Application of a short-time version of the Equalization-Cancellation model to speech intelligibility experiments with speech maskers

Rui Wan, Nathaniel I. Durlach, and H. Steven Colburn^{a)}

Hearing Research Center and Department of Biomedical Engineering, Boston University, Boston, Massachusetts 02215

(Received 24 April 2013; revised 3 June 2014; accepted 6 June 2014)

A short-time-processing version of the Equalization-Cancellation (EC) model of binaural processing is described and applied to speech intelligibility tasks in the presence of multiple maskers, including multiple speech maskers. This short-time EC model, called the STEC model, extends the model described by Wan *et al.* [J. Acoust. Soc. Am. **128**, 3678–3690 (2010)] to allow the EC model's equalization parameters τ and α to be adjusted as a function of time, resulting in improved masker cancellation when the dominant masker location varies in time. Using the Speech Intelligibility Index, the STEC model is applied to speech intelligibility with maskers that vary in number, type, and spatial arrangements. Most notably, when maskers are located on opposite sides of the target, this STEC model predicts improved thresholds when the maskers are modulated independently with speech-envelope modulators; this includes the most relevant case of independent speech maskers. The STEC model describes the spatial dependence of the speech reception threshold with speech maskers better than the steady-state model. Predictions are also improved for independently speech-modulated noise maskers but are poorer for reversed-speech maskers. In general, short-term processing is useful, but much remains to be done in the complex task of understanding speech in speech maskers. © 2014 Acoustical Society of America.
[\[http://dx.doi.org/10.1121/1.4884767\]](http://dx.doi.org/10.1121/1.4884767)

PACS number(s): 43.66.Pn, 43.66.Ba, 43.66.Dc [JFC]

Pages: 768–776

I. INTRODUCTION

A modified version of the Equalization-Cancellation (EC) model of Durlach (1963, 1972) is applied to speech intelligibility experiments that varied both the type and direction of interfering sounds with special attention given to conditions that were not well-described by an earlier model (Wan *et al.*, 2010). Specifically, the application of an earlier version of the EC model, referred to here as the steady-state Equalization-Cancellation model (SSEC), predicted poorer than observed relative performance in cases in which multiple interfering speech sounds were located both to the right and to the left of the target speech. In the current revision of the EC model, referred to here as the “short-time Equalization Cancellation” or STEC model, attention is given to stimuli comprising speech targets masked by multiple maskers. Maskers include speech-shaped noises, modulated speech-shaped noises, and speech maskers, and both the number of maskers and the directions of the maskers were varied for each type. The focus is on the spatial dependence of intelligibility thresholds on the locations of the maskers. In this STEC model, the equalization parameters are allowed to vary not only over frequency but also over time. More specifically, in the current study, the STEC model is again applied to results from the experiments of Hawley *et al.* (2004) and of Marrone *et al.* (2008), hereafter called the Hawley experiments and the

Marrone experiments. The Hawley experiments varied the type and directions of simultaneous, independent interfering sounds, including other speech signals, reversed-speech signals, speech-shaped noises, and speech-shaped noises that are independently modulated by envelope waveforms generated from (wideband) speech signals. The Marrone experiments included a speech target straight ahead masked by a pair of independent speech signals that are either co-located with the target or presented from opposite sides of the head (i.e., at 0°, ±15°, ±45°, or ±90° azimuth). These Marrone experiments also included conditions with reversed-speech maskers. The most significant effect of the short-time processing is the possibility for the model to cancel the dominant masker over time when the direction of the dominant masker varies in time as it does for well-separated and independently modulated maskers. These separated cases, for speech maskers and for independently modulated noise, are better fit by the STEC model. Unexpectedly, the separated cases for reversed-speech maskers in the Hawley experiment was better fit by the steady-state model.

The general organization of this paper is as follows: Sec. II reviews previous work on related topics; Sec. III describes the STEC model in detail; Sec. IV compares the predictions of the STEC model (as well as those of our earlier SSEC model) to empirical data; and Sec. V discusses these results as evidence for time-varying binaural processing parameters and the importance of informational masking (IM) (cf. Kidd *et al.*, 2008) in many speech-on-speech masking conditions (Wan, 2011).

^{a)}Author to whom correspondence should be addressed. Electronic mail: colburn@bu.edu

II. BACKGROUND

As outlined in our previous paper, the application of binaural models to speech intelligibility with spatially distributed maskers has a long history with early quantitative modeling contributions made by Zurek (1992). Zurek based his work on Levitt and Rabiner (1967a,b) and used the articulation index (AI) measure (Kryter, 1962; ANSI, 1969). Using a similar approach to data from multiple maskers, Culling and colleagues (Culling *et al.*, 2004) showed general agreement with the basic processing of the EC model. In a significant extension to reverberant rooms and to listeners with hearing impairments, Beutelmann and Brand (2006) applied the EC model with the Speech-Intelligibility Index (SII) measure (ANSI, 1997) to predict the intelligibility of speech masked by a single masker for different masker locations. They later (Beutelmann *et al.*, 2009) used their model to describe interesting experiments that showed independent processing across frequency bands. In another series of papers, Lavandier and colleagues (e.g., Lavandier and Culling, 2010; Lavandier *et al.*, 2012) modeled speech intelligibility with multiple wideband noise maskers, notably including effects of reverberation. They demonstrated the utility of the EC model to describe these cases with explicit attention to the head-related transfer functions and the accumulation of information across frequency. The models used in all of these studies were modified EC models with differences in the detailed assumptions in some studies. In general, these studies showed compatibility of the EC model processing with speech intelligibility data for various noise masking conditions.

More directly relevant to the short-time processing in the current study, there have also been applications of short-time processing algorithms to speech intelligibility predictions for monaural processing (Rhebergen and Versfeld, 2005; Rhebergen *et al.*, 2006) and for binaural processing (Beutelmann *et al.*, 2010). More specifically, Rhebergen and Versfeld (2005) extended the SII approach to explicitly include the temporal fluctuations of signal-to-noise ratio that are inherent with signals that fluctuate in time, such as amplitude-modulated noise or speech (even during single words). These authors analyzed the signal-to-noise ratio within temporal windows so that these fluctuations could be explicitly included in the SII analysis. A time-varying binaural processor was also assumed by Beutelmann, Brand, and Kollmeier (2010), who used the SII approach combined with binaural processing in a significant extension of their previous work. They again included reverberation and hearing-impaired listeners and varied the angle of the interfering speech-modulated noise relative to the target speech. In another study relevant to time-dependent processing, selective time-frequency processing is prominent in the experiments and analyses of Brungart and Iyer (2012). They find evidence for selective listening to the ear with the better signal-to-noise ratio (SNR) for each time-frequency tile. Most recently, a detailed analysis of short-time processing in speech intelligibility experiments with modulated maskers was presented for monaural conditions by Meyer and Brand (2013). These studies all suggest that our processing of masked speech takes advantage of temporal fluctuations. The study presented in this report applies a short-time

version of the EC model to speech intelligibility with multiple maskers, including speech maskers, focusing on the advantages of binaural processing that adjusts to the waveform properties as they vary in time. Specifically, the model is an extension of the model in our previous study (Wan *et al.*, 2010) with the addition of short-time processing capabilities. We believe that even though this is a clearly oversimplified model, the ability to analyze the effect of short-term processing on speech masked by speech may help in the development of more sophisticated models in the future.

III. SPECIFICATION OF THE STEC MODEL

The STEC model, shown in Fig. 1, has the same general model components as those included in the SSEC model (Wan *et al.*, 2010). The inputs to the model are the acoustic waveforms arriving at the two ears; each of these waveforms is passed through a filter bank consisting of one-third-octave bandpass filters (BPFs); the signals in both the binaural pathway and the two monaural pathways are corrupted by time-varying jitter in both time and amplitude, applied independently in each frequency channel [all jitters are zero-mean Gaussian processes with standard deviations $\sigma_\delta = 105\mu\text{s}$ and $\sigma_\epsilon = 0.25$ for the time and amplitude jitters, respectively, as chosen by Durlach (1963) to match binaural tone-in-noise detection data]; and the equalization and cancellation (EC) process occurs in each time-frequency unit independently.

The STEC model differs from the SSEC model (Wan *et al.*, 2010) in that the EC process in each frequency channel varies over time. More specifically, it is assumed that there exists a sliding time window that is rectangular in shape and 20 ms in duration and that there is an overlap between adjacent time windows of 50% (i.e., the duration of the overlap between adjacent time windows is 10 ms). The 20-ms duration of the window preserves more than enough envelope information within bands to allow good speech intelligibility (Drullman *et al.*, 1994).

The EC processing is assumed to take place independently in each time-frequency unit. A time-frequency unit is denoted $U(i, j)$, where i denotes the i th frequency channel and j the j th time frame. The equalization process within each time-frequency unit involves choosing values of the internal interaural time difference (ITD) and the internal interaural level difference (ILD), which are represented by τ and α , respectively. The optimal values of these equalization parameters for unit $U(i, j)$ (i.e., values that minimize the canceled noise component) are represented in the following equations by $\tau_o(i, j)$ and $\alpha_o(i, j)$, respectively, and they are chosen according to the equations

$$\begin{aligned}\tau_o(i, j) &= \underset{\tau}{\operatorname{argmax}} \{ \rho_{ij}(\tau) \}, \quad |\tau| < \frac{\pi}{\omega_i}, \text{ and} \\ \alpha_o(i, j) &= \sqrt{\frac{E_{NL_{ij}}}{E_{NR_{ij}}}},\end{aligned}\tag{1}$$

where $\rho_{ij}(\tau)$ is the normalized cross-correlation function of the jittered masker for the left and right ears within the unit $U(i, j)$, $E_{NL_{ij}}$ and $E_{NR_{ij}}$ are the masker energies for the left and right ears in $U(i, j)$, and ω_i is the center frequency in the

FIG. 1. EC model of Durlach (1963) modified to include time-varying jitter. In the short-time EC (STEC) model used here, the equalization parameters τ_o and α_o are adjusted to optimize cancellation within each time window.

i th frequency channel. The binaural output $Y_{ij}(t)$ for the unit $U(i, j)$ after the cancellation is given by

$$Y_{ij}(t) = W_j(t) \left\{ \frac{1}{\sqrt{\alpha_o(i, j)}} L_i \left(t + \frac{\tau_o(i, j)}{2} \right) - \sqrt{\alpha_o(i, j)} R_i \left(t - \frac{\tau_o(i, j)}{2} \right) \right\}, \quad (2)$$

where $L_i(t)$ and $R_i(t)$ denote the jittered waveform in the i th frequency band for the left and right ears, and $W_j(t)$ denotes the rectangular window function for the j th time frame, i.e.,

$$W_j(t) = \begin{cases} 1, & (j * 10) \text{ ms} \leq t < (j * 10) + 20 \text{ ms} \\ 0, & \text{otherwise.} \end{cases} \quad (3)$$

Subsequent to the EC processing for the units $U(i, j)$ in a given frequency band, temporal reconstruction is performed to generate a binaurally computed signal $B_i(t)$ in that frequency band according to the equation

$$B_i(t) = \frac{1}{2} \sum_j Y_{ij}(t). \quad (4)$$

Performance is computed from the SNRs for each frequency-band waveform $B_i(t)$ computed from the entire waveform. Thus the local temporal variation in SNR is not explicitly considered. The short-time variation in SNR could be considered with an explicit listen-in-the-gaps model, but this is not done in the current work.

The STEC decision device operates in a manner identical to the decision device in the SSEC model described previously (Wan *et al.*, 2010). Specifically, for each frequency band, it compares the SNRs between the output of the binaural processing and the two monaural outputs and selects the output with the best ratio. These optimal SNRs within each frequency band are then combined [by a weighted linear combination using the SII weights (ANSI, 1997)] (for SNRs between -15 and $+15$ dB) to form a value for the SII at each target level. This computed SII value, considered as a function of target level for fixed maskers, is used to predict the target speech intelligibility threshold using the same techniques as those used previously in the SSEC model (Wan *et al.*, 2010), i.e., by choosing a criterion value for the SII with the criterion held fixed for each type and number of maskers. Given this criterion value, predictions for other azimuths are calculated with the same criterion (i.e., no further fitting is

performed). Again, because of the difficulty of obtaining equations for the predictions in closed form, predictions are derived using simulation techniques. As in the work with the SSEC model, the differences among repeated simulations for the same condition were found to be so small [the standard error across repetitions (with statistically independent jitters) being less than 0.1 dB] that no error bars are plotted in the graphs showing the theoretical predictions of the data.

To apply the STEC model to the variety of types of maskers included in this study, a criterion value of the SII must be selected to predict the threshold for intelligibility.¹ Consistent with the original SSEC model, the SII threshold was chosen for each masker condition (i.e., for each type and number of maskers) to match one of the spatial arrangements of maskers. Although it might be possible (and we intend to explore this in the future) to model the advantages of listening in the gaps as well as the confusability of speech targets with speech maskers (i.e., informational masking) explicitly, this is not done in the current paper. Specifically, we chose the SII threshold to match the most-separated spatial condition for each number and type of masker. Thus the focus of this study is the dependence of the threshold for intelligibility on the *spatial* arrangements of the maskers and not the dependence on the type or number of maskers. In general, the choice of the reference condition as the condition with the maskers separated widely for each number and type of masker is an attempt to separate the temporal factors and the confusability factors from the binaural interaction effects that are explicit in the EC model. We plan to address these temporal and confusable factors in future work.

IV. APPLICATION OF STEC MODEL TO PSYCHOACOUSTIC DATA

In this section, the predictions of the STEC model are presented and compared to empirical results for a variety of speech intelligibility tasks as measured in the Hawley experiments and the Marrone experiments (Hawley *et al.*, 2004; Marrone *et al.*, 2008). These studies included multiple simultaneous maskers, a variety of masker azimuths, and several types of maskers created with noise and speech. We also explicitly compared the predictions of the STEC model to those of the SSEC model studied previously.

A. Application to the data of Hawley *et al.*

In the Hawley experiments (Hawley *et al.*, 2004), speech inputs were taken from the Harvard IEEE sentences

(Rothauser *et al.*, 1969), and maskers were sets of one, two, or three independent waveforms from various simulated directions. Masker types included speech-shaped noise (SSN); wideband-modulated speech-shaped noise (ModSSN), where the modulators were extracted from speech waveforms in the same IEEE corpus; speech signals (speech) chosen from the same corpus and spoken by the same (male) speaker; and time-reversed versions of the speech signals (RevSpeech). The speech-masker sentences in this study were chosen anew for each threshold, and the content, number, and directions of the interferers were fixed throughout the run. Also, during each measurement, the masker sentences were displayed on a monitor in front of the listener. In conditions that contained a non-speech interferer, “unintelligible” was printed on the monitor screen. All stimuli were calculated from anechoic HRTFs and presented over headphones.

The measured binaural speech reception thresholds (SRTs) for all these cases are plotted in Fig. 2 as squares with error bars showing one standard error. Within each row of panels, data from one type of interferer are shown as identified by the abbreviations in parentheses below each set of panels. The sets of numbers along the abscissa of each panel indicate the angle(s) of the interferer(s), arranged so that single-interferer data are shown in the leftmost panels, the two-interferer data are shown in the middle panels, and the three-interferer data are shown in the rightmost panels. The curves calculated from the models are shown as solid curves for the STEC model described in the preceding text and as dashed curves for the SSEC model of Wan *et al.* (2010).

For each of the graphs shown in Fig. 2, the SII criterion was chosen so that data and model are matched for the most separated target-masker configurations (the rightmost point along the axis in each case). The number in the upper right corner of each panel gives the SII criterion used to match this rightmost point within that panel. As noted in the preceding text, the criterion values are chosen for each experiment to accommodate differences in masker type and number of maskers, factors that are not included in the waveform processing basis of the model. The model (the original EC or the short-time EC) does not contain any mechanism for adjusting the threshold to compensate for the challenges of understanding speech in stimuli that sound more like speech; for example, informational masking (IM) effects are not explicitly included in the EC model. In considering the SII criterion values, note that the SII versus SNR slope (e.g., as plotted in Fig. 2 of Wan *et al.*, 2010) is approximately 0.03 SII units per decibel. Thus for example, a change in the SII criterion of 0.016 (as in the change from 0.374 to 0.358 between the upper two rows for the three-masker cases) corresponds to only about 0.5 dB in the SRT, so that the differences in the fitted SIIs for SSN and ModSSN are very small as expected according to Rhebergen and Versfeld (2005).

The comparisons of predictions and data in Fig. 2 show that the model predictions have a spatial dependence very similar to the data when the co-located condition is excluded. More specifically, the top row shows the SSN case, and all the data are well fit by both models, and there is no advantage to short-time EC operations as expected. The

FIG. 2. Simulated and measured binaural SRTs for different types (rows) and numbers (columns) of maskers. Masker types are indicated below each row of panels. Numbers along the abscissae specify the masker azimuths. Symbols are measurements from Hawley *et al.* (2004); error bars show one standard error. Predictions for the STEC model (solid lines), and the SSEC model (dashed lines) are shown with the SII criterion (values in upper right corners) matched to the most-separated conditions.

factors related to fluctuations in SNR, changes in the optimal equalization parameters over time, and confusions between target and masker sounds are not present in this co-located case. In the second row (the ModSSN case), almost all data are again well fit by both models with the exception that the steady-state EC predicts poorer than observed performance when maskers are on both sides of the target, i.e., the $(-30,90)$ and the $(-30,60,90)$ conditions. Note that co-located cases are also close to predicted values in the SSN and ModSSN data, consistent with the noise maskers being easily distinguished from the speech target, even with no spatial separation. For the speech-masker conditions, in the third row, results and predictions again show the benefit of time-dependent EC processing in the two widely separated-masker conditions: $(-30,90)$ and $(-30,60,90)$ have significantly lower predicted thresholds when short-time equalization adjustments are made, and the short-time model provides a better fit to these cases (viz., relative to the

cases when multiple maskers are at the same location). Also when the speech maskers and target are co-located (all straight ahead), an increase in masking (higher threshold) is observed for the multiple-masker conditions. This additional masking, which is not predicted by the spatial dependence of either model and which is a distinct difference from the noise-masker conditions, is consistent with the presence of greater IM when the maskers and targets are not distinguished by spatial position. Finally, for the reversed-speech-masker cases in the bottom row of panels, short-time processing again predicts lower thresholds for the split-masker conditions, but results are better predicted by the SSEC model. Although these differences are only a few decibels, which is within the variability as indicated by the error bars, these data are clearly better fit by the steady-state model. Overall, we interpret these data as showing the potential advantages of short-time processing for cases with maskers that have independently fluctuating levels and that are widely separated relative to the spatial selectivity of the binaural processor. Our hypothesis about the lack of benefit observed in the reversed masker case is that the temporal envelopes of the reversed maskers are different, such that (1) there is more forward masking, as measured by [Rhebergen,](#)

[Versfeld, and Dreschler \(2005\)](#) and (2) the shapes and unpredictability of the envelopes may reduce the listeners' abilities to switch the processing to cancel the stronger masker quickly enough to find a benefit. These hypotheses are discussed further in Sec. V.

An explicit demonstration of the functioning of the STEC model can be seen in the distributions of the values of the interaural parameters chosen to equalize for each processing window in the operation of the EC mechanism. These distributions for the speech-masker cases are illustrated in Fig. 3 for a low-frequency (500-Hz) channel and a high-frequency (4-kHz) channel. For the single masker cases shown in the top sets of panels, the internal interaural parameters are matched to the constant direction of the masker, and the distributions are very narrow around the values that compensate for the masker direction, values that depend on frequency. This behavior is illustrated by the distributions for a single speech masker in the upper two rows of distributions in Fig. 3, where distributions for τ_o ("Tau") are shown in the left set of panels (a) and those for α_o ("Alpha") are shown in the right set of panels (b). These distributions are all narrow, as expected.² The more informative distributions are seen in the two-speech-masker cases in the middle sets

FIG. 3. Distributions of equalization parameters τ_o (Tau(ms)) and α_o (Alpha(dB)) for 500-Hz and 4-kHz frequency bands for different azimuths of one, two, or three speech maskers (upper, middle, and lower sets, respectively). Sets of azimuths enclosed by parentheses at the top of each set of panels show masker directions. Ordinate values specify the number of occurrences (count) of the equalization parameters in the roughly 2.5-s stimulus.

of panels in the same figure. When maskers are on opposite sides, the equalization parameters would be expected to match each masker direction during the time intervals when that masker dominates. Here it is clear that the distributions for the case with maskers on each side of the target, i.e., the $(-30,60)$ case, are bimodal. The other two-masker cases are essentially unimodal, including the $(60,90)$ case where there are small differences in the cancellation delays for the two maskers, presumably corresponding to the slightly broader distribution of values for $(60,90)$ compared to $(90,90)$. For the three-speech-masker cases, shown in the third row of distributions in Fig. 3, the distributions are again bimodal in some cases, as expected. Specifically, the distributions are again spread over significant ranges when the masker directions include both sides of the target, notably the $(-30,60,90)$ cases. Finally, we note that the distributions for other masker types with significant speech-rate modulations are similar; they are not shown here, but are available upon request.

Overall, comparisons with the data from the Hawley experiments show that the STEC model provides a better description than the SSEC model for separated, independent speech maskers and for separated speech-modulated-noise maskers with independent envelopes. There are no significant differences between the models' predictions for the single-masker cases nor for unmodulated noise cases nor for multiple maskers that are spatially close together, all as expected. For the reversed-speech cases, the short-time model predicts relative improvements for the separated maskers that are not observed. This case is being further explored; for example, it is possible that the different envelope profiles (faster rise and slower decay for the forward-speech case relative to the reversed speech case) may reduce the advantages or make it difficult to adjust the equalization parameters optimally. Finally, it is notable that the predictions of both models (SSEC and STEC) are consistent with much of the data, and the steady-state version is much simpler to apply; however, there are cases for which short-time processing is needed to capture some aspects of the data and the STEC model is one way to do this.

B. Application to data of Marrone *et al.*

The model was also applied to the Marrone experiment (Marrone *et al.*, 2008), which used a different speech corpus than the Hawley experiment and also used a different set of masker directions. It was felt that applications to both data sets, as in Wan *et al.* (2010), would provide a stronger test of the model. More specifically, in the Marrone experiment (Marrone *et al.*, 2008), the degree of separation was varied in cases with maskers on both sides of the target and both targets and maskers were taken from the CRM corpus (Bolia *et al.*, 2000). In contrast to the diverse sentences of the IEEE corpus, the CRM corpus constrains each sentence to be of the form "Ready <name>, go to <color> <number> now." There was only a small number of choices for each variable item, and target and maskers are roughly synchronized. SRTs were measured for the speech-masker conditions and for a few reversed-speech-masker conditions. Although

measurements were taken in low- and high-reverberation environments, only the low-reverberation data are used here. Three different female talkers were used, one for the target voice and two others for the maskers. In contrast to the Hawley experiment discussed in the preceding text, there was no information provided to subjects about the content of the individual masker waveforms on each trial. The target voice was identified by the phrase "ready, baron," that started every target sentence to which the listener replied with a color and a number. Given these differences in the structures of the two studies, one might expect to see different amounts of informational masking (IM) and different roles of the short-time processing in the STEC model.

The application of the STEC model to these data again followed the procedures described in Wan *et al.* (2010) for the SSEC model except that short-time processing is used as described in Sec. II. The output of the model is evaluated using exactly the same method as that used in the previous section, employing the SII. The predicted and empirical data for these experiments with two speech maskers are shown in Fig. 4 along with the predictions made by the SSEC and the STEC models. In these predictions, the SII criterion is matched to the condition with the maskers at $\pm 90^\circ$. As discussed previously, the primary test of the current model is its ability to describe the dependence on the masker directions for the non-co-located conditions. (The co-located condition maximizes confusions between target and masker speech, an aspect of IM, which is not included in either model.)

As seen clearly in Fig. 4, the STEC model makes improved predictions (solid lines) for the spatial dependence of threshold compared to the SSEC model. In particular, the observed increase in threshold as the maskers move closer to the target is very close to that predicted by the STEC model (and not by the SSEC model). The basis within the STEC model for this behavior is discussed in the next paragraph. The only significant discrepancy between the data and the STEC predictions is in the relatively poor performance in the co-located $(0,0)$ condition relative to the separated-masker conditions. Consistent with the observations and

FIG. 4. Simulated and measured binaural SRTs for speech masker cases in a pseudo-anechoic environment. Symbols are the measurements from Marrone *et al.* (2008), and the error bar is one standard error. Predicted values are connected by solid lines (STEC model) and dashed lines (SSEC model). The number in the upper right corner of the plot gives the value of the SII criterion, which was chosen to match STEC prediction and data in the $(-90^\circ, 90^\circ)$ condition.

arguments of [Freyman et al. \(2007\)](#), the high threshold for the co-located condition is expected as a consequence of IM in addition to the energetic masking that is included in the STEC model predictions. As noted already in the preceding text, the STEC model has no mechanism to account for the IM.

When the predictions of the STEC model and the SSEC models are compared in Fig. 4, the spatial dependence of the predictions are similar (i.e., thresholds increase as the masker separation decreases), but the STEC model threshold increases faster as the separation between the maskers is reduced; this is more compatible with the data. This increase is expected within the STEC model because the benefit of short-time processing depends on the ability to cancel the stronger masker within short time intervals, and, as noted in the preceding text, there is minimal advantage when the maskers are close together. Thus the relative increase in the SRT that is seen with decreasing separation in the STEC model is actually a loss of the advantage of the STEC processing for larger separations.

Further insight into the discrepancy between the co-located threshold as observed and as predicted when the model parameter is chosen to fit the spatially separated thresholds can be gained by comparing the speech case with the reversed-speech case in the same study ([Marrone et al., 2008](#)). The data for the reversed-speech case were limited to the co-located (0,0) condition and the fully separated ($\pm 90^\circ$) condition and are not re-plotted here. [The spatial releases obtained by comparing the (-90° , $+90^\circ$) thresholds to the (0,0) thresholds for speech and reversed speech cases are shown in Fig. 10 of [Wan et al. \(2010\)](#).] The reversed-speech data (cf. Fig. 10, [Wan et al., 2010](#), or Fig. 3, [Marrone et al., 2008](#)) show a large decrease in the co-located threshold (about 12 dB) and relatively small reductions in the thresholds with the symmetrically displaced maskers (about 4 dB). The reversed-speech data are consistent with shape of the predictions of the STEC model, as given in Fig. 4, assuming no IM. A large reduction in IM with reversed-speech maskers has been observed in multiple other experiments, e.g., [Freyman et al. \(1999\)](#), [Rhebergen et al. \(2005\)](#), [Kidd et al. \(2010\)](#), and [Best et al. \(2012\)](#). The differences in the co-located thresholds and predictions for the Hawley speech experiment are smaller, presumably because there is less IM in that experiment, when the speech maskers sentences were provided to the listeners and IEEE sentences were used as maskers instead of the CRM sentences.

V. DISCUSSION

This paper describes a relatively straightforward generalization of the EC model ([Durlach, 1963, 1972](#)) to include time-dependent selection of equalization parameters. This STEC model used 20-ms windows and an idealized parameter-selection criterion to minimize the post-cancellation masker level in each temporal window. The model was applied to the data of [Hawley et al. \(2004\)](#) and of [Marrone et al. \(2008\)](#) and predicted lower thresholds for cases with multiple well-separated maskers when short-term masker levels were varying independently for each masker.

Overall the dependence of the thresholds on the locations of the maskers was better described by the STEC model than by the [Wan et al. \(2010\)](#) model in which the equalization parameters were chosen once for the full duration of the masker configuration. More specifically, predictions were fit better by the STEC model for the speech-masker cases when maskers were speech for both the Hawley and Marrone experiments. Performance with the STEC model was a poorer fit to the data when the maskers were reversed speech in the Hawley experiment, but STEC gave a better fit for the Marrone data with reversed-speech maskers. For speech-modulated noise in the Hawley experiment, the STEC model improved the fit for separated maskers, and in the unmodulated noise cases, there was no difference in the STEC versus SSEC predictions as expected. We conclude that the actual processing uses time-dependent processing, although the circumstances when this can be exploited are not yet understood. It is troubling for our hypothesis that the reversed-speech maskers do not show the STEC predicted behavior in the Hawley experiment although the differences in the shapes of forward and backward speech envelopes may be relevant. We plan to continue to pursue these questions with more refined (and therefore more complicated) models, including, for example, forward masking as explored by [Rhebergen et al. \(2006\)](#). In addition to the forward-masking effects, we also plan to explore the possibility that listeners' experience with speech may allow more efficient adjustment of equalization parameters within the EC model context.

The conclusion that time-frequency slices can be processed selectively is consistent with conclusions reached in the recent paper of [Brungart and Iyer \(2012\)](#). They compared performance for symmetric maskers with normal binaural listening to performance for listeners hearing a monaural signal that was constructed to include, for each time-frequency unit, the signal from the ear with the better SNR selected separately for each combination of time-interval and frequency-band. They found roughly equal performance in these two conditions, a result that is consistent with the ability to change the processing in individual time-frequency intervals. Their conclusion that time slices can be separately processed is consistent with our short-time processing. At a more refined level, a comparison between EC processing for each time slice, as assumed in our model, relative to better-ear selectivity, as presented to listeners in the Brungart and Iyer experiment, would be interesting to test explicitly and with a broader set of experimental conditions.

The distributions of the equalization parameters of the STEC model, i.e., the internal interaural time and intensity parameters, show bimodal distributions when the interferers are widely separated. This is consistent with what is expected when the cancellation process can cancel the masker in each time interval. Most noticeably, this occurs for conditions with maskers on each side of the target. For maskers close together, there is little advantage in adjusting the internal parameters and the changes in the distributions are insignificant as are any differences in the predicted thresholds for the two models. The degree of improvement with time-dependent processing relative to the separation of the maskers may depend on the level of internal noise. These

parameters were kept fixed at the values originally fit by Durlach (1963) for the current work.

Like most binaural processing models, the EC model predicts the dependence of thresholds on the interaural parameters according to the interactions of waveforms within frequency channels. This processing and these interactions generally ignore the nature of the waveforms so that the processing and the predicted advantages of various combinations of target and maskers are independent of whether the waveforms are noise, speech, or other sounds. These models ignore the confusions that are observed experimentally when the target and maskers are both speech signals and the listener reports a word from the masker as being part of the target. This “informational masking” or IM component (Brungart *et al.*, 2001; Durlach *et al.*, 2003) has not been explicitly included in binaural models and is not included in the models considered in this study, even though there is no doubt that IM is present in the experiments modeled here. In particular, the details of the Hawley and Marrone studies suggest different amounts of IM and more detailed comparison of the two may provide insight into how existing models succeed and fail and how they might be modified.

The effects of IM are seen prominently in conditions for which the maskers and target are co-located. It has been observed (e.g., Freyman *et al.*, 2007) that there are substantial confusions in these conditions and that the spatial separation of maskers from target resolves much of this confusion because the target is the source perceived as straight ahead and the maskers are perceived to the sides. As noted in the preceding text, the large differences in the Marrone experiment with reversed-speech interferers compared with speech interferers is consistent with the importance of IM in this experiment. The IM would be expected to be stronger in the Marrone experiment for multiple reasons. Perhaps most obviously, in the Hawley experiment, listeners were provided with the text of the masking speech waveforms during each threshold measurement of the speech-masker experiment, which would tend to reduce the IM in these cases. Also the Marrone experiment used the more restricted CRM corpus for both target and maskers, as opposed to the IEEE sentences in the Hawley experiment, which may also change the amount of IM. Of course, other factors such as talker differences would also have effects.

Finally, we note that, even when the SII criterion is matched to the thresholds for conditions when the maskers are most lateral (generally at $\pm 90^\circ$), there are differences in the value of the criterion for different types and numbers of maskers. For example, conditions with multiple maskers at these most lateral directions are matched with values of the SII criterion varying from 0.163 to 0.377 for different numbers and types of speech maskers. This range of SII corresponds to about 6 dB of SNR at threshold. These differences suggest a finer analysis of glimpses of the target speech might be important with explicit consideration of the ability of the listener to fill in the missing speech information while “listening in the gaps.” In general, there are multiple factors involved in the differences between masker types, and our modeling needs to be expanded to be able to understand the differences in the criterion values for the different conditions.

ACKNOWLEDGMENTS

The work reported in this document is a part of the doctoral dissertation of R.W. The authors appreciate useful comments from Gerald Kidd on an earlier version of this manuscript and substantial assistance from Andrew Brughera with the figures. The work reported here and the preparation of this document were supported by the National Institute of Deafness and Communication Disorders, Grant No. R01 DC00100.

¹As specified in the ANSI standard (ANSI, 1997) and as is usual in applications, the criterion for the SII is expected to depend on the details of the experiments, including procedures, speech material, and the type of maskers: “The intelligibility of received speech depends not only on the SII but also on other factors, such as the nature of the message being transmitted (e.g., its syntactic, semantic, linguistic, and contextual constraints) and the proficiency of the listeners and talkers.” In the current experiments, there are differences in the nature of the target speech (Marrone used CRM corpus and Hawley used IEEE sentences) as well as differences in the maskers (including speech maskers that are easily confusable with the target and modulated speech-shaped noises that are not easily confusable with the target).

²Note that the temporal jitter parameters δ_L and δ_R have distributions that are specified to be independent of frequency in the EC model. The standard deviations of the Gaussian distributions are fixed, independent of frequency, and the deviations become significant with respect to a period at 4 kHz so that the compensating values of the equalizing delay Tau have a wider distribution at 4 kHz even for a fixed masker direction.

- ANSI (1969). S3.5, American National Standard Methods for the Calculation of the Articulation Index (Acoustical Society of America, New York).
- ANSI (1997). S3.5, American National Standard Methods for the Calculation of the Speech Intelligibility Index (Acoustical Society of America, New York).
- Best, V., Marrone, N., Mason, C. R., and Kidd, G., Jr. (2012). “The influence of non-spatial factors on measures of spatial release from masking,” *J. Acoust. Soc. Am.* **131**, 3103–3110.
- Beutelmann, R., and Brand, T. (2006). “Prediction of speech intelligibility in spatial noise and reverberation for normal-hearing and hearing-impaired listeners,” *J. Acoust. Soc. Am.* **120**, 331–342.
- Beutelmann, R., Brand, T., and Kollmeier, B. (2009). “Prediction of speech intelligibility with frequency-dependent interaural phase differences,” *J. Acoust. Soc. Am.* **126**, 1359–1368.
- Beutelmann, R., Brand, T., and Kollmeier, B. (2010). “Revision, extension, and evaluation of a binaural speech intelligibility model,” *J. Acoust. Soc. Am.* **127**, 2479–2497.
- Bolia, R. S., Nelson, W. T., Ericson, M. A., and Simpson, B. D. (2000). “A speech corpus for multitalker communications research,” *J. Acoust. Soc. Am.* **107**, 1065–1066.
- Brungart, D., and Iyer, N. (2012). “Better-ear glimpsing efficiency with symmetrically placed interfering talkers,” *J. Acoust. Soc. Am.* **132**, 2545–2556.
- Brungart, D., Simpson, B. D., Ericson, M. A., and Scott, K. R. (2001). “Informational and energetic masking effects in the perception of multiple simultaneous talkers,” *J. Acoust. Soc. Am.* **110**, 2527–2538.
- Culling, J., Hawley, M., and Litovsky, R. (2004). “The role of head-induced interaural time and level differences in the speech reception threshold for multiple interfering sound sources,” *J. Acoust. Soc. Am.* **116**, 1057–1065.
- Drullman, R., Festen, J. M., and Plomp, R. (1994). “Effect of temporal envelope smearing on speech recognition,” *J. Acoust. Soc. Am.* **95**, 1053–1064.
- Durlach, N. I. (1963). “Equalization and cancellation theory of binaural masking-level differences,” *J. Acoust. Soc. Am.* **35**, 1206–1218.
- Durlach, N. I. (1972). “Binaural signal detection: Equalization and cancellation theory,” in *Foundations of Modern Auditory Theory*, edited by J. V. Tobias (Academic, New York), Vol. 2, Chap. 10.
- Durlach, N. I., Mason, C. R., Shinn-Cunningham, B. G., Arbogast, T. L., Colburn, H. S., and Kidd, G., Jr. (2003). “Note on informational masking,” *J. Acoust. Soc. Am.* **113**, 2984–2987.

- Freyman, R. L., Helfer, K. S., and Balakrishnan, U. (2007). "The role of perceived spatial separation in the unmasking of speech," *J. Acoust. Soc. Am.* **121**, 1040–1046.
- Freyman, R. L., Helfer, K. S., McCall, D. D., and Clifton, R. K. (1999). "A speech corpus for multitalker communications research," *J. Acoust. Soc. Am.* **106**, 3578–3588.
- Hawley, M. L., Litovsky, R. Y., and Culling, J. F. (2004). "The benefit of binaural hearing in a cocktail party: Effect of location and type of interferer," *J. Acoust. Soc. Am.* **115**, 833–843.
- Kidd, G., Mason, C. R., Richards, V. M., Gallun, F. J., and Durlach, N. I. (2008). "Informational masking," in *Auditory Perception of Sound Sources*, edited by William A. Yost, Arthur N. Popper, and Richard R. Fay (Springer, New York), pp. 143–189.
- Kidd, G., Jr., Mason, C. R., Best, V., and Marrone, N. L. (2010). "Stimulus factors influencing spatial release from speech on speech masking," *J. Acoust. Soc. Am.* **128**, 1965–1978.
- Kryter, K. D. (1962). "Validation of the Articulation Index," *J. Acoust. Soc. Am.* **34**, 1698–1702.
- Lavandier, M., and Culling, J. F. (2010). "Prediction of binaural speech intelligibility against noise in rooms," *J. Acoust. Soc. Am.* **127**, 387–399.
- Lavandier, M., Jelfs, S., Culling, J. F., Watkins, A. J., Raimond, A. P., and Makin, S. J. (2012). "Binaural prediction of speech intelligibility in reverberant rooms with multiple noise sources," *J. Acoust. Soc. Am.* **131**, 218–231.
- Levitt, H., and Rabiner, L. R. (1967a). "Binaural release from masking for speech and gain in intelligibility," *J. Acoust. Soc. Am.* **42**, 601–608.
- Levitt, H., and Rabiner, L. R. (1967b). "Predicting binaural gain in intelligibility and release from masking for speech," *J. Acoust. Soc. Am.* **42**, 820–829.
- Marrone, N., Mason, C. R., and Kidd, G. (2008). "Tuning in the spatial dimension: Evidence from a masked speech identification task," *J. Acoust. Soc. Am.* **124**, 1146–1158.
- Meyer, R. M., and Brand, T. (2013). "Comparison of different short-term speech intelligibility index procedures in fluctuating noise for listeners with normal and impaired hearing," *Acta Acust.* **99**, 442–456.
- 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. (2005). "Release from informational masking by time reversal of native and non-native interfering speech," *J. Acoust. Soc. Am.* **118**, 1274–1277.
- 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.
- Rothausen, E. H., Chapman, W. D., Guttman, N., Nordby, K. S., Silbiger, H. R., Urbanek, G. E., and Weinstock, M. (1969). "I.E.E.E. recommended practice for speech quality measurements," *IEEE Trans. Audio Electroacoust.* **17**, 227–246.
- Wan, R. (2011). "Investigation and evaluation of binaural models applied to speech perception in complex environments," Ph.D. dissertation, Boston University, Boston.
- Wan, R., Durlach, N. I., and Colburn, H. S. (2010). "Application of an extended equalization-cancellation model to speech intelligibility with spatially distributed maskers," *J. Acoust. Soc. Am.* **128**, 3678–3690.
- Zurek, P. M. (1992). "Binaural advantages and directional effects in speech intelligibility," in *Acoustical Factors Affecting Hearing Aid Performance*, 2nd ed., edited by G. A. Studebaker and I. Hochberg (Allyn and Bacon, Boston), 255–276.