Can physical metrics identify noise-reduction settings that optimize intelligibility?

Gaston Hilkhuysen & Mark Huckvale, CLEAR, University College London, London, UK

Introduction

Most noise reduction systems have parameters that need to be set for the best performance on any given combination of speech and noise. We have previously shown (Hilkhuysen *et al*, 2013) that the choice of settings can have large effects on intelligibility. Furthermore we showed that audio experts disagreed on their choice of best settings, and the settings chosen could even sometimes degrade intelligibility. We concluded that subjective impressions of intelligibility are not good enough to optimize the settings of a noise-reduction system.

On the other hand recent studies have reported high correlations between intelligibility scores obtained in listening experiments and intelligibility metrics based on the physical properties of the noise-reduced noisy signals (Table 1). In this paper we investigate whether these metrics could be used to optimize noise reduction for a particular signal.

Authors	*Conclusion*	*Evaluation*
Ludvigsen *et al.* (1993)	STI does not work	
Goldsworth & Greenberg (2004)	NCM is promising	
Ma & Loizou (2007)	$CSII_{mid}$ works best	r = 0.94; σ_{err}= 6%;
Taal *et al.* (2011a)	MCC works fine	r = 0.93; σ_{err}= 6%;
Taal *et al.* (2011b)	CSII and STOI work well	r = 0.92; σ_{err}= 6%; r = 0.92; σ_{err}= 8%;
Loizou & Ma (2011)	fAI works fine	r = 0.90; σ_{err}= 8%;
Jorgensen & Dau (2011)	sEPSM works fine	r = 0.99; σ_{err}= 0.5 dB;

Table 1. A short history of speech intelligibility metrics for noise-reduced noisy speech

Method

Hilkhuysen *et al.* (2013) measured the intelligibility of keywords in IEEE sentences for noise-reduced noisy speech in two noises using a commercial noise-reduction system. Effects of two noise-reduction parameters X ∈ {0, 13, 26, 39 dB} and Y ∈ {0, 33, 66, 99 %} were considered. The sixteen resulting combinations are denoted as set(X,Y). Figures 1a and 2a show contour plots of shifts in percentage word-correct relative to unprocessed noisy speech with car-cabin noise and babble, respectively. It can be seen that when both parameters were at nonzero values, processing deteriorated intelligibility. Thus the optimal intelligibility settings for the system was with X or Y set to zero.

Five metrics were evaluated:

1)The **Speech Intelligibility Index (SII)** (ANSI, 2008) is essentially based on a fractional signal-to-noise ratio in various auditory bands as in:

$$SII = \frac{1}{J}\sum_{j}^{J}\max\left(\min\left(\frac{SNR_j+15}{30},1\right),0\right) \qquad (1)$$

where SNR_j indicates the long-term average signal-to-noise ratio in the j^{th} auditory channel. To obtain these levels, speech with and without a 180 degree phase shift was added to the noise and both signals were processed by the commercial noise reduction system. The sum and difference of the two output signals were used to estimate the noise-reduced noise and speech, respectively (Hagerman & Olofsson, 2004).

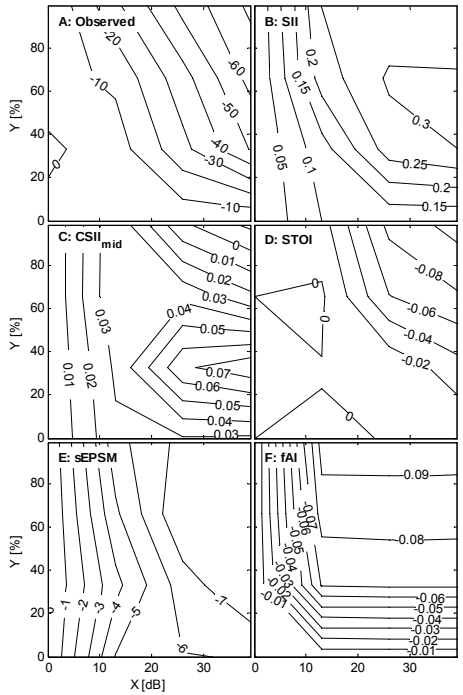

Figure 1. Contour plots representing shifts in (A) percentage word score and (B-F) values of intelligibility metrics as a function of combinations of two noise-reduction parameters settings for speech in car-cabin noise (SNR=-12dB).

2) The **Coherence Speech Intelligibility Index** based on mid levels of speech ($CSII_{mid}$) only includes segments with broadband levels of intact speech between 0 and –10 dB re: RMS (Kates & Arehart, 2005). SNR_j in Eq(1) is replaced by a signal distortion ratio (SDR_j). Distortion is quantified by correlating the complex spectra of the intact and the distorted speech across time frames as in:

$$|\gamma(i)|^2 = \frac{\left|\sum_k X_k(i)Y_k^*(i)\right|^2}{\sum_k |X_k(i)|^2 \sum_k |Y_k(i)|^2} \qquad (2)$$

where $|\gamma(i)|^2$ indicates the coherence, $X_k(i)$ and $Y_k(i)$ are the spectra of original and distorted speech time frames, * denotes the complex conjugate, *i* the frequency bin, and *k* the time frame. SDR_j is calculated by integrating $|\gamma(i)|^2$ across frequency bins as in:

$$SDR_j = \frac{\sum_i W_j(i)|\gamma(i)|^2 S_{yy}(i)}{\sum_i W_j(i)[1-|\gamma(i)|^2]S_{yy}(i)} \qquad (3)$$

with $S_{yy}(i)$ representing the power spectral density of the distorted signal. $W_j(i)$ denotes the weight of frequency bin *i* in auditory channel *j*.

3) The **Short-Time Objective Intelligibility metric** (STOI) (Taal et al., 2011) considers the spectrograms of the intact speech and the distorted speech. Pixels have sizes of 1 ERB by 13 ms and express levels in log powers. Temporal sequences of 30 adjacent pixels are correlated across signals, with the restriction that the pixel level of the distorted speech deviates less than -15 dB from the pixel level in corresponding intact speech. Lower pixel levels are raised to -15 dB. STOI is the average of all possible correlations.

4) The **speech-based Envelope Power Spectrum Model (sEPSM)** (Jorgensen & Dau, 2011) considers the power of the amplitude modulations at the output of a 22-channel cochlear filterbank feeding a 7-channel modulation filterbank. Let $Psn_{j,m}$ denote these powers for the noise-reduced speech plus noise signal in the j^{th} auditory channel and m^{th} modulation channel. The powers of the noise-reduced noise are expressed by $Pn_{j,m}$. The signal-to-noise ratio in the envelope domain is given by:

$$envSNR_{j,m} = \frac{\max(Psn_{j,m} - Pn_{j,m}, 10^{-30})}{Pn_{j,m}} \qquad (4)$$

sEPSM is defined as the RMS of $envSNR_{j,m}$.

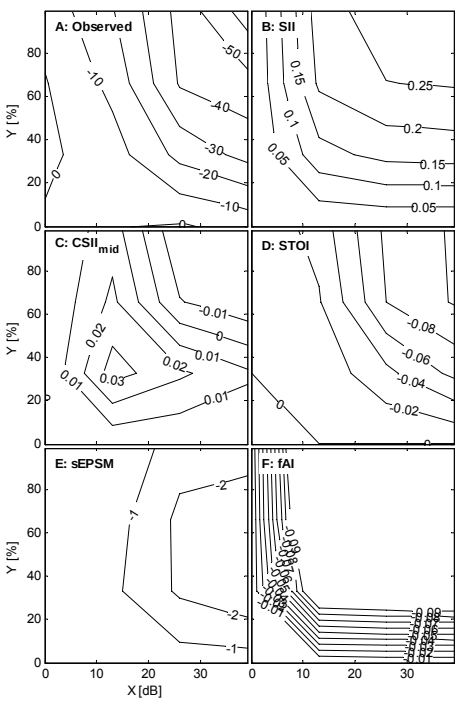

Figure 2. Same as Figure 1 but for speech in babble noise (SNR=-3dB).

5) The **Fractional Articulation Index (fAI)** (Loizou & Ma, 2011) divides signals into *j* auditory channels and *k* time frames. A spectral-temporal bin only contributes to intelligibility when its local SNR ($SNR_{j,k}$) exceeds 11 dB. If the noise reduction misestimates this SNR, intelligibility is compromised as in:

$$fSNR_{j,k} = \frac{\min(SNR_{j,k}, \hat{SNR}_{j,k})}{SNR_{j,k}} \qquad (5)$$

Here $SNR_{j,k}$ and $\hat{SNR}_{j,k}$ indicate the factual and estimated ratio of signal and noise powers in auditory channel *j* and time frame *k*. $fSNR_{j,k}$ is the fractional signal-to-noise ratio, which is zero when $SNR_{j,k}$ is less than 11 dB. $fSNR_{j,k}$ averaged across all auditory channels and time frames defines fAI.

Results

- **SII** incorrectly predicts improvements in intelligibility and is highest with set(39,66) and set(39,99) for speech in car-cabin noise and babble, respectively.
- **$CSII_{mid}$** incorrectly predicts maximal intelligibility results from noise suppression with set(39,33) and set(13, 33).
- **STOI** incorrectly predicts improvements for speech in car-cabin noise. However the metric correctly suggests that either X or Y should be kept at zero for speech in babble. It is the only metric with contour plots resembling the observed intelligibility shift.
- **sEPSM** correctly predicts no intelligibility improvements and suggests X should be kept at zero, however it is insensitive to changes in Y for speech in car-cabin noise.
- **fAI** is the only metric to identify that parameter combinations which include either X or Y set to zero do not deteriorate intelligibility and are optimal.

Discussion

Of the five metrics considered only fAI identified the optimal settings for both car-cabin and babble. It should however be noted that fAI can never predict intelligibility improvements, despite these having been observed (Tsoukalas *et al*.,1997; Arehart *et al*., 2003). STOI and sEPSM performed well with babble, but failed in car-cabin noise. Nevertheless STOI appears to be the best estimator for intelligibility shifts induced by noise reduction. In both noises SII and $CSII_{mid}$ predicted improvements that were not observed.

Overall the high correlations reported in Table 1 contrast with the poor predictions visible in Figures 1 and 2. The fact that the current study involved only one signal-to-noise ratio per noise type may account for this discrepancy.

Acknowledgements

This work was undertaken when the first author was at the Centre for Law Enforcement Audio Research, a joint research centre hosted at Imperial College London and UCL funded by the United Kingdom Home Office. Visit www.clear-labs.com for publications.