4pPP3. Physiological prediction of masking release for normal-hearing and hearing-impaired listeners

Ian C. Bruce*, Agnès C. Léger, Brian C. Moore and Christian Lorenzi

*Corresponding author’s address: Department of Electrical & Computer Engineering, McMaster University, Room ITB-A213, Hamilton, L8S 4K1, Ontario, Canada, ibruce@ieee.org

Léger et al. (2012c) measured the intelligibility of speech in steady and spectrally or temporally modulated maskers for stimuli filtered into low- (< 1.5 kHz) and mid-frequency (1−3 kHz) regions. Listeners with high-frequency hearing loss but near to clinically normal audiograms in the low- and mid-frequency regions showed poorer performance than a control group with normal hearing, but showed preserved spectral and temporal masking release. Here, we investigated whether a physiologically accurate model of the auditory periphery (Zilany et al., 2009) can explain these masking release data. Intelligibility was predicted using the Neurogram SIMilarity (NSIM) metric of Hines and Harte (2010 and 2012). This metric can make use of either an “all-information” neurogram with small time bins or a “mean-rate” neurogram with large time bins. The average audiograms of the different groups of listeners from the study of Léger et al. were simulated in the model by applying different mixes of outer and/or inner hair cell impairment. Very accurate predictions of the human data for both normal-hearing and hearing-impaired groups were obtained from the all-information NSIM metric (i.e., taking into account phase-locking information) with threshold shifts produced predominantly by OHC impairment (and minimal IHC impairment).

Published by the Acoustical Society of America through the American Institute of Physics
INTRODUCTION

“Masking release” corresponds to the improvement of speech intelligibility because of some change in the masker or in the target/masker relationship. One form of monaural masking release is observed for modulated compared with unmodulated maskers, where the masker can be either temporally modulated (temporal masking release) or spectrally modulated (spectral masking release). Léger et al. (2012c) evaluated temporal and spectral masking release for speech filtered into low- and mid-frequency regions. They found that hearing-impaired listeners with normal or near-normal audiograms in these frequency regions had reduced overall speech perception but preserved masking release. The purpose of this present study is to determine the physiological basis for these results by use of a computational model of the auditory periphery (Zilany et al., 2009).

The experiments of Léger et al. (2012c) utilized four sets of 48 Vowel-Consonant-Vowel (VCV) stimuli in a consonant identification task. In Experiment I, they studied the perception of Low-Frequency Speech, that is, speech lowpass filtered at 1.5 kHz, while Experiment II investigated Mid-Frequency Speech (speech bandpass filtered between 1 and 3 kHz). For each experiment, different groups of normal hearing (NH) and hearing impaired (HI) listeners were tested. The mean audiograms of the four groups of listeners are shown in Fig. 1. Speech identification was measured in quiet and in six different masker noises presented at three different signal-to-noise ratios (−6, −3 and 0 dB SNR):

- Notionally steady (unmodulated) speech-shaped noise.
- Temporally-modulated noises (8-Hz square wave, 100% modulation depth) at a:
  - 50% duty-cycle (DC), or
  - 25% duty-cycle.
- Spectrally-modulated noises created by passing noise through an array of gammatone filters, each with a bandwidth of 1 ERBN (Glasberg and Moore, 1990), and setting to zero the output of:
  - one filter out of every two (1ERBN/2),
  - two adjacent filters out of every four (2ERBN/4), or
  - three adjacent filters out of every four (3ERBN/4).

The consonant identification results are shown in Fig. 2. The results in the left column for the NH listeners indicate that they received a moderate amount of masking release from the introduction of moderate spectral dips in the masker (1ERBN/2 and 2ERBN/4) relative to the steady-noise condition (unmodulated). Further masking release was produced by larger spectral dips in the masker (3ERBN/4). Providing temporal dips in the masker also produced masking release, with a noise duty-cycle of 50% giving consonant identification results similar to those for moderate spectral dips. Introducing larger temporal dips gave further masking release, with a noise duty-cycle of 25% producing identification scores similar to those for the case of larger spectral dips. It should be noted that for the conditions with temporal dips in the noise, the mean identification scores do not vary greatly over the range of SNRs from −6 to 0 dB, in contrast to the steady-noise and spectral-dip conditions, which show a fairly strong effect of SNR. The results for the HI listeners (right column of Fig. 2) show the same general patterns of masking release but with an overall deficit in all listening conditions, which suggests that they demonstrated supra-threshold deficits in the tested frequency regions of normal or near-normal audiometric thresholds. Note that although the HI groups had normal or near-normal audiometric thresholds, their absolute thresholds were higher than for the NH groups, which suggests that elevation of audiometric threshold within the range conventionally considered as “normal” can be associated with deleterious effects on speech intelligibility in quiet and in noise.
Figure 1: Audiograms of the listener groups from Léger et al. (2012c). The top panel is for Experiment I, in which the listeners were presented with Low-Frequency Speech (<1.5 kHz), while the bottom panel is for Experiment II, in which different groups of listeners were presented with Mid-Frequency Speech (1–3 kHz). The curves show mean air conduction thresholds of the test ears of the NH and HI groups of listeners. Error bars indicate ±1 SD. Horizontal dashed lines show the limits of normal (≤20 dB HL) and near-normal (≤30 dB HL) audiometric thresholds. The yellow shaded areas show the frequency limit of the stimuli used in the two experiments. (Adapted from from Léger et al. (2012c) with permission from the Acoustical Society of America © 2012).

Figure 3 shows that an acoustic-based speech intelligibility predictor, the extended speech intelligibility index (ESII) of Rhebergen and Versfeld (2005) and Rhebergen et al. (2006), is able to partially describe the masking release for the cases of temporal dips in the speech-shaped noise. However, the ESII is unable to predict the masking release for the cases of spectral dips in the noise, nor the pattern of SNR-dependence for the different noise conditions. The ESII does predict the overall drop in consonant identification scores for HI listeners (right column of Fig. 3), but this is achieved in an ad hoc fashion—because the ESII is an acoustic-based metric, it cannot directly incorporate the effects of cochlear pathology. This motivates the use of a physiologically-based speech intelligibility predictor, as described in the next section.
FIGURE 2: Mean consonant identification scores for VCVs, expressed in rationalized arcsine units (RAU), for the NH (left panels) and HI (right panels) listeners of Experiments I (top panels) and II (bottom panels), plotted as a function of SNR (in dB). Conditions were: in quiet (crosses), in temporally-modulated noise (squares), in spectrally-modulated noise (circles and asterisks), and in unmodulated noise (triangles). Error bars indicate ±1 standard error. (Adapted from Léger et al. (2012c) with permission from the Acoustical Society of America © 2012).

**Physiological model predictions**

The auditory-periphery model of Zilany et al. (2009) has been well validated against a wide range of physiological data for simple and complex stimuli, including speech, and can take into account the effects of impairment of inner (IHCs) and outer (OHCs) hair cells. The response of a population of model auditory nerve (AN) fibers can be represented as a “neurogram”, in which the spike counts for AN fibers tuned to different characteristic frequencies (CFs) are plotted as a function of time. If the AN response is calculated with large time bins, then the resulting representation is referred to as a “mean-rate” neurogram, as the information about spike timing is not preserved. In contrast, if small time bins are utilized, then an “all-information” neurogram is generated, which maintains the information about the fine timing of spikes.

In order to predict speech intelligibility from the neurogram representation, we use the Neurrogram SIMilarity metric (NSIM) of Hines and Harte (2010, 2012) to determine how close the neural representation of a VCV stimulus processed as in the Léger et al. (2012c) experiments is to a template neurogram (in this study, the normal model response to the respective intact stimulus at 65 dB SPL in quiet). This metric is based on the Structural SIMilarity (SSIM) index of Wang et al. (2004), which provides a more accurate prediction of image quality than the pixel-by-pixel mean-squared error. The applicability of this metric for speech in quiet at different presentation levels and for different degrees of OHC and IHC impairment has been demonstrated by Hines and Harte (2010, 2012).
In this study we utilized “mean-rate” neurograms with a time step of 6.4 ms and “all-information” neurograms with a time step of 0.16 ms. At each CF the spike count was summed for fifty model AN fibers with a physiologically-realistic mix of spontaneous rates and corresponding thresholds (Liberman, 1978). The CFs were chosen to match the center frequencies of the filters used in the ESII. In addition, the same frequency band importance function as used in the ESII was utilized to weight the NSIM scores before averaging over time and frequency. Furthermore, a time-frequency window was applied to the neurograms such that only the neural response to i) the central consonant portions of the VCVs and ii) the filtered speech frequencies (that is, below 1.5 kHz for the LowF speech and 1–3 kHz for the MidF speech) contributed to the NSIM calculation.

We explored a number of different patterns of pathology that would be consistent with the mean group audiograms shown in Fig. 1. For an average listener, approximately two-thirds of the threshold shift measured in the audiogram can be attributed to OHC impairment and one-third to IHC impairment (Plack et al., 2004), so we utilized this as our default model impairment. However, there is substantial variability across individuals (Plack et al., 2004; Moore and Glasberg, 2004; Lopez-Poveda and Johannesen, 2012), so we also ran simulations in which all of the threshold shift was produced by IHC impairment and others in which OHC impairment accounted for almost all of the threshold shift. In addition, we investigated cases of partially missing high-spontaneous-rate, low-threshold AN fibers, consistent with the observations of Liberman and Dodds (1984), and of completely missing low-spontaneous-rate, high threshold fibers, as described by Kujawa and Liberman (2009).
In general, the “all-information” NSIM results better predicted the patterns of masking release in the human consonant identification scores than did the “mean-rate” NSIM, suggesting that for each CF, neural phase locking to the speech temporal fine structure conveys crucial identification cues. The left panels of Fig. 4 show that the NSIM metric gives better predictions of the human data than does the ESII. Of the different variants in cochlear pathology that we investigated, the case of primarily OHC impairment (with minimal IHC impairment) was closest to the human data. These predictions are shown in the right panel of Fig. 4.

The dominance of OHC impairment in creating the observed supra-threshold deficit is consistent with the results of the follow-up study of Léger et al. (2012a), in which spectral smearing was applied to simulate reduction of frequency selectivity for a group of NH listeners. It was found that substantial spectral smearing could explain the general reduction in consonant perception, but Léger et al. (2012a) estimated that the cochlear tuning of the HI groups in the earlier study (with the audiograms shown in Fig. 1) would not be broad enough to produce such substantial spectral smearing. However, this method of simulating broadened tuning with NH listeners does not take into account the loss of cochlear “two-tone” suppression that occurs with OHC impairment (in addition to the broadened baseline tuning). Thus, the results obtained with the physiological predictor in this paper indicate that OHC impairment could indeed explain the majority of the supra-threshold deficits in consonant perception but preserved temporal and spectral masking release measured by Léger et al. (2012c). Recently, Léger et al. (2012b) observed a significant correlation between otoacoustic emissions (OAE) and speech identification scores in a second follow-up study that tested similar NH and HI listeners using a similar task, providing further evidence that OHC impairment might be causing the observed supra-threshold deficits in speech intelligibility in both quiet and noise.

![Figure 4: “All-information” NSIM predictions. OHC impairment was applied to the model to match the respective group audiograms. The plotting conventions match those used for the data in Fig. 2.](image-url)
CONCLUSIONS

The physiologically-based speech intelligibility predictor (NSIM) gave a better prediction of spectral and temporal masking release than did the acoustical-based model (ESII) for the speech identification scores of NH and HI listeners tested in frequency regions of near-normal audiometric thresholds. The best fit was obtained when the predictions were computed using small time bins, which suggests a role for temporal fine structure information. Modeling the cochlear pathology using primarily OHC impairment (with minimal IHC impairment) gave the best predictions for the human data, which suggests that this pathology could underly the supra-threshold deficits demonstrated by the HI listeners in Léger et al. (2012c).

ACKNOWLEDGMENTS

Programming assistance from M. R. Wirtzfeld. Funded by the Erasmus Mundus ACN Training Network, NSERC Discovery Grant 261736, the Royal Society (International Joint Project, 2009R3), 7FP-SME222291 Dual-Pro program, MRC (UK), and a CIFRE grant from ANRT and Neurelec.

REFERENCES


