Challenging the Speech Intelligibility Index: Macroscopic vs. Microscopic Prediction of Sentence Recognition in Normal and Hearing-impaired Listeners

Tim Jürgens, Stefan Fredelake, Ralf M. Meyer, Birger Kollmeier, Thomas Brand
Medizinische Physik, Carl-von-Ossietzky Universität Oldenburg, Germany
{tim.juergens, stefan.fredelake, ralf.m.meyer, birger.kollmeier, thomas.brand}@uni-oldenburg.de

Abstract
A “microscopic” model of phoneme recognition, which includes an auditory model and a simple speech recognizer, is adapted to model the recognition of single words within whole German sentences. “Microscopic” in terms of this model is defined twofold, first, as analyzing the particular spectro-temporal structure of the speech waveforms, and second, as basing the recognition of whole sentences on the recognition of single words. This approach is evaluated on a large database of speech recognition results from normal-hearing and sensorineural hearing-impaired listeners. Individual audiometric thresholds are accounted for by implementing a spectrally-shaped hearing threshold simulating noise. Furthermore, a comparative challenge between the microscopic model and the “macroscopic” Speech Intelligibility Index (SII) is performed using the same listeners’ data. The results are that both models show similar correlations of modeled Speech Reception Thresholds (SRTs) to observed SRTs.

Index Terms: speech intelligibility, auditory model, sentence recognition test, hearing impairment

1. Introduction
The Speech Intelligibility Index (SII) [1] is widely used to predict human speech recognition (HSR) in different noise conditions or for subjects with different audiometric hearing losses. The SII can be called a “macroscopic” model, as it uses only the long-term spectra of speech and noise separately, whereas the particular temporal structure of speech and noise is disregarded. Speech intelligibility is predicted using a weighted sum over the Signal-to-Noise-Ratios (SNRs) in different frequency bands, resulting in an SII value between 0 and 1. The weighting factors are tabulated and depend on the context or the articulation style of the speech material used [1].

Figure 1 displays averaged audiogram data of the NH group and two groups of HI listeners (black lines), including the ranges between the 5th and 95th percentiles. The first group of HI listeners showed nearly normal hearing at low frequencies (≤ 30 dB HL between 125 Hz and 1 kHz) and hearing loss at higher frequencies (HI-H). The second group showed a hearing loss both at low and high frequencies (HI-LH). Listeners were paid for their participation in the experiments.

2. Measurements
2.1. Subjects
15 normal-hearing (NH) listeners aged from 24 to 34 years and 48 sensorineural hearing-impaired (HI) listeners aged from 17 to 82 years participated in this study. In 51 listeners both ears were tested separately, resulting in a total of 114 investigated ears. NH listeners showed pure-tone thresholds of not more than 15 dB Hearing Level (HL) using standard audiometry (IEC60645-1). The SII value is transformed to a speech recognition rate in percent using a nonlinear function that depends on the speech material.

A psychoacoustically-driven, “microscopic” model of HSR, on the other hand, models the recognition of single phonemes [2] by analyzing the particular spectro-temporal structure of speech and noise. An “internal representation” (IR) is computed from the waveform of the speech/noise-mixture using an auditory model and employing a simple speech recognizer. Thus, it mimics the individual auditory signal processing in a much more realistic way than the SII. The main goal of this study is first, to adapt this microscopic model from phonemes to sentences and second, to compare the predictive power of this modeling approach with that of the SII. For the comparative challenge of the two models, an ambitious speech recognition data set is used with perceptually similar (rather than physically equal) acoustic measurement conditions for all listeners. This means that signals with higher levels were used for hearing-impaired listeners to ensure equal loudness perception of these signals.

2.2. Apparatus
All stimuli were presented monaurally via Sennheiser HDA 200 headphones that were free-field equalized using an FIR-filter with 801 coefficients, while the listeners were seated in a sound-insulated booth. The headphones were connected to a computer-controlled audiometry workstation that was developed within a German joint research project on speech audiometry [3].
2.3. Speech Intelligibility Measurements

Speech intelligibility in stationary ICRA1 noise [4] was measured using the Oldenburg sentence test [5] that is part of the Oldenburg Measurement Applications (OMA) software by HörTech gGmbH. The Oldenburg sentence test consists of German sentences with a fixed syntactic structure name-verb-number-adjective-object, e.g., 'Peter gets five wet cars', spoken by a male speaker. Each word of the sentence was chosen from ten alternatives, respectively. Such sentences were combined in lists consisting of 30 sentences each that were optimized with respect to equal speech intelligibility [5]. Within one measurement run, one list of sentences was presented. An adaptive procedure [6] was used to measure the Speech Reception Threshold (SRT), i.e. the SNR at 50% speech recognition rate for the sentences of this list as follows. After the presentation of each sentence, the level of the speech was adaptively varied in two randomly interleaved tracks. One track converged at 80% and the other track converged at 20% speech recognition rate. Both tracks started with an SNR of 0 dB. After each run the SRT was calculated by fitting a logistic function with the parameters SRT and slope to all collected data using a maximum likelihood estimator [6]. During the measurements the level of the noise was fixed at a level that individually corresponded to medium loudness. This means that all listeners were tested under perceptually similar, rather than physically equal conditions. At least two test lists were measured as training in advance. Subjects were asked to repeat each presumably understood word after presenting the whole sentence (open test). An investigator marked the correctly recognized words using a touch screen response box.

3. Modeling

3.1. Speech Intelligibility Index (SII)

The SII was calculated according to [1] using the long-term spectrum of the ICRA1 background noise and the long-term spectrum of the complete speech material of the Oldenburg sentence test [5]. The critical frequency band method was used and the standard speech spectrum level for stated vocal effort was chosen according to ‘normal’ speech articulation. Individual audiogram data, interpolated at the center frequencies of the critical frequency bands, was used to calculate the equivalent hearing threshold level. As critical band importance function the values for SPeech In Noise (SPIN) were chosen. The modeled psychometric function, i.e. SII-values for each listener as a function of SNR, was calculated using the same fixed noise level as in the measurements and speech levels in the range of 40 to 100 dB. After each run the SII was calculated according to [1] using the long-term speech recognition rate for the sentences of this list as follows. Within one measurement run, one list of sentences was measured using the Oldenburg sentence test [5] that is part of the Oldenburg Measurement Applications (OMA) software by HörTech gGmbH. The Oldenburg sentence test consists of German sentences with a fixed syntactic structure name-verb-number-adjective-object, e.g., 'Peter gets five wet cars', spoken by a male speaker. Each word of the sentence was chosen from ten alternatives, respectively. Such sentences were combined in lists consisting of 30 sentences each that were optimized with respect to equal speech intelligibility [5]. Within one measurement run, one list of sentences was presented. An adaptive procedure [6] was used to measure the Speech Reception Threshold (SRT), i.e. the SNR at 50% speech recognition rate for the sentences of this list as follows. After the presentation of each sentence, the level of the speech was adaptively varied in two randomly interleaved tracks. One track converged at 80% and the other track converged at 20% speech recognition rate. Both tracks started with an SNR of 0 dB. After each run the SRT was calculated by fitting a logistic function with the parameters SRT and slope to all collected data using a maximum likelihood estimator [6]. During the measurements the level of the noise was fixed at a level that individually corresponded to medium loudness. This means that all listeners were tested under perceptually similar, rather than physically equal conditions. At least two test lists were measured as training in advance. Subjects were asked to repeat each presumably understood word after presenting the whole sentence (open test). An investigator marked the correctly recognized words using a touch screen response box.

3.2. Microscopic model

The microscopic model of speech recognition was implemented very similar to the approach of Jürgens and Brand (2009) [2] for NH listeners and was extended to HI listeners and to a sentence test in the present study. A word from the Oldenburg sentence test, mixed with ICRA1 background noise with an SNR ranging from –15 to 15 dB in 3 dB steps is added to a hearing threshold simulating noise that is spectrally shaped to the individual audiogram data of the listener’s ear (cf. Figure 2). Subsequently, the Perception Model (PeMo, [7]) computes an IR from this signal. The PeMo implementation used in the present study consists of a gammatone filter bank with 27 frequency channels ranging from 236 Hz to 7469 Hz center frequency. The gammatone filterbank models the peripheral filtering in the cochlea.

A haircell-model computes the temporal envelope in each frequency channel and adaptation loops emphasize on- and offsets of the signal. A modulation filterbank with four modulation channels evaluates low speech modulations up to about 20 Hz. Consecutively, the IR is downsampled to a sampling frequency of 100 Hz and thus contains a feature-matrix of 27 frequency channels and four modulation frequency channels at each 10 ms time step. PeMo is capable of modeling psychoacoustical data, e.g. of forward and backward masking experiments, and modulation detection in normal-hearing listeners [7].
lower part of Figure 3). A Dynamic-Time-Warp (DTW) speech recognizer computes pairwise the Lorentzian distance ("perceptive" distance, cf. [2]) between IRtest and the IRs in the vocabulary by locally stretching and compressing the time axes. That word from the vocabulary with the smallest perceptive distance to the test word is taken as the recognized one. Note that the exact speech waveform to recognize is always also contained in the vocabulary, i.e. the detector stage is assumed to be optimal (cf. [2]). However, the waveforms of the speech/noise mixtures that enter PeMo are always different due to different temporal passages of the background noise and the hearing threshold simulating noise. For each test word the recognition procedure was conducted nine times using different temporal passages of background noise and hearing threshold simulating noise. The speech recognition rate for a given listener and SNR was then calculated as the average over the nine repetitions, different parts of the sentence, and different sentences. The whole calculation was performed on a computer cluster of the University of Oldenburg.

4. Results and comparison

Figure 4: 20 observed (grey solid lines) and one modeled psychometric function (triangles and black solid line) of speech intelligibility of NH listeners using the microscopic model.

Figure 4 shows the modeled recognition rates (triangles) using the microscopic model for a NH listener with 0 dB HL at all audiometric frequencies. Note that the modeled recognition rates were corrected for the random hit rate of 10% that is inherent in this modeling approach, but not inherent in the open-set speech intelligibility measurements. A fit to the modeled recognition rates using a logistic function (psychometric function, black solid line, cf. [2]) results in the optimal fit parameters SRT = -1.9 dB SNR and slope = 11.3 %/dB. Additionally, the psychometric functions of the 20 NH ears are plotted as grey solid lines. Substantial inter-individual differences (about 5 dB) in the SRT between NH ears can be observed. The modeled psychometric function shows an SRT that is about 5 dB higher than the average SRT of the NH listeners. Furthermore, the slope of the modeled psychometric function is slightly shallower than the slopes of the psychometric functions of the NH ears.

Figure 5 presents the predicted SRTs using the SII (upper panel) and the microscopic model (lower panel) versus the observed SRTs for NH and HI listeners. Dashed lines indicate the 95% confidence boundaries that were calculated as ± twice the standard deviation of the test-retest SRT-difference (1.4 dB) measured in a subset of the listeners. The SII shows a correlation of $r^2 = 0.25$ ($p < 0.001$) using Pearson’s correlation coefficient $r$. 82% of the predicted SRTs fall within the confidence boundaries of the measurement procedure. HI-LH data (red crosses) show the largest inter-individual variation in both, observation and prediction. NH data (green triangles) show only inter-individual variation in the observations, but almost no variation in the predictions. Most of the data are clustered in a region that covers about 3 dB in the predictions and about 10 dB in the observations. The microscopic model shows a correlation of $r^2 = 0.28$ ($p < 0.001$). 57% of the data points fall within confidence boundaries. The microscopic model, as well as the SII, is not able to predict the variation in NH listener’s data according to different audiometric thresholds. However, using the microscopic model, HI data points show less clustering than observed using the SII, which indicates that the microscopic model predicts larger differences of SRTs due to individual audiometric thresholds and testing conditions (i.e. background noise levels) of the HI listeners. Concerning the individual slopes of the psychometric functions, the microscopic model shows only a poor correlation of $r^2 = 0.09$ ($p < 0.01$). On average, the modeled psychometric functions (average slope 10%/dB) are shallower than the observed psychometric functions (16%/dB).

5. Discussion

SII and microscopic model show similar correlations between predicted and observed SRTs. This indicates that it is possible
to achieve the same performance as the SII, concerning individual differences of HI listeners, using a psychoacoustically-driven microscopic model. However, there is a difference between the models regarding the number of SRT values in confidence intervals. The fact that over 80% of the SII-predicted SRT values fall within confidence boundaries was achieved by assuming an SII-value of 0.24 at the SRT. This value is adjustable and was set in order to reproduce the average SRT of all listeners. With the microscopic model such an optimization was not necessary or rational. However, both models are not able to model the remarkable inter-individual differences of listeners with nearly the same audiometric thresholds (e.g., of NH listeners), which indicates the existence of an important, not adequately modeled individual factor on speech intelligibility. Some of the following factors might explain parts of the differences between measurements and predictions.

Semantic context effects might be responsible for the predicted SRTs of NH listeners being higher and for the predicted psychometric functions being shallower than the respective observed value. Although the sentences of the Oldenburg sentence test contain low semantic context, listeners might still benefit in their recognition performance due to co-articulation between subsequent words and due to the prosody of the sentence, which cannot be used by the model. The amount of this benefit might be subject-dependent and thus might explain parts of the remaining variance in the data. Too shallow psychometric functions are also reported in [8] when modeling human sentence recognition using an auditory model and an information-theoretic framework. Within the scope of their framework, the authors of [8] attributed the too shallow modeled psychometric function to a non-optimal probabilistic speech model they used. However, since an “optimal detector” approach is used in the present study, this reason does not hold here.

The microscopic model assumes the Oldenburg sentence test to be a closed test, although the measurement was performed as an open test. Modeled psychometric functions that show a random hit probability of 10% at low SNRs are scaled to cover the whole range of possible recognition rates (cf. Figure 4). A closed test approach was also used in [2] and has the advantage that only limited speech material is needed as possible response alternatives for the speech recognizer. Using this approach for the open speech test used here seems to be feasible, since a study comparing the results of the open and closed version of the Oldenburg sentence test for NH listeners revealed no significant differences, as long as the listeners have been trained prior to the test [9].

The individual measurement conditions might be a factor responsible for the prediction of better speech recognition performance (i.e., lower SRTs) for some of the HI listeners than for NH listeners by the microscopic model (red crosses in the SRT observed in NH listeners).

6. Conclusions

The microscopic model of human sentence recognition applied to speech recognition data of normal-hearing and sensorineural hearing-impaired listeners shows similar performance as the standard SII. However, the different modeling blocks of the microscopic model aim at mimicking human speech processing much more closely than the SII. Furthermore, the microscopic model has the potential to be extended to model the effects of context of the speech material on speech recognition and to investigate how different individual aspects of hearing impairment affect sentence recognition.

7. Acknowledgements

Thanks to Sven Kissner and the Hörzentrum Oldenburg GmbH for the execution of the measurements. This work is supported by SFB TR 31 “The active auditory system” and the “Audiologie-Initiative Niedersachsen”.

8. References


