Predicting masking release of lateralized speech

ALEXANDRE CHABOT-LECLERC*, EWEN N. MACDONALD, AND TORSTEN DAU

Hearing Systems Group, Department of Electrical Engineering, Technical University of Denmark, Kgs. Lyngby, Denmark

Lőcsei et al. (2015) [Speech in Noise Workshop, Copenhagen, 46] measured speech reception thresholds (SRTs) in anechoic conditions where the target speech and the maskers were lateralized using interaural time delays. The maskers were speech-shaped noise (SSN) and reversed babble with 2, 4, or 8 talkers. For a given interferer type, the number of maskers presented on the target's side was varied, such that none, some, or all maskers were presented on the same side as the target. In general, SRTs did not vary significantly when at least one masker was presented on the same side as the target. The largest masking release (MR) was observed when all maskers were on the opposite side of the target. The data in the conditions containing only energetic masking and modulation masking could be accounted for using a binaural extension of the speech-based envelope power spectrum model [sEPSM; Jørgensen *et al.*, 2013, J. Acoust. Soc. Am. 130], which uses a short-term equalization-cancellation process to model binaural unmasking. In the conditions where informational masking (IM) was involved, the predicted SRTs were lower than the measured values because the model is blind to confusions experienced by the listeners. Additional simulations suggest that, in these conditions, it would be possible to estimate the confusions, and thus the amount of IM, based on the similarity of the target and masker representations in the envelope power domain.

INTRODUCTION

Listeners benefit from listening with two ears compared to a single ear in complex listening situations. This binaural benefit is usually explained in terms of "better-ear" (BE) and binaural unmasking (BU) concepts. The former relies on interaural level differences (ILDs) caused by the acoustical "shadow" cast by the head, which creates an advantageous signal-to-noise ratio (SNR) at the ear contra-lateral to the masker. In the latter, the interaural time differences (ITDs) give the hearing system the ability to increase the effective SNR by "cancelling" some of the masker signals (equalization-cancellation (EC) theory; Durlach, 1963).

The BE benefits are typically modeled in terms of audibility (Beutelmann *et al.*, 2010; Lavandier and Culling, 2010; Wan *et al.*, 2014), with a decision metric such as the speech intelligibility index (SII; ANSI, 1997). In other words, those models consider only energetic masking (EM), where EM is defined as masking of the

*Corresponding author: alech@elektro.dtu.dk

Proceedings of ISAAR 2015: Individual Hearing Loss – Characterization, Modelling, Compensation Strategies. 5th symposium on Auditory and Audiological Research. August 2015, Nyborg, Denmark. Edited by S. Santurette, T. Dau, J. C. Dalsgaard, L. Tranebjærg, and T. Andersen. ISBN: 978-87-990013-5-4. The Danavox Jubilee Foundation, 2015.

peripheral representation of the signal. However, Stone *et al.* (2012) showed that noises that are typically considered “steady”, such as speech-shaped noise (SSN), actually behave more as modulation maskers than as energetic maskers, i.e., they provide “modulation masking” (MM). Yet, EM and MM may not be sufficient to account for speech intelligibility data for some masker types, such as speech, in which case the unaccounted-for masking is labeled as “informational masking” (IM). According to Watson (2005), IM can be divided into two categories, uncertainty and similarity. Uncertainty is explained as a listener’s inability to identify the target, whereas similarity prevents a listener from segregating the target and the masker. Multiple factors can reduce the similarity between target and masker, such as spatial separation and fundamental frequency (F_0) information, and thus reduce IM (Bronkhorst, 2000).

The present study investigated the contributions of MM and IM and their interactions in an ITD-only binaural condition with a variable number of maskers (Lőcsei *et al.*, 2015) using a binaural extension of the multi-resolution speech-based envelope power spectrum model (mr-sEPSM; Jørgensen *et al.*, 2013; Chabot-Leclerc *et al.*, 2015). The mr-sEPSM framework considers MM using the SNR in the envelope domain ($\mathrm{SNR}_{\mathrm{env}}$) as the decision metric and was shown to account well for intelligibility where IM was not the dominating factor, such as with SSN maskers, sinusoidally modulated maskers, or multi-talker babble. Here, the maskers under consideration were SSN and time-reversed speech maskers, the latter known to produce informational masking, although not as much as regular speech (Rhebergen *et al.*, 2005). In particular, the focus was to analyze how well the $\mathrm{SNR}_{\mathrm{env}}$ metric could capture the intelligibility change as a function of the total number of maskers and the masker configuration and what could be attributed to IM.

MODEL DESCRIPTION

The structure of the proposed model is presented in Fig. 1. It consists of two monaural realizations of the mr-sEPSM (Jørgensen *et al.*, 2013) and a binaural unmasking pathway implemented as an EC process (Wan *et al.*, 2014).

The model takes as input the noisy speech and the noise-alone signals for each ear. Each signal is processed through a filterbank of 22 gammatone filters covering the frequency range from 63 Hz to 8 kHz with a third-octave spacing. The sub-band envelopes are then extracted using half-wave rectification followed by a fifth-order Butterworth low-pass filter with a cutoff frequency of 770 Hz (Breebaart *et al.*, 2001). Jitter in the time and amplitude domain is applied independently to each sub-band envelope to limit the efficacy of the EC process; all jitters are zero-mean Gaussian processes with standard deviations of $\sigma_\delta = 105$ µs for the temporal jitter and of $\sigma_\varepsilon = 0.25$ for the amplitude jitter (Durlach, 1963). In the monaural pathways, the envelopes are further processed by a modulation filterbank consisting of eight second-order band-pass filters with octave spacing between 2 and 256 Hz. A third-order low-pass filter with a 1-Hz cutoff frequency is applied in parallel to the filterbank.

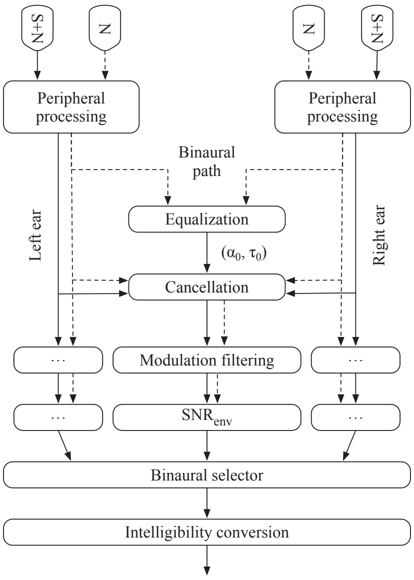

Fig. 1: Structure of the proposed model.

Only modulation filters with center frequencies below one-fourth of their respective peripheral-filter center frequency are used (Verhey *et al.*, 1999). The output of each modulation filter is then divided into non-overlapping segments of duration inversely proportional to the modulation filter's characteristic frequency, i.e., the output of the 4 Hz filter is divided into 250-ms segments. The power of each segment is calculated as its variance and the lower limit of the envelope power is set to -30 dB relative to 100% modulation. The $\mathrm{SNR}_{\mathrm{env}}$ for each segment, i, peripheral channel, p, and modulation channel, n, is calculated as:

$$\mathrm{SNR}_{\mathrm{env},i}(p,n) = \frac{P_{\mathrm{env},S+N,i}(p,n) - P_{\mathrm{env},N,i}(p,n)}{P_{\mathrm{env},N,i}(p,n)}, \qquad \text{(Eq. 1)}$$

where $P_{\mathrm{env},S+N}$ is the power of the noisy speech mixture and $P_{\mathrm{env},N}$ is the power of the noise alone.

The binaural unmasking stage is implemented as described in Wan *et al.* (2014). The jittered envelopes at the output of the peripheral filterbank are the inputs to the EC process, which is applied independently in each channel as well as in short 20-ms time frames, k. For each time-frequency frame, the equalization stage selects the optimal ITD, τ_0, and ILD, α_0, using the following equations:

$$\tau_0 = \arg\min_{\tau}\{\rho\}, \ \ |\tau| < \frac{\pi}{\omega}, \text{ and} \qquad \text{(Eq. 2)}$$

$$\alpha_0 = \sqrt{\frac{E_{N,L}}{E_{N,R}}}, \qquad \text{(Eq. 3)}$$

where ρ is the normalized cross-correlation function of the left and right ears within the frame, $E_{N,L}$ and $E_{N,R}$ are the masker energies for the left and right ear, respectively, and ω is the center frequency of the channel of interest. Subsequently, the sub-band signal, B_p, is reconstructed for each channel by summing over all frames.

The unmasked outputs for the noisy speech and the noise alone are then used as inputs to the modulation filtering stage of the mr-sEPSM and processed similarly to the monaural pathways, yielding a binaurally unmasked $\mathrm{SNR_{env}}$, $\mathrm{BU\text{-}SNR_{env}}$.

A selection stage then selects the best $\mathrm{SNR_{env}}$ of the left, right and binaural pathways, yielding the complete model's output, the $\mathrm{B\text{-}SNR_{env}}$. The $\mathrm{B\text{-}SNR_{env}}$ is then averaged across time, and combined optimally across modulation and peripheral filters:

$$\mathrm{B\text{-}SNR_{env}} = \left[\sum_{p=1}^{22}\sum_{n=1}^{9} \mathrm{B\text{-}SNR^2_{env}}(p,n)\right]^{1/2}. \qquad \text{(Eq. 4)}$$

The final $\mathrm{B\text{-}SNR_{env}}$ is then converted to intelligibility using a Gaussian psychometric function. The left- and right-ear pathways are combined and converted similarly, yielding alternate model outputs for each ear.

More details about the mr-sEPSM framework and the EC process implementation can be found in Jørgensen *et al.* (2013) and Wan *et al.* (2014), respectively.

METHODS

In this experiment, the speech and masker signals were lateralized individually to the left or right using fixed 33-sample delays (687.5 μs) and the spatial distribution of maskers was systematically varied. The speech material was the DAT corpus (Nielsen *et al.*, 2014), sampled at 48 kHz and recorded by female speakers. The DAT corpus consists of unique meaningful Danish sentences built as a fixed carrier sentence with two interchangeable target words. The maskers were either of one stationary SSN, denoted as $\mathbf{S}_{xy}$ conditions, or 2, 4, or 8 time-reversed sentences from the GRID corpus (Cooke *et al.*, 2006), denoted as $\mathbf{C}_{xy}$ conditions, where y is the total number of maskers and x is the number of maskers on the same side as the target. Both the SSN and the GRID material were shaped to have the same long-term spectrum as the target speech material. The maskers were either all on the same side as the target (e.g., $\mathbf{C}_{44}$), half on the same side (e.g., $\mathbf{C}_{24}$), or all on the opposite side (e.g., $\mathbf{S}_{04}$). The target level was fixed at 65 dB SPL and the maskers were summed before their levels were adjusted to the desired SNR. Model predictions were calculated for 30 randomly selected sentences and for SNRs ranging from −12 to 9 dB in 3-dB steps. The predicted SRT was the average across target sentences. The mean and standard deviation of the psychometric function were fitted to minimized the square error between the "left-ear" of the model and the word-scores as a function of SNR in the collocated condition ($\mathbf{S}_{11}$), as measured by Lőcsei *et al.* (2015).

RESULTS

Figure 2 shows the speech reception thresholds (SRTs) measured by Lőcsei *et al.* (2015) (open squares), the predictions by the proposed model (B-sEPSM; filled squares), as well as the predictions by the left- and right-ear outputs of the B-sEPSM (left- and right-pointing triangles, respectively) for each masker type and configuration. In the $\mathbf{S}_{x1}$ conditions with SSN maskers, the B-sEPSM predicted SRTs lower than the data by 0.5 to about 3 dB, but captured the MR when the maskers were moved to the opposite side. In the $\mathbf{C}_{x8}$ condition, the B-sEPSM accurately captured the MR when 4 and then all 8 reversed-speech maskers were lateralized to the other side. In the $\mathbf{C}_{x4}$ condition, the B-sEPSM predicted a similarly progressive MR as in the $\mathbf{C}_{x8}$ condition, as 2 or all 4 maskers were lateralized to the other side. This is in contrast to the data, where the SRT was constant at about -2.5 dB when 4 or 2 of the maskers were on the same side as the target and then there was about 5 dB of MR once all maskers were on the other side. In the $\mathbf{C}_{x2}$ condition, the B-sEPSM predicted constant SRTs of about -10 dB, irrespective of the masker arrangement. In contrast, the data SRTs were about the same when 2 or 1 masker(s) were collocated with the target at about -4 dB — not significant differences, $p < 0.05$ (Lőcsei *et al.*, 2015) — and then decreased by 4 dB once all maskers were on the other side, similar to the $\mathbf{C}_{x4}$ condition. The SRTs predicted by the left- and right-ear models (left- and right-pointing triangles) depended only on the total number of masker and masker type, irrespective of their configuration. The SRTs were highest in the $\mathbf{C}_{x8}$ and lowest in the $\mathbf{C}_{x2}$ condition, consistent with the increased number of dips in the two-masker condition. Overall, the Pearson correlation coefficient between the B-sEPSM predictions and the data was 0.78 and the mean absolute error was 2.24 dB.

DISCUSSION

The B-sEPSM could account well for the MR due to lateralization in the conditions with the SSN masker ($\mathbf{S}_{x1}$ conditions) and also accurately predicted the SRTs and MR in the $\mathbf{C}_{x8}$ conditions. However, the model was "too good" once the number of maskers was small enough such that IM became the dominating factor, i.e., in the conditions $\mathbf{C}_{x4}$ and $\mathbf{C}_{x2}$. A possible explanation framework has been put forward by Best *et al.* (2013), where it was suggested that intelligibility has a "lower limit" (of SRT) corresponding to the EM/MM present in the condition. In this case, the model's failure can be explained by the fact that it is blind to IM, and thus predicts the lower limit of intelligibility, given EM and MM only.

It is assumed that the mr-sEPSM framework has "perfect segregation" due to its access to the noisy-speech mixture and the noise-alone signals. Therefore, if most of the IM is due to confusion caused by the similarity between the target and maskers, and not to uncertainty about the target, then the B-sEPSM is blind to those confusions (Watson, 2005). An estimate of those confusions in the model would allow it to account for some of the IM in the listener. A possible approach would be to use a model of streaming, such as Elhilali and Shamma (2008) or

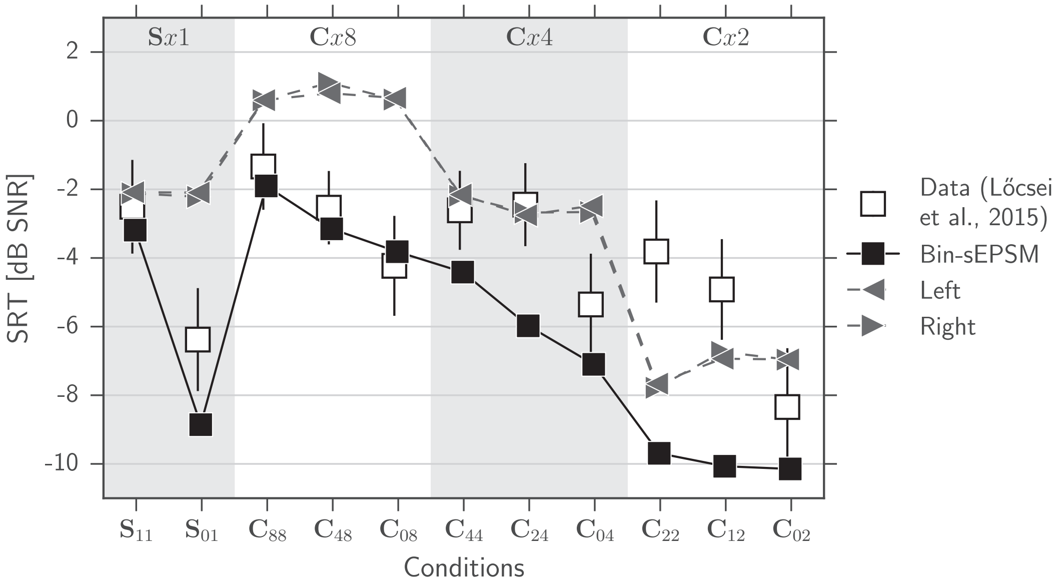

Fig. 2: Measured SRTs (open squares; Lőcsei *et al.*, 2015) and predictions by the B-sEPSM (filled squares) and the 'left-' and 'right-ear' models (triangles) for each condition. **S** conditions are with SSN maskers and **C** conditions are with reverse-babble maskers.

Christiansen *et al.* (2014), and to combine its output with the intelligibility model's output; a single-stream percept would lead to worse intelligibility than a multi-stream percept. Although that approach might prove powerful and possibly more realistic, it would greatly increase the complexity of the models, to the extent that two internal representations would be required. Figure 3 shows a potential similarity measure, calculated as a "modulation distance" between the speech estimate (i.e., $(S+N)-N$) and the noise-alone representations, as a function of the SNR and for different masker configurations. Given the three-dimensional representation of the envelope power as a function of sub-band frequency, modulation frequency, and time frames, the "modulation distance" is calculated as the Euclidean distance between the sub-band and modulation frequency representation (i.e., a 2D matrix) of the speech estimate and the noise for each time frame: The "distance" is then averaged across all time frames.

In Fig. 3, the black lines show the distance for the $\mathbf{C}_{x2}$ condition, where most IM was observed. The distance was largest in $\mathbf{C}_{02}$ condition (dashed line), whereas the distances for conditions $\mathbf{C}_{22}$ and $\mathbf{C}_{12}$ (solid and dotted lines) were almost the same. This mirrors the data, where an MR was observed once all maskers were not collocated with the target, i.e., confusions were resolved once spatial cues were available. In contrast, the distance varied much less as a function of masker location when MM was the dominating factor, such as in the SSN maskers conditions (dark gray lines, $\mathbf{S}_{x1}$) and in the eight-reversed speech masker conditions (light gray lines, $\mathbf{C}_{x8}$).

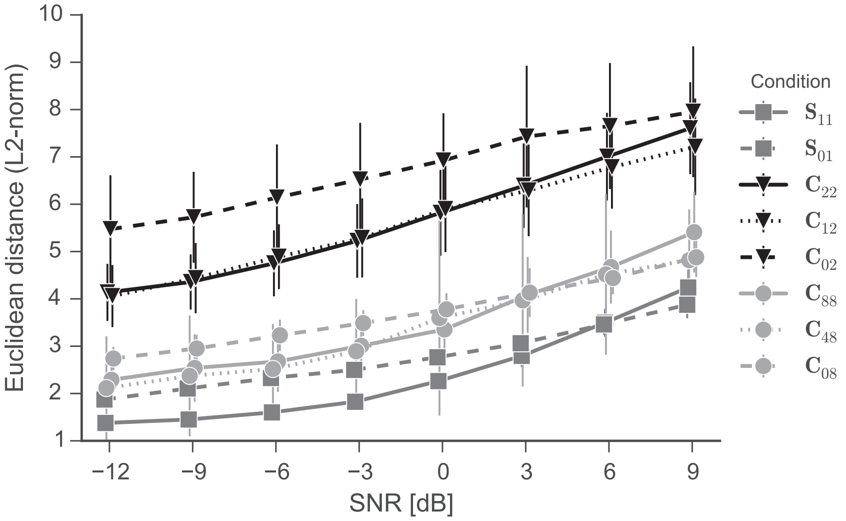

Fig. 3: Euclidean distance between the speech estimate and the noise in the envelope power domain, as a function of SNR. Each line represents a different condition.

In summary, the B-sEPSM could accurately predict SRTs when the dominating factor was modulation masking, but failed when IM became more prevalent. It seems that similarity information between the target estimate and the maskers is available in the multi-resolution envelope power representation and that it could be used to account for some of the IM. However, more work is required in order to combine this information with the binaural model predictions.

ACKNOWLEDGMENTS

This research was supported in part by the National Science and Engineering Research Council of Canada (NSERC), Phonak, and the Technical University of Denmark.

REFERENCES

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

Best, V., Thompson, E.R., Mason, C.R., and Kidd, G. (**2013**). “An energetic limit on spatial release from masking,” J. Assoc. Res. Otolaryngol., **14**, 603-610.

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.

Breebaart, J., van de Par, S., and Kohlrausch, A. (**2001**). “Binaural processing model based on contralateral inhibition. I. Model structure,” J. Acoust. Soc. Am., **110**, 1074-1088.

Bronkhorst, A. (**2000**). “The cocktail party phenomenon: A review of research on speech intelligibility in multiple-talker conditions,” Acta Acust. United Ac., **86**, 117-128.

Chabot-Leclerc, A., MacDonald, E.N., and Dau, T. (**2015**). “Predicting binaural speech intelligibility using the signal-to-noise ratio in the envelope power domain,” J. Acoust. Soc. Am., submitted.

Christiansen, S.K., Jepsen, M.L., and Dau, T. (**2014**). “Effects of tonotopicity, adaptation, modulation tuning, and temporal coherence in “primitive” auditory stream segregation,” J. Acoust. Soc. Am., **135**, 323-333.

Cooke, M., Barker, J., Cunningham, S., and Shao, X. (**2006**). “An audio-visual corpus for speech perception and automatic speech recognition,” J. Acoust. Soc. Am., **120**, 2421-2424.

Durlach, N. (**1963**). “Equalization and cancellation theory of binaural masking-level differences,” J. Acoust. Soc. Am., **35**, 1206-1218.

Elhilali, M. and Shamma, S.A. (**2008**). “A cocktail party with a cortical twist: How cortical mechanisms contribute to sound segregation,” J. Acoust. Soc. Am., **124**, 3751-3771.

Jørgensen, S., Ewert, S.D., and Dau, T. (**2013**). “A multi-resolution envelope-power based model for speech intelligibility,” J. Acoust. Soc. Am., **134**, 436-446.

Lavandier, M. and Culling, J.F. (**2010**). “Prediction of binaural speech intelligibility against noise in rooms,” J. Acoust. Soc. Am., **127**, 387-399.

Lőcsei, G., Hefting Pedersen, J., Laugesen, S., Santurette, S., Dau, T., and MacDonald, E.N. (**2015**). “Lateralized speech perception, temporal processing and cognitive function in NH and HI listeners,” Poster presented at Speech in Noise Workshop (Copenhagen, Denmark).

Nielsen, J.B., Dau, T., and Neher, T. (**2014**). “A danish open-set speech corpus for competing-speech studies,” J. Acoust. Soc. Am., **135**, 407-420.

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.

Stone, M.A., Füllgrabe, C., and Moore, B.C.J. (**2012**). “Notionally steady background noise acts primarily as a modulation masker of speech,” J. Acoust. Soc. Am., **132**, 317-326.

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

Wan, R., Durlach, N.I., and Colburn, H.S. (**2014**). “Application of a short-time version of the equalization-cancellation model to speech intelligibility experiments with speech maskers,” J. Acoust. Soc. Am., **136**, 768-776.

Watson, C.S. (**2005**). “Some comments on informational masking,” Acta Acust. United Ac., **91**, 502-512.