90–119.

Goldsworthy, R., and Greenberg, J. (**2004**). "Analysis of speech-based speech transmission index methods with implications for nonlinear operations," J. Acoust. Soc. Am. **116**(6), 3679–3689.

Holube, I., Fredelake, S., Vlaming, M., and Kollmeier, B. (**2010**). "Development and analysis of an International Speech Test Signal (ISTS)," Int. J. Audiol. **49**, 891–903.

Holube, I., and Kollmeier, B. (**1996**). "Speech intelligibility prediction in hearing-impaired listeners based on a psychoacoustically motivated perception model," J. Acoust. Soc. Am. **100**, 1703–1716.

Houtgast, T., Steenekem, H. J. M., and Plomp, R. (**1980**). "Predicting speech intelligibility in rooms from the modulation transfer function. I. General room acoustics," Acustica **46**(1), 60–72.

ISO (**2005**). ISO 389-7, *Reference Zero for the Calibration of Audiometric Equipment—part 7: Reference Threshold of Hearing under Free-field and Diffuse-field Listening Conditions* (International Organization for Standardization, Geneva, Switzerland).

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**(3), 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**(1), 436–446.

Kates, J., and Arehart, K. (**2005**). "Coherence and the speech intelligibility index," J. Acoust. Soc. Am. **117**(4), 2224–2237.

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.

Kohlrausch, A., Fassel, R., and Dau, T. (**2000**). "The influence of carrier level and frequency on modulation and beat-detection thresholds for sinusoidal carriers," J. Acoust. Soc. Am. **108**(2), 723–734.

Kryter, K. D. (**1962**). "Methods for the calculation and use of the articulation index," J. Acoust. Soc. Am. **34**(11), 1698–1702.

Langner, G., and Schreiner, C. E. (**1988**). "Periodicity coding in the inferior colliculus of the cat. II. Topographical organization," J. Neurophysiol. **60**(6), 1823–1840.

Ludvigsen, C., Elberling, C., and Keidser, G. (**1993**)."Evaluation of a noise reduction method: Comparison between observed scores and scores predicted from STI," Scan. Audiol. **22**(38), 50–55.

Ludvigsen, C., Elberling, C., Keidser, G., and Poulsen, T. (**1990**). "Prediction of intelligibility of non-linearly processed speech," Acta Otolaryngol. Suppl. **469**, 190–195.

Nielsen, J. B., and Dau, T. (**2009**). "Development of a Danish speech intelligibility test," Int. J. Audiol. **48**, 729–741.

Patterson, R. D., Nimm-Smith, I., Holdworth, J., and Rice, P. (**1987**). "An efficient auditory filterbank based on the gammatone function," presented at the Speech-Group Meeting of the Institute of Acoustics on Auditory Modelling.

Payton, K. L., and Braida, L. D. (**1999**). "A method to determine the speech transmission index from speech waveforms," J. Acoust. Soc. Am. **106**(6), 3637–3648.

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.

Rhebergen, K. S., Versfeld, N. J., and Dreschler, W. A. (**2009**). "The dynamic range of speech, compression, and its effect on the speech reception threshold in stationary and interrupted noise," J. Acoust. Soc. Am. **126**(6), 3236–3245.

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. **19**(7), 2125–2136.

Verhey, J. L., Dau, T., and Kollmeier, B. (**1999**). "Within-channel cues in comodulation masking release (CMR): Experiments and model predictions using a modulation-filterbank model," J. Acoust. Soc. Am. **106**(5), 2733–2745.

Viemeister, N. F., and Wakefield, G. H. (**1991**). "Temporal integration and multiple looks," J. Acoust. Soc. Am. **90**(2), 858–865.

Wagener, K., Josvassen, J. L., and Ardenkjaer, R. (**2003**). "Design, optimization and evaluation of a Danish sentence test in noise," Int. J. Audiol. **42**(1), 10–17.

Wang, D. L. (**2005**). "On ideal binary mask as the computational goal of auditory scene analysis," in *Speech Speech Separation by Humans and Machines* (Kluwer, Dordrecht, the Netherlands), pp. 181–197.