Predicting the intelligibility of processed noisy speech based on the signal-to-noise ratio in the modulation domain

Torsten Dau, Søren Jørgensen

Technical University of Denmark, 2800 Lyngby, Denmark, Email: tdau@elektro.dtu.dk

Introduction

A major challengein current hearing research is the "noise reduction paradox" that refers to the apparent mismatch between predicted and actual speech intelligibility following noise reduction signal processing. The Speech Transmission Index (STI) and related models successfully predict effects of linear distortions, such as noise and reverberation [1], but fail to predict the effects of nonlinear signal processing and noise reduction, such as spectral subtraction [2]. The STI considers the reduction in speech modulations as the critical physical characteristic related to speech intelligibility. However, recent investigations suggest that the ratio of speech to noise energy in the modulation domain might be a crucial indicator [3]. Here, a new model for predicting the intelligibility of processed noisy speech is proposed based on the envelope power spectrum model (EPSM) originally developed to account for modulation detection data [4]. The model estimates the speech-to-noise envelope power ratio at the output of a modulation filterbank, and relates this metric to speech intelligibility using an optimal detector. Model predictions are compared to data from the literature as well as new experimental data of noisy speech subjected to spectral subtraction.

Model description

The processing stages of the speech-based envelope power spectrum model (sEPSM) are shown in Fig. 1(A). The stages consist of a gammatone bandpass filterbank followed by envelope extraction via Hilbert transformation and a modulation bandpass filterbank. The envelope signal-to-noise ratio, $\mathrm{SNR_{env}}$, is calculated from the long-term integrated envelope power at the output of each modulation filter and the resulting values are combined optimally across modulation filters and across gammatone filters. The overall $\mathrm{SNR_{env}}$ is converted to percent correctly recognized speech using the concept of an "ideal observer" which contains two parameters that reflect the number of response alternatives and the redundancy of a given speech material.

The scheme for predicting intelligibility of noisy speech is shown in Fig. 1(B), whereby noisy speech and noise alone are subjected separately to the same transmission-channel processing, such as reverberation or spectral subtraction, and then analyzed by the sEPSM. Here, the noise alone represents an estimate of the intrinsic noise within the noisy speech. Figure 1(C) illustrates the effect of the transmission-channel processing on $\mathrm{SNR_{env}}$ and on the corresponding predicted percent correct as a function

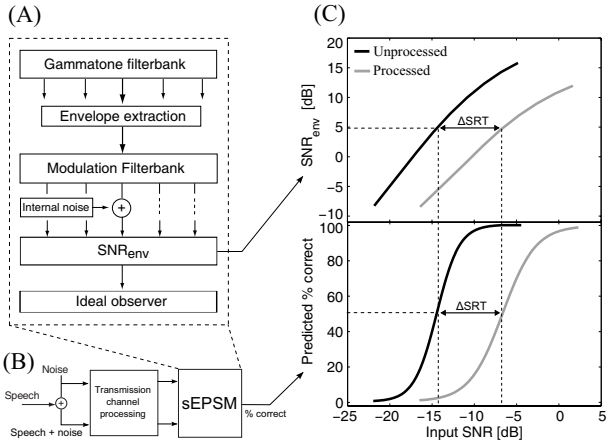

Figure 1: Block-diagram of the model structure. See the text for a description of each stage.

of the input SNR. The change in the speech recognition threshold, ΔSRT, is predicted by the corresponding shift (in terms of the input SNR) at the 50 % point of the predicted psychometric function.

Method

Model predictions were compared to intelligibility data of noisy speech for three Danish speech materials: (i) the CLUE [5] sentence material, (ii) the DANTALE II [6] sentence material, and (iii) the DANTALE [7] word material. Own speech intelligibility data were collected for noisy speech processed by spectral subtraction, measuring the SRT using the CLUE sentence-test. The spectral subtraction was defined as:

$$\hat{S}(f) = \left[P_{S+N}(f) - \alpha \hat{P}_N(f)\right]^{1/2} \qquad (1)$$

where $\hat{S}(f)$ denotes the estimated clean-speech power spectrum, $\hat{P}_N(f)$ is an estimate of the noise power spectrum, $P_{S+N}(f)$ is the power spectrum of the noisy speech and α denotes the over-subtraction factor which controls the amount of subtraction. The experimental parameter was α which took the values: 0, 0.5, 1, 2, 4 or 8.

Results

Figure 2 shows the measured (normative) psychometric functions (solid curves) for the three speech materials. The corresponding simulations are indicated as filled symbols. For all three speech materials, the predicted percent correct values are in very good agreement with

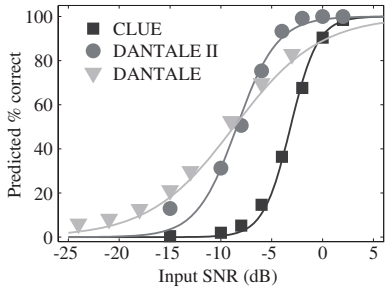

Figure 2: Psychometric functions based on measured data (solid lines) and predicted % correct (filled symbols).

the measured values. This results from the calibration of the ideal observer, whereby the model was adjusted to account for the differences between the three materials. The calibrated values were kept fixed when predicting ΔSRT. Figure 3 shows ΔSRT as a function of the over-subtraction factor α. The open squares represent measured data, averaged across 4 normal-hearing listeners, where the SRT in the reference condition was obtained at an SNR of -3.3 dB, consistent with [5]. In all cases of spectral subtraction ($\alpha > 0$), ΔSRT was between 1.5 to 2.5 dB, reflecting a reduced speech intelligibility compared to the reference condition without spectral subtraction ($\alpha = 0$). Such decrease in intelligibility is consistent with data from [2]. The filled squares in Fig. 3 represent

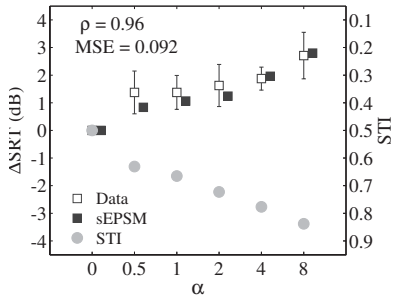

Figure 3: Measured (open symbols) and predicted (filled symbols) change in speech intelligibility with spectral subtraction.

the predicted results, showing an increase of ΔSRT by 1 to 2.5 dB which agrees very well with the measured data. In contrast, the corresponding STI is increased in all conditions of spectral subtraction, compared to the reference condition, thus predicting an increase in speech intelligibility. The STI thus fails to account for these data.

Discussion

The sEPSM followed the idea presented in [3], that a modulation noise floor can influence intelligibility. In contrast to [3], the proposed model (i) includes peripheral and modulation filtering inspired by the human auditory system, (ii) uses a statistically ideal observer as decision device and, in particular, (iii) hypothesizes that the reduction in speech intelligibility of processed noisy speech is caused mainly by the intrinsic noise fluctuations in the envelope of the noisy speech waveform, not by so-called "spurious" modulations, arising from the interaction between the speech and the noise waveforms. The critical distinction between all STI-based models and the sEPSM is that the STI considers the difference between the clean-speech and the noisy-speech envelope power, quantified by the modulation transfer function (MTF). In contrast, the sEPSM considers the difference between the noisy-speech envelope power and the intrinsic noise envelope power within the noisy speech. This means that the sEPSM is sensitive to distortions such as spectral subtraction, where the intrinsic noise envelope power is increased more than the speech envelope power. This is not captured by the STI because it neglects the intrinsic noise fluctuations.

Weighting of individual frequency bands is a common feature of speech intelligibility prediction metrics such as the standardized SII and STI. In contrast, the sEPSM does not apply any explicit weighting at all, apart from limitations in terms of absolute sensitivity. The agreement between the predicted and the measured intelligibility (cf. Fig. 2 and Fig. 3) therefore suggests that an explicit frequency weighting might not be necessary to account for the data if the metric that is assumed to be related to speech intelligibility is appropriate.

The sEPSM framework might also be useful in other conditions, such as low- and high-pass filtered noisy speech. Furthermore, the model can be applied to nonlinear conditions that have not been considered in the present study, such as amplitude compression, peak-clipping, phase-jitter and phase-shifts.

References

[1] Houtgast, T., Steeneken, M., Plomp, R.: Predicting speech intelligibility in rooms from the modulation transfer function. I. general room acoustics. J. Acoust. Soc. Am. 46 (1980), 60-72

[2] Ludvigsen, C., Elberling, C., and Keidser, G.: Evaluation of a noise reduction method-comparison between observed scores and scores predicted from sti, Scand. Audiol. Suppl. 38 22 (1993), 50-55.

[3] Dubbelboer, F. and Houtgast, T.: The concept of signal-to-noise ratio in the modulation domain and speech intelligibility, J. Acoust. Soc. Am. 124 (2008), 3937-3946.

[4] Ewert, S. and Dau, T.: Characterizing frequency selectivity for envelope fluctuations, J. Acoust. Soc. Am. 108 (2000), 1181-1196.

[5] Nielsen, J. B. and Dau, T.: Development of a danish speech intelligibility test, Int. J. Audiol. 48 (2009), 729-741.

[6] Wagener, K., Josvassen, J. L., and Ardenkjaer, R.: Design, optimization and evaluation of a danish sentence test in noise, Int. J. Audiol. 42 (2003), 10-17.

[7] Keidser, G.: Normative data in quiet and in noise for "dantale"-a danish speech material, Scand. Audiol. 22 (1993), 231-236.