*Original Article*

Trends in Hearing
2016, Vol. 20: 1–12

DOI: 10.1177/2331216516669919
tia.sagepub.com

# A Binaural Grouping Model for Predicting Speech Intelligibility in Multitalker Environments

**Jing Mi[1] and H. Steven Colburn[1]**

**Abstract**

Spatially separating speech maskers from target speech often leads to a large intelligibility improvement. Modeling this phenomenon has long been of interest to binaural-hearing researchers for uncovering brain mechanisms and for improving signal-processing algorithms in hearing-assistive devices. Much of the previous binaural modeling work focused on the *unmasking* enabled by binaural cues at the periphery, and little quantitative modeling has been directed toward the grouping or source-separation benefits of binaural processing. In this article, we propose a binaural model that focuses on grouping, specifically on the selection of time-frequency units that are dominated by signals from the direction of the target. The proposed model uses Equalization-Cancellation (EC) processing with a binary decision rule to estimate a time-frequency binary mask. EC processing is carried out to cancel the target signal and the energy change between the EC input and output is used as a feature that reflects target dominance in each time-frequency unit. The processing in the proposed model requires little computational resources and is straightforward to implement. In combination with the Coherence-based Speech Intelligibility Index, the model is applied to predict the speech intelligibility data measured by Marrone et al. The predicted speech reception threshold matches the pattern of the measured data well, even though the predicted intelligibility improvements relative to the colocated condition are larger than some of the measured data, which may reflect the lack of internal noise in this initial version of the model.

**Keywords**

binaural hearing, EC model, grouping

Date received: 21 June 2016; accepted: 28 August 2016

Many empirical studies have shown that speech intelligibility can be improved by spatially separating the maskers from the target (Arbogast, Mason, & Kidd, 2005; Freyman, Helfer, McCall, & Clifton, 1999; Kidd, Mason, & Gallun, 2005). This intelligibility improvement is called spatial release from masking or SRM (Arbogast et al., 2005). As decreased hearing ability in multisource environments is a frequent complaint by hearing-impaired listeners, particularly when using hearing aids or cochlear implants, modeling SRM in complex sound environments is of interest to hearing researchers. Better understanding of the psychophysical issues involved in this process could lead to a better understanding of neural mechanisms involved and might stimulate the design of better sound-processing algorithms for hearing-assistive devices.

Previous binaural models of SRM (Beutelmann & Brand, 2006; Beutelmann, Brand, & Kollmeier, 2010; Lavandier & Culling, 2010; Levitt & Rabiner, 1967; Wan, Durlach, & Colburn, 2010, 2014; Zurek, 1992) focused on the *binaural unmasking* aspects of processing, where unmasking refers to the enhancement of signal-to-noise ratio (SNR) by suppression of masker components with binaural processing. The Equalization-Cancellation (EC) model of Durlach (1963) is the most widely used model of binaural unmasking and has been used in several SRM modeling efforts. Levitt and Rabiner (1967) first adapted the EC model to predict improvement in speech intelligibility in broadband noise by applying the EC model separately in each frequency band and calculating the

[1]Boston University, Boston, MA, USA

**Corresponding author:**
Jing Mi, Boston University, 44 Cummington Mall, Room 427, Boston, MA 02215-1300, USA.
Email: jingmi@bu.edu

corresponding SNR. These SNRs were then combined across frequency with the standard band-importance function (American National Standards Institute [ANSI], 1997) to calculate the Speech Intelligibility Index (SII). This band-by-band unmasking approach with the SII frequency combination was applied by Zurek (1992) to describe the dependence of speech intelligibility on the direction of unmodulated speech-shaped noise. This EC model was further developed to also apply to speech masked by other masker types including modulated noise, multiple speech, and reversed speech maskers (Beutelmann & Brand, 2006; Lavandier & Culling, 2010; Wan et al., 2010, 2014). These models' predictions have shown general agreement with experimental data when nonspeech maskers are involved. When the maskers are speech, however, the predicted SRM is much smaller than measured data and shows little dependence on spatial separation relative to the data (Wan et al., 2010), which is possibly due to not considering the amplitude fluctuations of the speech maskers. In particular, with two maskers in different positions, a dominant masker can be cancelled (approximately), but the dominant masker varies over time. To exploit the amplitude fluctuations of speech signals, researchers developed short-time EC (STEC) models which perform EC calculations in short time frames (Beutelmann et al., 2010; Wan et al., 2014). For the speech-on-speech experiment, the dependence of speech reception threshold (SRT) on spatial separation can be successfully predicted by the STEC model, with the exception of the collocated condition; thus, the model fails to predict the large amount of SRM that is observed when maskers change from being collocated with to being separated from the target. With a spatial separation as small as 15°, the measured SRM can be as large as 10 dB (Marrone, Mason, & Kidd, 2008), while predicted SRM is no larger than 1 dB (Wan et al., 2014).

This unexplained SRM for the speech maskers case has been attributed to the existence of informational masking (Wan et al., 2014). Informational masking (Kidd, Mason, Richards, Gallun, & Durlach, 2008) is a broad concept that generally refers to the confusability between target and maskers. To avoid confusion of the target with maskers in a multitalker mixture, listeners use cues like pitch or spatial location to distinguish target speech components from masker speech components and to group target elements together across time and frequency. This process is referred to by Bregman (1990) as grouping. Unlike the abundant works on the psychoacoustic modeling of binaural unmasking, relatively little work has been done on the psychoacoustic modeling of binaural grouping (cf., review by Bronkhorst, 2015). Most literature on grouping using binaural cues is found in the Computational Auditory Scene Analysis (CASA) domain (Jiang, Wang, Liu, & Feng, 2014; Lyon, 1983; Mandel, Weiss, & Ellis, 2010; Roman, Srinivasan, & Wang, 2006; Roman, Wang, & Brown, 2003); however, those studies focused on engineering solutions for source separation rather than proposing a physiologically plausible binaural grouping model. Thus, those models are seldom applied to predict data from psychoacoustical experiments.

The study reported here proposes a grouping model based on binaural cues and combines the grouping model with the coherence-based SII (CSII; Kates & Arehart, 2005) to predict SRM measured by Marrone et al. (2008). The model uses EC processing to estimate the strength of the signal from the target direction. Specifically, signals from the left and right channels are equalized with the equalization parameters chosen to match the known (or postulated) target direction. Then, the equalized signals are subtracted (cancelled) to eliminate the signal from the target direction and the size of the residual is evaluated. If a time-frequency (T-F) region is dominated by target, the cancellation is likely to be successful with a small residue. In other words, if a T-F region has much less energy after EC processing compared with the energy before EC, that T-F region is likely to be dominated by the target. The output of the binaural model consists of the combined components from these target-dominated T-F regions. The intelligibility of the resulting output is evaluated with the CSII measure. The proposed grouping model predicts a 6 to 10 dB larger SRM compared with the STEC unmasking model (Wan et al., 2014) and this prediction correlates well with human performance in Marrone et al. (2008). Another significant difference between the current model and past models is the ability to perform the binaural processing without a priori knowledge of the stimulus waveforms. In the STEC model, for example, this knowledge is used to choose the equalization parameters that could maximally cancel the maskers in each time-frequency unit. In the current model, the only a priori knowledge assumed is the direction of the source of interest and the associated head-related transfer function (HRTF).

## Description of the Binaural Grouping Model

The grouping model proposed here is fundamentally a mechanism using binaural cues to select time-frequency intervals of the input waveforms that are dominated by the target. Target direction and the associated interaural time/level differences (ITD/ILD) of target direction are assumed to be known a priori (i.e., the HRTF is assumed known for the target direction). A block diagram of the proposed model is shown in Figure 1. The model consists of the following four stages: (a) a linear filter model of

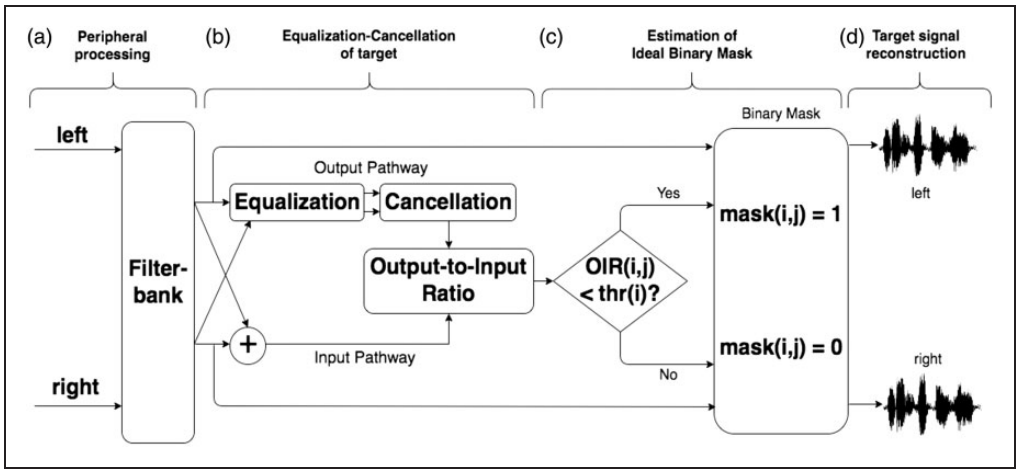


**Figure 1.** Diagram of the proposed binaural grouping model. The input to the model is binaural multitalker mixtures. The model consists of four stages: (a) peripheral processing, (b) equalization-cancellation of target, (c) estimation of ideal binary mask, and (d) target signal reconstruction.

the auditory periphery; (b) an EC process applied to time slices of the filtered signals to show the relative strength of the target signal in each time-frequency unit; (c) estimation of a binary mask based on reduction of the signals by the EC processing; and (d) reconstruction of the binaural target signal by applying the binary mask to the input signals.

The input of the model is a binaural mixture that is generated by two or more speech sources at different locations. If the binaural mixture is from collocated speech sources, the model will not be able to use binaural cues to group target elements together. So, for the collocated case, the binaural signals will simply pass through the linear filter stage and get reconstructed with an all-unity binary mask, essentially skipping the second and the third stages. The collocated stimuli still need to go through peripheral processing because the auditory filterbank in the peripheral processing stage could introduce temporal distortions to the original signal. So if a speech intelligibility metric that is sensitive to temporal alignment is used for evaluating intelligibility of the model's output, all the stimuli have to go through peripheral processing stage to avoid temporal misalignment.

The peripheral processing stage is simulated with a bank of bandpass filters. The filterbank here includes 32 gammatone filters implemented in the MATLAB Auditory Toolbox (Slaney, 1998). The center frequencies of the filters range from 80 Hz to 6 kHz and are spaced uniformly on a logarithmic scale.

The goal of the second stage is to cancel out the target signal from the mixture. In each frequency channel, the left-filtered waveform and right-filtered waveform are equalized and cancelled (EC) with the ITD and interaural amplitude ratio corresponding to the target direction chosen as the EC parameters. Equation (1) below summarizes the EC process:

$$Y_i(t) = \frac{1}{\sqrt{\alpha_i}} L_i\left(t + \frac{\tau_i}{2}\right) - \sqrt{\alpha_i} R_i\left(t - \frac{\tau_i}{2}\right) \quad (1)$$

In this equation, $L_i(t)$ and $R_i(t)$ represent the filtered left-ear and right-ear waveforms for the $i$th frequency channel before EC processing; the variables $\tau_i$ and $\alpha_i$ stand for the intrinsic ITD and amplitude ratio between two ears for the target direction; and $Y_i(t)$ represents the $i$th channel output after EC processing. In the equalization step, the left and right signals are time aligned and amplitude adjusted using the interaural parameters of target direction. Then, in the cancellation step, the difference between the equalized left and right waveforms is calculated, which represents the residual of the mixture after cancelling the target signal. Note that by cancelling the target signal, the masker signal might be boosted when the target signal has a large ILD; however, unlike classic use of the EC model, this EC processing is only used as a method for cancelling the target in order to estimate its relative strength compared with the maskers. The EC output is not directly used as the model's output. The internal noise that is present (Durlach, 1963) in classic EC implementation is not taken into consideration here, primarily because this model is intended to be a proof-of-concept model for grouping using binaural cues. In the future, a more refined model should include internal noise in the binaural processing. Note also that this model can be also applied as a CASA algorithm, and any internal noise would be omitted for that purpose.

The goal of the third stage is to estimate an ideal binary mask (IBM) based on the target cancellation result described previously. The IBM is an energy-based

binary mask that preserves the time-frequency (T-F) regions with positive SNR and that silences the T-F regions with negative SNR. The IBM has been shown to be a reasonable goal for source-segregation algorithms in CASA (Wang, 2005). We adopted this concept from CASA for psychoacoustic modeling. In the proposed model, each filtered signal was divided into 20-ms time slices using Hamming windowing with 50% overlap. In each time-frequency unit, the output-to-input ratio (OIR) is calculated as specified in Equation (2) (cf., Roman et al., 2006):

$$\mathrm{OIR}(i,j) = 10 \times \log_{10} \frac{\int \left| Y_{i,j}(t) \right|^2 \mathrm{d}t}{0.5 \times \left( \int \left| L_{i,j}(t) \right|^2 \mathrm{d}t + \int \left| R_{i,j}(t) \right|^2 \mathrm{d}t \right)} \quad (2)$$

In Equation (2), $Y_{i,j}(t)$ represents the target-cancelled output in the $i$th frequency channel and the $i$th time slice; and $L_{i,j}(t)$ and $R_{i,j}(t)$ represent the left- and right-filtered signals of that T-F unit before target cancellation. So, the denominator represents the average input energy of the EC stage and the numerator is the output energy of the EC stage, both computed for each T-F unit. The OIR variable is used to indicate the relative strength of the target in a T-F unit (Roman et al., 2006). Suppose a T-F unit consists of only target signal; the numerator will then be approximately 0 due to the nearly perfect cancellation and OIR will go to minus infinity (as a decibel measure). Otherwise, when sources from other directions dominate a T-F unit, cancellation of target will not effectively suppress the other sources and OIR will stay relatively large. In other words, OIR is an indicator of SNR in the original T-F unit. Because the IBM is generated by imposing a threshold on SNR for the binary decision, a decision threshold $D(f)$ is imposed on OIR to create the OIR-based mask. The decision threshold $D(f)$ is a function of frequency $f$ due to the frequency dependency of binaural cues. For the binary decision, if a T-F unit has an OIR greater than the threshold $D(f)$, it will be labeled as 1 in the estimated binary mask; otherwise, it will be labeled as 0. The setting of the threshold $D(f)$ is critical to the model. How the threshold is set in the proposed model and how human listeners could potentially set the threshold internally will be discussed in next section.

The last step is the application of the estimated binary mask to the original binaural mixture. Those T-F units that are labeled as 1 will be preserved and the T-F units that are labeled as 0 will be replaced by zero. The same binary mask will be applied to the left-ear and right-ear mixtures separately. The masked binaural signals are summed across frequency at each ear and the summed broadband binaural signals form the output of the binaural model. The signal at the better ear, namely the ear with higher SNR before processing, goes on to be evaluated by the speech intelligibility model.

## Specification of Model Parameters

To study and evaluate the proposed model, we simulate a set of binaural stimuli using HRTFs measured in anechoic conditions. In the simulation, a female target talker was placed in the front, and two female masker talkers were placed symmetrically at $\pm 60^\circ$ to the target talker. The Coordinate Response Measure corpus (Bolia, Nelson, Ericson, & Simpson, 2000) was used as speech source for all talkers; the masker talkers and the target talkers were different female voices from the same corpus. The simulated scenarios are designed to be similar to the experiment done by Marrone et al. (2008) for convenient comparison purposes.

## The Relationship Between SNR and OIR

Figure 2 shows a scatter plot of individual T-F units' SNR (SNR measured before the application of HRTFs) and OIR values for a frequency channel centered at 950 Hz with a 120-Hz bandwidth. The plotted data are computed from 10 sets of three-talker mixtures. Figure 2(a) shows the case when the target is in front. It can be seen that, in the positive SNR region, OIR has an approximately linear relationship with SNR. The more target energy in a T-F unit, the more energy will get cancelled in the EC step; thus, higher SNR leads to lower OIR. In the negative SNR region, OIR does not vary much with SNR because the proportion of cancelled target energy over total mixture energy is becoming negligible. Especially, in the very low SNR region (SNR below $-20$ dB), the OIR only deviates slightly around 10 dB and is not correlated with SNR. Figure 2(b) shows the case when the target is at $60^\circ$. It can be seen when the target is off-front, the OIR is still a good indicator of the SNR. A similar pattern between SNR and OIR has been observed for the other frequency channels. Based on these observations, it can be concluded that OIR is a good indicator of the polarity of SNR in the anechoic condition.

## Optimal Threshold Setting

As mentioned in the second section, a frequency-dependent OIR threshold needs to be chosen for the estimation of the IBM in the third stage of the processing. Setting the threshold properly is crucial for the predicted intelligibility by the model. We define the optimal threshold as the threshold that leads to the most accurate estimation of the IBM. The optimal threshold can be

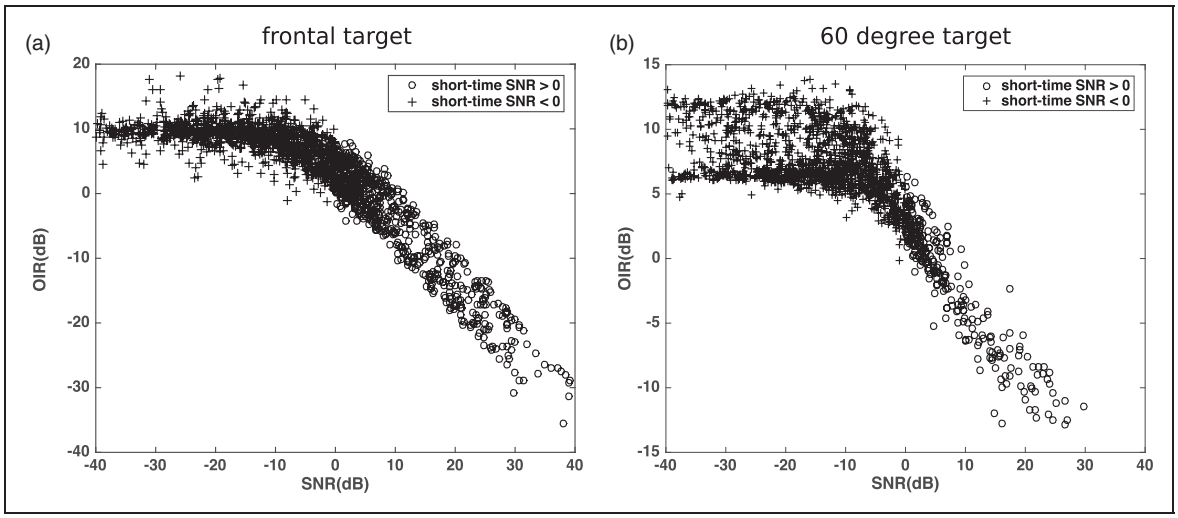


**Figure 2.** Scatterplots of the OIR to the SNR for each time slices in a frequency channel centered at 950 Hz. Circle symbol represents the T-F unit with positive SNR and cross symbol represents the T-F unit with negative SNR. The three talkers are at 60, 0, and −60 degrees, respectively. (a) The target talker is at front. (b) The target talker is at 60 degrees.
*Note*. OIR = output-to-input ratio; SNR = signal-to-noise ratio; T-F = time-frequency.

influenced by many factors, including frequency, spatial configuration, and room characteristics. First, as mentioned earlier, the relationship between SNR and OIR varies due to the frequency-dependence of binaural cues; thus, the optimal threshold needs to be set for each frequency channel individually. Second, the binaural interactions of multiple sources change when the spatial configuration change; therefore, the optimal thresholds need to be set for each spatial configuration as well. And last, room characteristics like reverberation have a huge impact on the binaural properties of the sound, so the threshold also depends on the room characteristics. The currently proposed binaural grouping model is only intended to model empirical data measured in anechoic conditions and there is no further discussion here of how optimal thresholds are affected by room acoustics. This important question will be addressed in future work.

In the model analyzed here, the optimal threshold is defined as the threshold that minimizes the difference between the estimated binary mask and the IBM. To quantify the difference, two types of error are counted: false-positive errors and false-negative errors. A false-positive error is made when a T-F unit is labeled as 1 in the estimated mask while it is labeled as 0 in the IBM. That happens when a T-F unit with negative SNR has OIR below the threshold. A false-negative error is made when a T-F unit is labeled as 0 in the estimated mask while it is labeled as 1 in the IBM. That happens when a T-F unit with positive SNR has OIR above the threshold. Based on the definitions of the two types of error, we then calculate the receiver-operating characteristic (ROC) curve by moving the OIR threshold from one direction to the other direction. With equal weight placed on false-positive error and false-negative error, the point with minimal error rates is identified on the ROC curve. We choose to give equal weights to the two types of error, but the threshold could be easily adjusted to accommodate different weights for different error types. Past studies have argued that false-positive error is more detrimental to speech intelligibility in binary mask-processed sound (Li & Loizou, 2008; Yu, Wójcicki, Loizou, Hansen, & Johnson, 2014); however, there is still controversy (Kressner & Rozell, 2015) and no definitive conclusion has been reached. So whether a different weighing of error could lead to intelligibility improvement of the model's output could be a future research direction.

Figure 3(a) shows how the error rates of different frequency channels vary, assuming optimal thresholds for the previously described simulated scenario: anechoic room, two maskers symmetrically located, and an overall SNR of −3 dB. First, it can be observed that, especially for frequencies above 500 Hz, the error rates are below a level of 0.35, which is slightly worse than the state-of-art performance of a binaural-cue-based source-segregation algorithm (Jiang et al., 2014). Second, the particularly high error rates below 500 Hz are expected from two factors. The first factor is the broader distribution of energy over the time-frequency units. This is due to the facts that speech energy is more densely distributed over the frequency bandwidths in low-frequency auditory filters compared with in high-frequency auditory filters (Lewicki, 2002; Mi & Colburn, 2015) and that the narrower bandwidths of low-frequency auditory filters lead to a wider spread of speech energy in time. Thus, energy distributions from different sources tend to overlap with

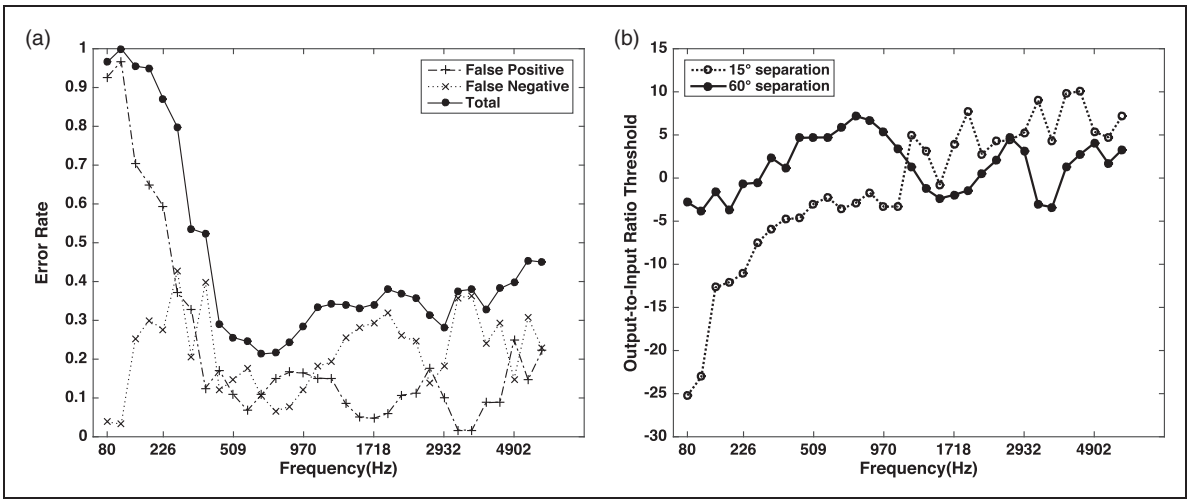


**Figure 3.** (a) Error rates of the estimated binary mask compared with the ideal binary mask. The two dashed lines represent false-positive and false-negative error rates, respectively. The solid line represents the total error rate. (b) Optimal OIR thresholds of each frequency channels for two different spatial configurations: 1. Maskers are 15° symmetrically separated from target; 2. Maskers are 60° symmetrically separated from the target.
*Note*. OIR = output-to-input ratio.

each other in time and in frequency. Therefore, below 500 Hz, there is little chance that one T-F unit is clearly dominated by one source. As a result, the grouping of T-F units will not be very useful in this region. The second factor is that the naturally occurring ITDs for spatial separations are not strongly frequency dependent, but the resulting IPDs depend on the center frequency. This implies that the phase differences for different sources are smaller at lower frequencies and would provide less cancelation in the subtraction process. Figure 3(b) shows the optimal thresholds for two spatial configurations: (a) target at front, two maskers separately at $\pm 15^{\circ}$; (b) target at front, two maskers separately at $\pm 60$ degrees. Note first that the variation of optimal threshold can be as large as 8 dB across frequency channels and thus setting the threshold to a value independent of frequency would have adverse impact on the performance. Note also that the change in spatial configuration may cause a threshold change as large as 10 dB. The difference between the thresholds for these two spatial configurations also shows frequency dependency.

The determination of the optimal threshold in the third stage must be specified and this may be difficult for an unfamiliar condition. The method described in the previous paragraphs only applies when the waveforms of the target and maskers are known for a period of time to allow the estimation of the thresholds. In real-life situations, the opportunity to get separate estimates of target and maskers is rare; however, there is evidence showing that binaural speech intelligibility in a background of maskers can be improved by preexposure to the listening environment (Brandewie & Zahorik, 2010; Kidd, Arbogast, Mason, & Gallun, 2005). Hence, a listening history-based solution is proposed for finding suboptimal thresholds. The idea is that the units with OIR values at the lower end of the OIR distribution are always desirable. Thus, a short history of OIR could be accumulated to estimate OIR distributions for each frequency channel. Then, the proportion of target-dominant units over the whole accumulation time period should be estimated. For example, in the simulated three-speaker scenario, the proportion of target-dominant T-F units is estimated to be $100 \div 3 \approx 33\%$. This estimation is made based on the assumption that the speech signals are orthogonal to each other in the time-frequency domain and the speech signals are roughly equal in long-term power. In that case, the chance of a T-F unit dominant by target source is one third. The assumption of orthogonality is overidealized but supported to some extent by analysis of speech in the time-frequency domain (Yilmaz & Rickard, 2004). With the accumulated OIR distribution and the estimated percentage, the accumulated OIR are ranked from the lowest value to the highest value. Also, an index is calculated by multiplying the estimated percentage (33% is used here) by the total number of time units. Finally, the indexed value of the ranked OIR distribution is used as the threshold for the corresponding frequency channel. Figure 4 shows how the estimated thresholds change with increasing accumulation time for the simulated scenario. As can be seen, the estimated thresholds converge to the optimal thresholds within 1 s for frequency channels for most frequency region. The strongest deviations are in the frequency channels below 500 Hz. However, in that frequency region, binaural

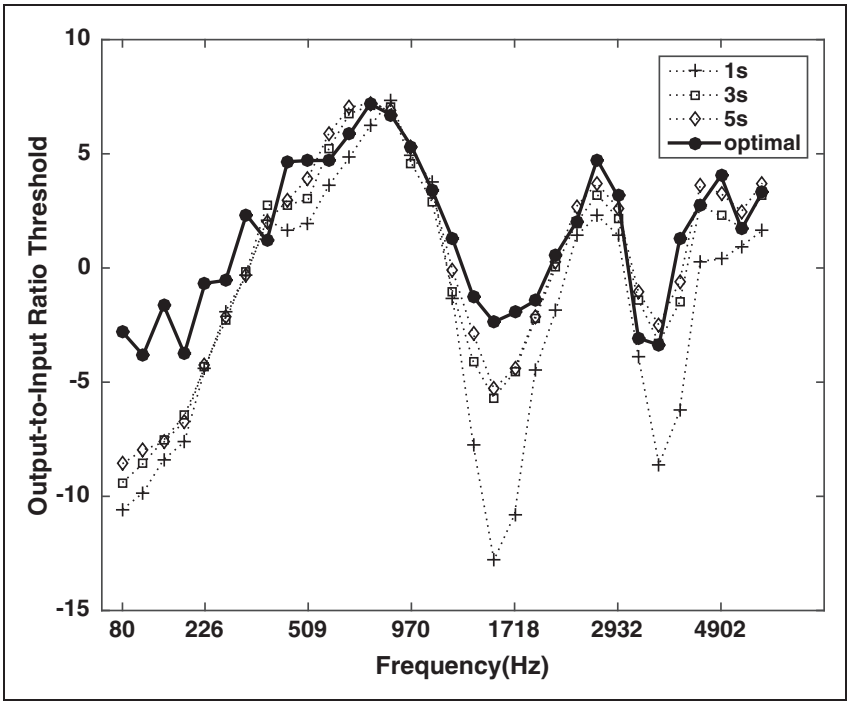


**Figure 4.** Estimated OIR thresholds at different accumulation times. The dash lines show the OIR thresholds estimated in the way proposed in the text with different amounts of accumulation time. The solid line shows the optimal thresholds that were chosen to minimize error rates compared with the ideal binary mask.
*Note.* OIR = output-to-input ratio.

grouping itself probably won't be very useful in enhancing intelligibility, so the threshold is not that important. Overall, this OIR-history-based method provides a good solution to the threshold-setting problem.

## Model Predictions Compared With Psychoacoustic Data

The model was applied to predict human performance in the speech intelligibility tasks described in Marrone et al. (2008). Specifically, they measured the binaural SRTs for a target speech sentence masked by two speech maskers that were located symmetrically with respect to the frontal target. The experiments were conducted using loudspeakers in a large sound booth (12'4" long, 13 wide, 7'6" high) with very low reverberation (6.3 dB direct-to-reverberant ratio and 0.06 s reverberation time). The target and masker sentences are from the CRM corpus with different female talkers. The CRM sentences follow the structure 'Ready < call sign > go to < color > <number > now'. The < color > and < number > choices are made randomly (without replacement for the three speech signals) for each of the presentations. Subjects are asked to report the color and number spoken by the front target, which always has *Baron* as the call sign. In this article, we simulated their experimental condition with anechoic head-related impulse responses from the CIPIC database. Marrone's data were specifically chosen here for modeling for two reasons: First, in their experiments, the maskers are symmetrically distributed with respect to the target; there is no monaural *better-ear* acoustical advantage to confound the analysis of the binaural system's role in SRM. Second, both target and maskers are spoken by female talkers in their study; therefore, pitch separation will play a small role in performance. This is good for the analysis here because pitch separation could also lead to release from informational masking, which would be another cue dimension in addition to the spatial cues.

For the predictions here, as is described earlier, the binaural mixture first went through the binaural grouping model as described earlier and the CSII (Kates & Arehart, 2005) of the model output was calculated. Only the key words (color and number) portion of output is used for CSII calculation because the other part of the sentence has little effect on the performance. To calculate the CSII, the first step is to calculate the magnitude-squared coherence of output signal to target signal in each T-F unit; then, the signal-to-distortion ratio (SDR) is derived from the coherence value; finally, SII was calculated by replacing the SNR with the SDR without changing the importance weightings of frequency bands (ANSI, 1997). The detailed calculation of CSII is described in Kates and Arehart (2005).

The CSII measure was chosen because it has been shown to be one of the best predictors for speech intelligibility in fluctuating noise conditions (Ma, Hu, & Loizou, 2009). Although SII is the most widely used speech intelligibility model (Hawley, Litovsky, & Culling, 2004; Wan et al., 2010), it is not suitable for evaluating the intelligibility of the binary mask-processed sound. For example, suppose in a frequency channel, there is only one T-F unit preserved and the SNR of that T-F unit is very high; this condition would lead to a high SNR estimation (the estimated SNR is equal to the SNR of the one preserved T-F unit) of the output of that channel even though the true long-term SNR of that frequency channel might be low. Compared with SII, CSII won't have this problem because it is based on the coherence between the output signal to the target signal in all time intervals. So if too little target is preserved, the coherence of the output signal to desired target will be low.

The CSII is based on the SII; and, like the SII, the CSII value depends on multiple factors like speech materials and masker types. Thus, a CSII criterion has to be chosen for the specific experiment setting for predicting SRT. Here, the CSII criterion was chosen such that the SRT of the collocated condition matched the empirical data. Figure 5(a) shows two SNR-CSII curves, one for the collocated condition and other for the model-processed spatially separated condition. In the measured data, the SRT for the collocated condition (labeled in Figure 5(a) as SRTr) is around 3 dB, so the CSII criterion was set as 0.4 to match the measured data of the collocated condition. Using this criterion, a prediction for the 15° separation could be made by identifying the point matching the CSII criterion on SNR-CSII curve for the 15° separation. The predicted SRT (labeled in Figure 5(a) as SRTp) is approximately −10 dB.

Figure 5(b) shows both the measured and predicted binaural SRTs for different angles of spatial separations. Each prediction is the mean and standard deviation of 25 repetitions with different target and masker stimuli. The standard deviation, represented by the shaded area, is calculated from the standard deviation of CSII by assuming a local linear relationship between CSII and SRT. In both measured and predicted data, the largest SRT change happens when maskers change from collocated with the target to 15° symmetrically separated from the target. A separation as small as 15° can lead to a SRM as large as 10 dB in the measured data and 13 dB in the predicted data. Further separations between target and maskers only generate an additional 3 dB of SRM. In general, the predicted SRT matches the pattern of measured SRT very closely (for convenience of seeing the pattern, predictions with (−90,90) as reference condition are plotted in Figure 5(b), represented by the dashed curve; as can be seen, by choosing a different CSII criterion, the predicted SRT match most of the measured SRTs except the collocated SRT). However, the predicted SRT for separations of 15° and more is always 3 to 4 dB lower than the measured data. This offset could be due to multiple reasons. The first possibility is that binaural noise is not considered in carrying out the EC processing in the current implementation (Durlach, 1963), which would reduce the benefit of using binaural cues; second, the OIR thresholds involved in the prediction are set optimally while it is unclear whether listeners could operate the selection of target components optimally; finally, the application of the binary mask in the model only allows the selected components to pass through while completely shutting off all the sound in other T-F units. It's unlikely that the brain carries out a completely binary operation, so the unwanted components might still be able to distract the listeners. Whether to replace the silenced region with appropriate noise or use a nonbinary mask could be addressed within future versions of the model.

## Discussion

This article proposes a binaural grouping model that is straightforward in implementation and that accurately predicts the pattern of SRM of the measured data in Marrone et al. (2008). In the grouping model, a binary decision is made for each individual T-F unit to determine whether it is dominated by the target-direction component or not. The key feature for the decision making is the amount of energy decrease after EC processing is carried out to cancel the target. The energy decrease is measured by the OIR. A large energy decrease is expected only when the ITD and ILD match those for the target direction and when the interaural coherence (IC) is relatively high. Lacking any of the three conditions will not result in a large energy decrease when the target is cancelled using EC processing. Thus, the OIR measure can be considered as combining the three most important binaural features: ITD, ILD, and IC. This is also one particular benefit of using E-C processing rather than cross-correlation method to calculate the measure used for binary mask estimation. Suppose cross-correlation method is used to calculate ITD and IC for each T-F unit; then, combined with the high-frequency-dominant ILD cue, the binary decision has to be made on a two- or three-dimension feature space, namely ITD + ILD or ITD + ILD + IC. Modeling the statistical distribution of the multiple binaural features is not easy, which is likely to require prelearning of the distribution (Roman et al., 2003). Therefore, by using EC processing, the binary decision can be made based only on a single variable, OIR, which is less complex than making decision on a multiple-dimensional feature space.

The proposed model is an initial effort for quantitative modeling of the grouping stage of auditory

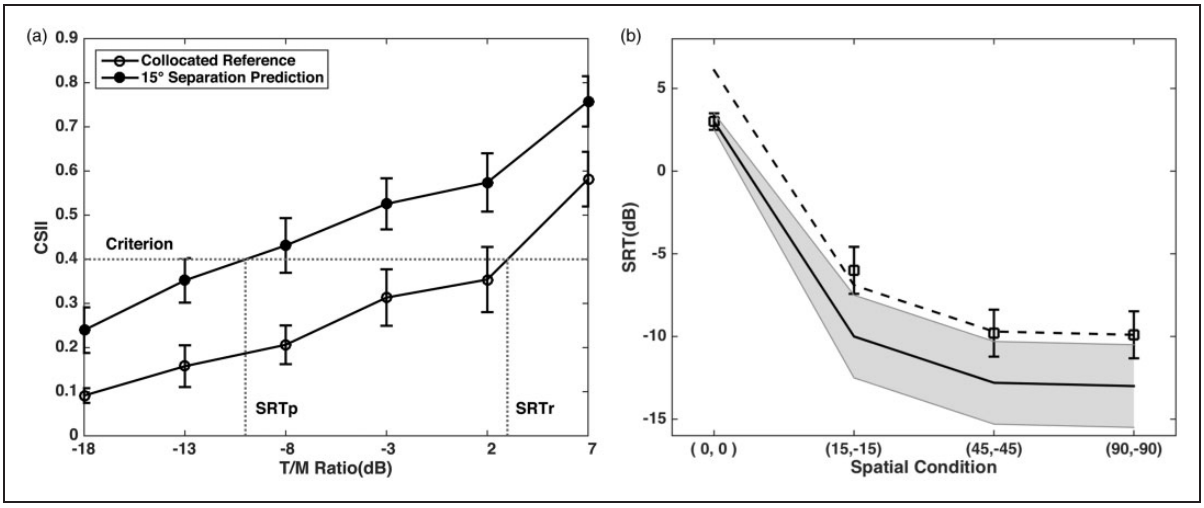


**Figure 5.** (a) CSII-SNR curves. The CSII values are calculated for the binaural grouping model's output at different Target-to-Masker (T/M) Ratio (approximately 3 dB higher than SNR in the three-talker condition). The line with open symbols is for the collocated condition and the line with solid symbols is for 15° separation condition. (b) Simulated and measured speech reception threshold. Symbols are the measured data replotted from Marrone et al. (2008). The horizontal axis labels the locations of maskers in degree. The solid curve is the model prediction using the (0,0) case as reference, and the dashed curve is the prediction using the (−90,90) case as reference. The shaded area shows one standard deviation of the predictions.
*Note.* CSII = Coherence-based Speech Intelligibility Index; SNR = signal-to-noise ratio.

processing (cf., review by Bronkhorst, 2015). Although studying binaural cue-based grouping has a long history in CASA (Jiang et al., 2014; Lyon, 1983; Mandel et al., 2010; Roman et al., 2006; Roman et al., 2003), the main goal of those studies is to develop and optimize source-separation algorithms for computing devices. Therefore, those algorithms usually require pretraining for each listening environment and are computationally intense. For example, the algorithm by Mandel et al. (2010) optimizes localization and separation of sources simultaneously and achieves state-of-the-art source-separation performance. However, it is a batch-processing algorithm and is 32 times slower than real time measured by computational power of that time. Although not biologically plausible, previous CASA studies set a good basis for quantitative modeling of human performance. The grouping model proposed here is inspired by the target-cancellation idea of Roman et al. (2006). Unlike most CASA algorithms that segregate streams from all the sources simultaneously, Roman's algorithm only segregates the target stream from background sources. This is very similar to human perception of foreground and background sound. This model proposed here makes several improvements to Roman's algorithm. In Roman et al. (2006), a few seconds of clean target is required to adaptively train a filter to cancel the target, which might be unrealistic to operate in a real environment. To circumvent this problem, the proposed model uses the HRTF of target direction and EC processing, a heuristic method to cancel the target. Another problem is that the OIR thresholds are set as a constant for all frequency channels and all conditions in Roman et al. (2006), which could harm the separation performance as shown in the Results section. The current model addresses this issue by proposing a history-based threshold-setting strategy. With these improvements, the proposed model requires minimal amounts of pretraining and computational resources, making it biologically plausible and computationally sensible.

Previous psychoacoustic binaural modeling that focused on binaural-enabled *unmasking* failed to predict the big improvement in speech intelligibility for even small degrees of separation of the speech target and maskers. Besides the failure on the modeling side, experiments have also shown that the SRM in speech-on-speech experiment does not mainly depend on an unmasking-based mechanism (Schoenmaker, Brand, & van de Par, 2016; Schoenmaker & van de Par, 2016). Instead of focusing on unmasking, the proposed model focused on grouping based on binaural cues, which represents a subsequent stage after peripheral and brainstem levels. The SRM predicted by the proposed model fits the measured data from the speech-on-speech experiment well, which suggests that binaural-cue-based grouping, rather than unmasking, contributed most to the improved speech intelligibility when speech sources are spatially separated. Many factors, like attention, can actively interact with the grouping process (Best, Ozmeral, Kopčo, & Shinn-Cunningham, 2008; Kidd, Arbogast, et al., 2005). Although attention is not explicitly modeled here, it could play a role at multiple stages of this model, including which sound direction to cancel

and determining the optimal threshold. For example, once the target direction changes, attention resources are possibly required to analyze the short history of OIR to determine the optimal threshold. To combine an attention model with the current binaural model would be an interesting direction to pursue.

Binaural cues are the only grouping cues used in the proposed model; however, the role of binaural cues in grouping is controversial. Although binaural cues could lead to large SRM, previous studies have also shown the ITD cue plays a weak role in grouping on short-time scales (~100 ms; Culling & Summerfield, 1995; Schwartz, McDermott, & Shinn-Cunningham, 2012). Culling and Summerfield (1995) showed that listeners could not identify synthetic vowel-like sounds through across-frequency grouping based on common ITD; however, the stimuli they used are unmodulated narrowband noises which are unlike natural sounds and Stern, Trahiotis, and Ripepi (2006) have shown that, by adding natural amplitude or frequency modulation in the synthetic vowels, listeners could use common ITD to group sound components together across frequency. In another study of ITD's role in grouping, Schwartz et al. (2012) used more complex synthetic stimuli that share similarities to natural sounds while lacking grouping cues like harmonicity and comodulation. Although they pointed out that ITD plays a weak role in promoting target segregation, the effect of ITD is statistically significant in their data. Moreover, their complex stimuli are not sparse in time-frequency domain like natural speech. In their second experiment, they showed that by reducing the spectral–temporal overlap between target and masker, the effectiveness of ITD in promoting source segregation was improved significantly. Therefore, some aspects of their study support the present model's assumption that for stimuli that are sparse in time-frequency domain (like speech), binaural cues can be used to distinguish target components from the sound mixture. The proposed model only uses binaural cues as a starting point for psychoacoustic modeling of grouping; other grouping cues available in natural speech, like pitch and common onsets/offsets, would also play a role in speech-on-speech experiments.

In addition to being a psychoacoustic model of binaural grouping, the model also has the potential to be a target-enhancing algorithm in hearing-assistive devices. Hearing-impaired listeners have difficultly picking out the target in complex listening situations, and efforts have been devoted to developing source-segregation algorithms to alleviate users' difficulties in noisy situations. The model proposed here is easy to implement and does not require many computational resources, making it plausible as a signal processor in hearing aids. Additionally, the parameters in the model, including target direction and OIR thresholds, could be adjusted intuitively, unlike the nonintuitive parameters in more complicated algorithms. This could provide users the benefit of adjusting the related hearing-aid settings according to their own preferences. For example, if a user prefers more direction-focused speech, the OIR thresholds could be set to a lower level to allow fewer T-F units through. In contrast, if a user prefers a more complete sound image with a slight enhancement of the target, the OIR threshold could be set to a higher level to filter out fewer T-F units. Users could also adjust the OIR thresholds according to different room acoustics, like the adaptation to room acoustics by normal-hearing listeners (Brandewie & Zahorik, 2010). Future work on adapting the model into a practical algorithm could focus on identifying the optimal threshold for noise-masker case and studying the model's behavior in reverberant condition.

This model raises many questions. On the psychoacoustical modeling side, future work is needed to determine the general utility of this approach to modeling of speech-on-speech situations. For example, binary masks could also be derived based on pitch, common onsets/offsets, and comodulation; how that could be done in a straightforward way and how to combine the binary masks derived using different cues in a probabilistic way remain open questions. Also, the validity of using binary masks for psychoacoustic modeling needs to be examined. Listeners have shown the ability to listen in gaps; however, whether the listening-in-gaps strategy operates in the brain in a similar way to applying a binary mask to the original mixture is not clear. It seems likely that the human hearing system adopts a more complex and powerful grouping mechanism than simply applying binary masks. But for initial exploration of psychoacoustic modeling of grouping, the application of binary masks is a reasonable simplification of the problem. On the engineering solution side, future work on adapting the model into a practical algorithm could focus on identifying the optimal threshold for different types of noise maskers and studying the model's behavior in reverberant conditions.

## Declaration of Conflicting Interests

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

## Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This research is supported by NIH/NIDCD Grant DC000100.

## References

American National Standards Institute. (1997). *American National Standard Methods for the Calculation of the Speech Intelligibility Index*. Melville, NY: Author.

Arbogast, T. L., Mason, C. R., & Kidd, G. (2005). The effect of spatial separation on informational masking of speech in normal-hearing and hearing-impaired listeners. *The Journal of the Acoustical Society of America*, *117*(4), 2169–2180.

Best, V., Ozmeral, E. J., Kopčo, N., & Shinn-Cunningham, B. G. (2008). Object continuity enhances selective auditory attention. *Proceedings of the National Academy of Sciences*, *105*, 13174–13178.

Beutelmann, R., & Brand, T. (2006). Prediction of speech intelligibility in spatial noise and reverberation for normal-hearing and hearing-impaired listeners. *The Journal of the Acoustical Society of America*, *120*(1), 331–342.

Beutelmann, R., Brand, T., & Kollmeier, B. (2010). Revision, extension, and evaluation of a binaural speech intelligibility model. *The Journal of the Acoustical Society of America*, *127*(4), 2479–2497.

Bolia, R. S., Nelson, W. T., Ericson, M. A., & Simpson, B. D. (2000). A speech corpus for multitalker communications research. *The Journal of the Acoustical Society of America*, *107*(2), 1065–1066.

Brandewie, E., & Zahorik, P. (2010). Prior listening in rooms improves speech intelligibility. *The Journal of the Acoustical Society of America*, *128*(1), 291–299.

Bregman, A. S. (1990). *Auditory scene analysis: The perceptual organization of sound*. Cambridge: MIT Press.

Bronkhorst, A. W. (2015). The cocktail-party problem revisited: Early processing and selection of multi-talker speech. *Attention, Perception, & Psychophysics*, *77*(5), 1465–1487.

Culling, J. F., & Summerfield, Q. (1995). Perceptual separation of concurrent speech sounds: Absence of across-frequency grouping by common interaural delay. *The Journal of the Acoustical Society of America*, *98*(2), 785–797.

Durlach, N. I. (1963). Equalization and cancellation theory of binaural masking-level differences. *The Journal of the Acoustical Society of America*, *35*(8), 1206–1218.

Freyman, R. L., Helfer, K. S., McCall, D. D., & Clifton, R. K. (1999). The role of perceived spatial separation in the unmasking of speech. *The Journal of the Acoustical Society of America*, *106*(6), 3578–3588.

Hawley, M. L., Litovsky, R. Y., & Culling, J. F. (2004). The benefit of binaural hearing in a cocktail party: Effect of location and type of interferer. *The Journal of the Acoustical Society of America*, *115*(2), 833–843.

Jiang, Y., Wang, D., Liu, R., & Feng, Z. (2014). Binaural classification for reverberant speech segregation using deep neural networks. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *22*(12), 2112–2121.

Kates, J. M., & Arehart, K. H. (2005). Coherence and the speech intelligibility index. *The Journal of the Acoustical Society of America*, *117*(4), 2224–2237.

Kidd, G., Arbogast, T. L., Mason, C. R., & Gallun, F. J. (2005). The advantage of knowing where to listen. *The Journal of the Acoustical Society of America*, *118*, 3804–3815.

Kidd, G., Mason, C. R., & Gallun, F. J. (2005). Combining energetic and informational masking for speech identification. *The Journal of the Acoustical Society of America*, *118*(2), 982–992.

Kidd, G., Mason, C. R., Richards, V. M., Gallun, F. J., & Durlach, N. I. (2008). Informational masking. In W. A. Yost, A. N. Popper, & R. R. Fay (Eds.), *Auditory perception of sound sources* (pp. 143–189). New York, NY: Springer.

Kressner, A. A., & Rozell, C. J. (2015). Structure in time-frequency binary masking errors and its impact on speech intelligibility. *The Journal of the Acoustical Society of America*, *137*(4), 2025–2035.

Lavandier, M., & Culling, J. F. (2010). Prediction of binaural speech intelligibility against noise in rooms. *The Journal of the Acoustical Society of America*, *127*(1), 387–399.

Levitt, H., & Rabiner, L. R. (1967). Predicting binaural gain in intelligibility and release from masking for speech. *The Journal of the Acoustical Society of America*, *42*(4), 820–829.

Lewicki, M. S. (2002). Efficient coding of natural sounds. *Nature Neuroscience*, *5*(4), 356–363.

Li, N., & Loizou, P. C. (2008). Factors influencing intelligibility of ideal binary-masked speech: Implications for noise reduction. *The Journal of the Acoustical Society of America*, *123*(3), 1673–1682.

Lyon, R. F. (1983). *A computational model of binaural localization and separation*. Paper presented at the Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP '83 (pp. 1148–1151), Paris, May 1982.

Ma, J., Hu, Y., & Loizou, P. C. (2009). Objective measures for predicting speech intelligibility in noisy conditions based on new band-importance functions. *The Journal of the Acoustical Society of America*, *125*(5), 3387–3405.

Mandel, M. I., Weiss, R. J., & Ellis, D. P. W. (2010). Model-based expectation-maximization source separation and localization. *Transactions on Audio Speech and Language Processing*, *18*(2), 382–394.

Marrone, N., Mason, C. R., & Kidd, G. (2008). Tuning in the spatial dimension: Evidence from a masked speech identification task. *The Journal of the Acoustical Society of America*, *124*(2), 1146–1158.

Mi, J., & Colburn, H. S. (2015). Role of high-frequency component in source segregation in multi-talkers condition. *The Journal of the Acoustical Society of America*, *137*(4), 2207.

Roman, N., Srinivasan, S., & Wang, D. (2006). Binaural segregation in multisource reverberant environments. *The Journal of the Acoustical Society of America*, *120*(6), 4040–4051.

Roman, N., Wang, D., & Brown, G. J. (2003). Speech segregation based on sound localization. *The Journal of the Acoustical Society of America*, *114*(4), 2236–2252.

Schoenmaker, E., Brand, T., & van de Par, S. (2016). The multiple contributions of interaural differences to improved speech intelligibility in multitalker scenarios. *The Journal of the Acoustical Society of America*, *139*(5), 2589–2603.

Schoenmaker, E., & van de Par, S. (2016). Intelligibility for binaural speech with discarded low-SNR speech components. In P. v. a. n. Dijk, D. Başkent, E. Gaudrain, E. d. e. Kleine, A. Wagner, & C. Lanting (Eds.), *Physiology, psychoacoustics and cognition in normal and impaired hearing* (pp. 73–81). Cham, Switzerland: Springer.

Schwartz, A., McDermott, J. H., & Shinn-Cunningham, B. (2012). Spatial cues alone produce inaccurate sound

segregation: The effect of interaural time differences. *The Journal of the Acoustical Society of America*, *132*(1), 357–368.

Slaney, M. (1998). *Auditory toolbox: A MATLAB toolbox for auditory modeling work*. Technical Report 1998–010 (Interval Research Corporation, Palo Alto, CA), pp. 1–52.

Stern, R. M., Trahiotis, C., & Ripepi, A. M. (2006). Fluctuations in amplitude and frequency enable interaural deleys to foster the identification of speech-like stimuli. In S. Greenberg, P. Divenyi, & G. Meyer (Eds.), *Dynamics of speech production and perception* (pp. 143–151). Amsterdam, The Netherlands: IOS Press.

Wan, R., Durlach, N. I., & Colburn, H. S. (2010). Application of an extended equalization-cancellation model to speech intelligibility with spatially distributed maskers. *The Journal of the Acoustical Society of America*, *128*(6), 3678–3690.

Wan, R., Durlach, N. I., & Colburn, H. S. (2014). Application of a short-time version of the Equalization-Cancellation model to speech intelligibility experiments with speech maskers. *The Journal of the Acoustical Society of America*, *136*(2), 768–776.

Wang, D. (2005). On ideal binary mask as the computational goal of auditory scene analysis. In P. Divenyi (Ed.), *Speech separation by humans and machines* (pp. 181–197). New York, NY: Springer.

Yilmaz, O., & Rickard, S. (2004). Blind separation of speech mixtures via time-frequency masking. *IEEE Transactions on Signal Processing*, *52*(7), 1830–1847.

Yu, C., Wójcicki, K. K., Loizou, P. C., Hansen, J. H. L., & Johnson, M. T. (2014). Evaluation of the importance of time-frequency contributions to speech intelligibility in noise. *The Journal of the Acoustical Society of America*, *135*(5), 3007–3016.

Zurek, P. M. (1992). Binaural advantages and directional effects in speech intelligibility. In G. A. Studebaker, & I. Hochberg (Eds.), *Acoustical factors affecting hearing aid performance*. Boston, MA: Allyn & Bacon.