SIGNAL PROPERTIES REDUCING INTELLIGIBILITY OF SPEECH AFTER NOISE REDUCTION

Gaston Hilkhuysen and Mark Huckvale

Centre for Law Enforcement Audio Research (CLEAR)
Speech, Hearing & Phonetic Sciences,
University College London, UK
E-mail: {g.hilkhuysen, m.huckvale}@ucl.ac.uk

ABSTRACT

The effect of noise reduction on the intelligibility of speech in noise is poorly understood. Although the SNR of noisy speech is improved by the removal of more noise than speech from the signal, the expected increase in intelligibility does not typically occur. To account for these deleterious effects we present an orthogonal decomposition of the signal intensity envelopes at the output of a filterbank. The noisy speech envelopes are decomposed into components indicating (1) the coherence of speech across audio bands; (2) the distortion of the speech envelope; and (3) the speechiness of the noise. By modelling the results of a listening experiment we show that envelope distortion can largely account for the deleterious effects of noise reduction; although reduced coherence could also play a role at low SNRs. There was little evidence for the idea that increased speechiness of the noise contributed to the poorer intelligibility after noise reduction.

1. INTRODUCTION

At the output of a filterbank, speech is observed to show considerable envelope modulations which appear to distinguish it from other everyday sounds. It has been long supposed that these modulations are essential to intelligibility; and channels that correctly transmit intensity modulations of speech give high intelligibility [1]. These modulations have also been widely used in physical measures that predict speech intelligibility from the signal. For example, the Speech Intelligibility Index (SII) is calculated from the signal-to-noise ratios in multiple audiobands over 30 dB ranges located symmetrically around the RMS of speech [2]. Likewise the calculation of the Speech Transmission Index (STI) is based on changes in the modulation depths per audio band [3]. Experience shows that both SII and STI provide good estimates of intelligibility after linear processing, such as reverberation, filtering and additive noise, but worse estimates after non-linear transforms, such as dynamic range compression [4] or noise suppression [5].

To account for the deleterious effects of multiband dynamic range compression on intelligibility, Stone and Moore [6] distinguish three possible mechanisms:

(1) Speech modulations within an audio band may be degraded by the processing. These would affect the modulations considered in intelligibility models like the SII and STI. However, while those models are particularly concerned with masking of low intensity speech modulations, Stone and Moore suggest that there may also be changes in the envelopes at high-levels which might be made clear by correlating envelopes before and after processing.

(2) The coherence of speech modulations across audio bands may be affected. This mechanism focuses on the co-modulations in speech envelopes across audio bands. In the SII, the contribution of each audio band to intelligibility is assumed to be independent from all other bands, although it has been acknowledged that speech modulations between channels are correlated [7]. Whereas in the latest implementation of STI [3] these co-modulations are interpreted as adding to the redundancy of speech, others [6, 7, 8] have argued that they may help to perceptually separate the speech from the noise, a phenomenon known as auditory grouping [9]. According to this latter view, a process that impairs coherence in speech modulations would reduce intelligibility.

(3) The noise may obtain a speech-like character. This mechanism also concerns the perceptual separation of speech from noise. Noise with modulations similar to speech will be more difficult to distinguish from the speech, presumably leading to reduced intelligibility. The mechanism could account for the greater impact on intelligibility of multi-speaker babble compared to Gaussian-noise with the same long-term spectrum. In [10] it was reported that the intelligibility of high-pass filtered speech was disturbed by the introduction of speech-like modulations in an off-frequency low-pass filtered masking noise. Both phenomena suggest that ”speechy” noise deteriorates intelligibility.

A previous study [6] addressed the contributions of the three mechanisms to the deleterious effects of dynamic range compression. However, the decomposition of the modulations in that study was not orthogonal. This meant that a similar distortion could be attributed to different mechanisms, which obstructed a clear view on their relative magnitude. To try to obtain a better understanding of speech intelligibility after noise suppression, we investigate which of these three mechanisms play a role in the detrimental effects of spectral subtraction, using an orthogonal decomposition of noisy speech intensity envelopes.

2. ENVELOPE DECOMPOSITION

The DC component of the intensity envelope represents the average level in an audio band. Here we assume that levels in all audio bands are well above hearing threshold, meaning that the DC holds no consequences for intelligibility. Therefore the DC components were excluded from all calculations to be presented. The clean-speech intensity envelope $s_i(t)$ in audio band i can be divided into a unique modulation and a modulation shared with other frequency bands according to:

$$s_i(t) = s_i^u(t) + s_i^s(t), \tag{1}$$

where the superscripts u and s denote the unique and the shared modulations, respectively. Shared modulations are found with regression on the modulations in band i by the modulations in all other bands:

$$s_i^s(t) = \hat{s}_i(t) = \sum_j \beta_j s_j(t), \quad j \neq i. \tag{2}$$

Since all DC components were ignored, the regression in (2) does not include a constant. The unique clean speech modulations are collected in the error term of this regression. Since both unique and shared modulations are signals, their relative levels can be expressed as:

$$\mathrm{USSR}_i = 10\log_{10}\left(\frac{\sum_t [s_i^u(t)]^2}{\sum_t [s_i^s(t)]^2}\right), \tag{3}$$

where USSR stands for the unique to shared speech modulation ratio. It expresses the coherence of the speech envelope in a particular band with the speech envelopes in other auditory bands, similar to within source modulation coherence [6] A low USSR value indicates a strong coherence of the band with other auditory bands.

Mixing speech with noise, possibly followed by non-linear processing such as dynamic range compression or noise reduction may add noise to the clean speech modulations, which can be written

$$ns_i(t) = w_u s_i^u(t) + w_s s_i^s(t) + n_i(t), \tag{4}$$

with $ns_i(t)$ representing the noisy speech modulations and $n_i(t)$ representing the noise modulation. Both weights in (4) may be determined from the correlations between the unique and shared modulations of clean speech with the noisy speech envelope. For additive noise, one would expect equal effects on both components, hence stable USSR. However, non-linear processing could result in different effects for unique and shared speech envelope components and consequently USSR would change.

We previously labelled the ratio between the speech and noise modulations as the signal-to-noise ratio in the modulation domain, which is closely related to fidelity to envelope shape [6]. It is given by:

$$\mathrm{SNR}_i^{\mathrm{mod}} = 10\log_{10}\left(\frac{\sum_t [w_u s_i^u(t) + w_s s_i^s(t)]^2}{\sum_t [n_i(t)]^2}\right). \tag{5}$$

In the decomposition of the noise modulations, one quantifies the speech-like character of the noise:

$$n_i(t) = n_i^e(t) + n_i^m(t), \tag{6}$$

where the e and m superscripts indicate the modulations exclusive and mutual with speech, hence:

$$n_i^m(t) = \hat{n}_i(t) = \sum_j \lambda_j s_j(t), \quad j \neq i. \tag{7}$$

The noise modulations that a particular band has in common with speech in other audio bands are found by linear regression: the noise envelope in an audio band is predicted by the clean speech envelopes in all other bands. Since both exclusive and mutual noise modulations are signals, one can express their relative ratios as

$$\mathrm{EMNR}_i = 10\log_{10}\left(\frac{\sum_t [n_i^e(t)]^2}{\sum_t [n_i^m(t)]^2}\right), \tag{8}$$

defining the exclusive to mutual noise modulation ratio (EMNR). A low EMNR indicates noise with a high speech-like character.

Combining equations (4) and (6) we obtain:

$$ns_i(t) = w_u s_i^u(t) + w_s s_i^s(t) + n_i^e(t) + n_i^m(t), \tag{9}$$

which is an orthogonal decomposition of the noisy speech envelope. Since the four components in (9) have zero correlations, their effects on speech intelligibility can be studied independently, in contrast to the mechanisms presented in [6].

3. EXPERIMENT

Envelopes were generated for 17 adjacent 1/3-octave audio bands with band-centre frequencies ranging from 0.160 to 6.35 kHz covering the audio frequencies that contribute most to intelligibility [2, 3]. The filter bank consisted of zero-phase hamming-windowed sync filters with complementary skirts and over 60 dB oct^{-1} slopes. After band-pass filtering, intensity envelopes were extracted by squaring the magnitude of the Hilbert transform and subsequent limiting the modulations to 6 octave bands with centre frequencies ranging from 1 to 32 Hz, while applying a -3 dB oct^{-1} slope on the 6-octave wide band pass. This pinking of the envelope gives rise to a log-frequency weighting of the modulation frequencies as advocated by Dau *et al.* [11], instead of the linear weighting obtained without this envelope colouration. USSR, SNR$_i^{\mathrm{mod}}$ and EMNR were calculated using a fixed 128 s fragment of concatenated IEEE sentences [12, 13] combined with a fixed car noise fragment of equivalent duration. Speech was mixed to the noise at five different levels, ranging from -21 to -9 dB SNR in 3 dB steps. To study the effects of noise suppression on signal ratios, calculations were performed on noisy speech before and after noise reduction by spectral subtraction. The spectral subtraction implementation [14] in VOICEBOX [15], which uses the minimum statistics method [16] to estimate the noise spectrum, was utilised.

Intelligibility was addressed by presenting IEEE sentences to a group of 20 listeners at the 5 SNR levels previously mentioned for speech in car noise with and without noise suppression, while scoring the number of correct keywords in their responses.

4. RESULTS

Figure 1 shows USSR as a function of SNR in the audio band with a centre frequency of 0.63 kHz. Plots of these functions in other audio bands exhibit similar behaviour. Open en filled markers represent USSR levels before and after noise suppression, respectively. From Fig. 1, it emerges that before noise suppression the unique and shared speech modulations are deteriorated equally at different SNRs: i.e. additive noise equally affects both components of the speech envelope. This does not hold for speech envelopes after noise suppression. In this case USSR is higher, especially at lower SNRs. In other words, noise suppression appears to deteriorate the coherence of speech.

Figure 2 shows $\mathrm{SNR}_i^{\mathrm{mod}}$ as a function of SNR of SNR, again for the audio band with a centre frequency of 0.63 kHz. Marker styles conform to Fig. 1. At low SNRs noise reduction removes envelope distortions introduced by the noise, hence $\mathrm{SNR}_i^{\mathrm{mod}}$ increases after noise reduction. At high SNRs the inverse holds: noise reduction gives rise to additional distortion of speech envelope, already distorted by the noise.

Figure 3 visualizes the speechiness of the noise expressed in EMNR once again for the audio band with a centre frequency of 0.63 kHz. Marker styles denote processing conditions equivalent to Fig. 1. While applying noise reduction, EMNR reduces, indicating that the fluctuation in the noise obtain a speech-like character. In contrast to USSR and $\mathrm{SNR}_i^{\mathrm{mod}}$, the effect of noise reduction on EMNR varies largely across audio bands.

Figure 4 shows the results of the listening experiment with word scores represented on the ordinate. In contrast to most other studies where results are expressed in percentages of words correct, we prefer to display $\log_2$ odds, where odds are the ratio between the number of correct and incorrect responses. We labelled this quantity performance level with units in Berkson. For psychometric functions that follow a logistically shaped curve, this scale gives rise to a more linear relationship than the traditional percentage scale. The right axis reflects percentage values corresponding to the performance levels indicated on the left axis. Curves displayed in Fig. 4 are known as performance functions. The abscissa indicates intelligibility as predicted from two speech intelligibility models. Open circles represent the values from traditional SII calculations based on the SNR in all audio bands. To calculate the SNR after noise reduction, displayed as filled circles, the attenuation factors were determined on the basis of the noisy speech signal, while these factors were applied to the corresponding frames of the speech and noise separately. The long-term average RMS levels of these speech and noise signals determined the SNRs in the audio bands, which were subjected to an SII calculation. Squares indicate the predicted intelligibility based on a measure called $\mathrm{SII}^{\mathrm{mod}}$. Figure 2 shows that the $\mathrm{SNR}_i^{\mathrm{mod}}$ is monotonically related to SNR. Consequently the $\mathrm{SNR}_i^{\mathrm{mod}}$ after noise reduction can be expressed as an “equivalent SNR” in a particular audio band before noise reduction. The latter is the SNR before noise reduction that gives rise to a similar amount of envelope distortion as present in the signal after noise reduction. These equivalent SNRs were used in subsequent SII calculation, leading to the $\mathrm{SII}^{\mathrm{mod}}$. For speech before noise suppression $\mathrm{SII}^{\mathrm{mod}}$ equals SII, hence is not displayed in Fig. 4.

Given an optimal intelligibility model, performance functions for speech in noise with and without noise suppression should coincide. In that case there exists a one-to-one relation between the predicted and observed intelligibility, independently of the presence of noise suppression. For predictions based on SII, this is evidently not the case. Noise reduction removed more energy from the noise than from the speech, hence increased the apparent SNR. This resulted in an increase in SII, leading to the prediction that the intelligibility after noise reduction should increase, while the observed intelligibility dropped. For predictions based on $\mathrm{SII}^{\mathrm{mod}}$, the one-to-one relationship between observed and predicted intelligibility appears to hold at high performance levels, where SNR is high. However at low performance levels $\mathrm{SII}^{\mathrm{mod}}$ still overestimate intelligibilities.

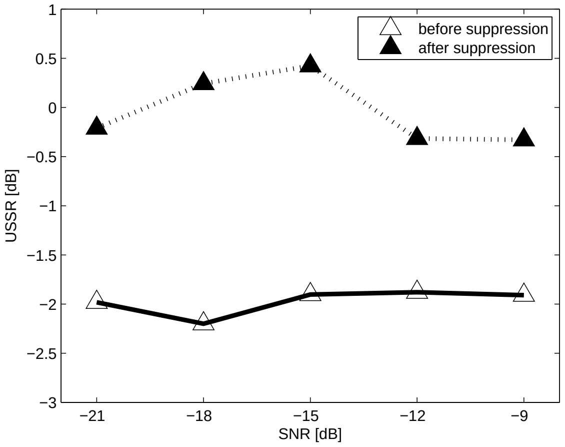

Figure 1: USSR as a function of SNR.

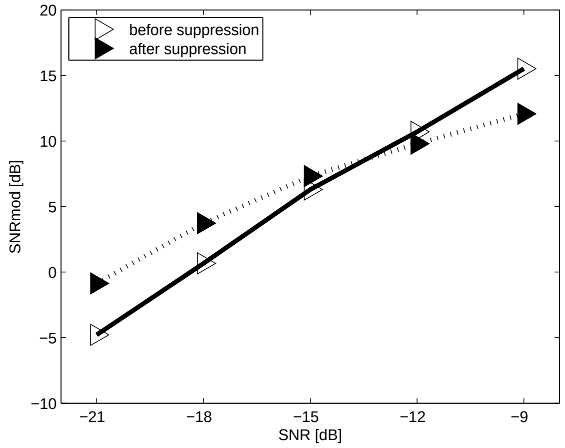

Figure 2: $\mathrm{SNR}^{\mathrm{mod}}$ as a function of SNR.

5. DISCUSSION AND CONCLUSIONS

From the performance functions based on the SII as displayed in Fig. 4, one may conclude that noise suppression successfully removes more energy from the noise than from the speech, resulting in a higher SNR per audio band than before noise reduction, and consequently higher SII values. But unfortunately, these higher SNRs do not result in improved intelligibility, in contrast to what one might predict from SII calculations: the intelligibility of speech in noise after noise reduction is poorly predicted from the long-term average levels of speech and noise.

Figures 2 and 4 suggest that the deleterious effects of noise suppression on speech intelligibility are largely induced by distortion of the speech envelope. It may be that envelope distortion results in impoverished consonant identi-

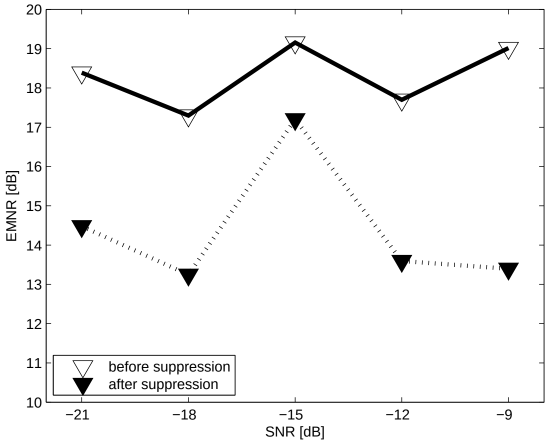

Figure 3: EMNR as a function of SNR.

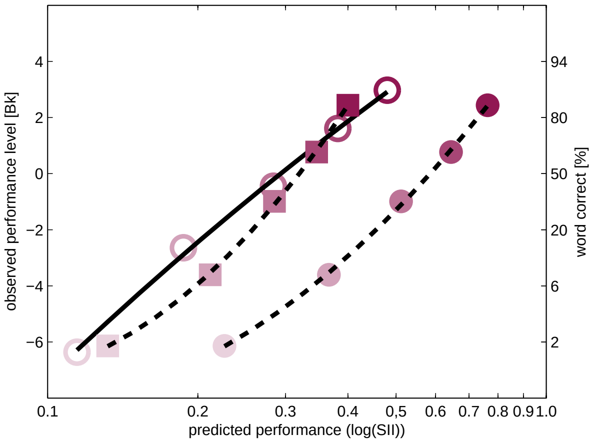

Figure 4: Observed versus modelled intelligibilities.

fication, particularly important to the intelligibility of speech in noise [17, 18, 19]. However, based on the SII$^{\mathrm{mod}}$ measure, one would predict an increase in intelligibility at low SNRs - where noise suppression increases SNR$_i^{\mathrm{mod}}$ - while the observed performance is in fact reduced.

To account for this limitations of SII$^{\mathrm{mod}}$, we introduced the USSR and EMNR measures. It appears that, at low SNRs, noise reduction diminishes the coherence in speech, giving rise to higher USSR values. This could account for the fact that even at low SNRs, intelligibility is still deteriorated by noise suppression, notwithstanding an increase in SNR$_i^{\mathrm{mod}}$.

On the other hand, the fact that the noise obtains a speech-like character after noise reduction at high SNRs, appears to have little effect on the intelligibility. If the speechiness of the noise had deleterious effects on intelligibility, one would expect a misfit in the performance functions before and after noise suppression at low and high SNRs, where the noise after noise suppression has low EMNRs, indicating a speech like character.

Unfortunately, the current data only allow for a qualitative examination of the effects of USSR and EMNR on intelligibility. But following (9), it is possible to manipulate the different components independently, which may contribute to the development of an intelligibility model that better accounts for the effects of noise suppression on intelligibility. Such a model could contribute to the development of noise suppression algorithms that improve intelligibility, since noise suppressor design and subsequent parameter adjustment could be optimized for a given signal without the need for time-consuming listening experiments.

REFERENCES

[1] R. V. Shannon, F. G. Zeng, V. Kamath, J. Wygonski, and M. Ekelid, “Speech recognition with primarily temporal cues,” *Science*, vol. 270, pp. 303–304, 1995.

[2] *Methods for the Calculation of the Speech Intelligibility Index*, ANSI Std. S3.5–1997, Rev. R2007.

[3] *Sound system equipment - Part 16: Objective rating of speech intelligibility by speech transmission index*, EIC Std. 60 268-16:2003.

[4] I. M. Noordhoek and R. Drullman, “Effect of reducing temporal intensity modulations on sentence intelligibility,” *J. Acoust. Soc. Am.*, vol. 101, pp. 498–502, 1997.

[5] C. Ludvigsen, C. Elberling, and G. Keidser, “Evaluation of noise reduction method: Comparison between observed scores and scores predicted from STI,” *Scan. Audiology*, vol. 39, pp. 50–55, 1993.

[6] M. A. Stone and B. C. Moore, “Quantifying the effects of fast-acting compression on the envelope of speech,” *J. Acoust. Soc. Am.*, vol. 121, pp. 1654–1664, 2007.

[7] O. Crouzet and W. A. Ainsworth, “On the various influences of envelope information on the perception of speech in adverse conditions: An analysis of between-channel envelope correlation,” in *Workshop on Consistent and Reliable Cues for Sound Analysis*, Aalborg, Denmark, Sep. 2001.

[8] J. B. Allen, “How do humans process and recognize speech?” *IEEE Trans. Speech Audio Process.*, vol. 2, pp. 567–577, 1994.

[9] A. S. Bregman, *Auditory Scene Analysis, The Perceptual Organization of Sound.* MA:MIT Press, 1990.

[10] F. Apoux and S. P. Bacon, “Modeling auditory processing of amplitude modulation.” *J. Acoust. Soc. Am.*, vol. 123, pp. 1665–1672, 2008.

[11] T. Dau, B. Kollmeier, and A. Kohlrausch, “Modeling auditory processing of amplitude modulation. i. detection and masking with narrow-band carriers,” *J. Acoust. Soc. Am.*, vol. 102, pp. 2892–2905, 1997.

[12] E. H. Rothauser, W. D. Chapman, N. Guttman, M. H. L. H. K. S. Sordby, H. R. Silbiger, G. E. Urbanek, and M. Weinstock, “IEEE recommended practice for speech quality measurements,” *IEEE Trans. Audio Electroacoust.*, vol. 17, pp. 225–246, 1969.

[13] M. W. Smith and A. Faulkner, “Perceptual adaptation by normally hearing listeners to a simulated “hole” in hearing,” *J. Acoust. Soc. Am.*, vol. 120, pp. 4019–4030, 2006.

[14] M. Berouti, R. Schwartz, and J. Makhoul, “Enhance-

ment of speech corrupted by acoustic noise," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 4, 1979, pp. 208–211.

[15] D. M. Brookes, "VOICEBOX: A speech processing toolbox for MATLAB," 1997. [Online]. Available: http://www.ee.imperial.ac.uk/hp/staff/dmb/voicebox/voicebox.html

[16] R. Martin, "Noise power spectral density estimation based on optimal smoothing and minimum statistics," *IEEE Trans. Speech Audio Process.*, vol. 9, pp. 504–512, Jul. 2001.

[17] L. Xu, C. S. Thompson, and B. E. Pfingst, "Relative contributions of spectral and temporal cues for phoneme recognition," *J. Acoust. Soc. Am.*, vol. 117, pp. 3255–3267, 2005.

[18] L. Xu and Y. Zheng, "Spectral and temporal cues for phoneme recognition in noise," *J. Acoust. Soc. Am.*, vol. 122, pp. 3255–3267, 2007.

[19] E. Buss, L. N. Whittle, J. H. Grose, and J. W. H. III, "Masking release for words in amplitude-modulated noise as a function of modulation rate and task," *J. Acoust. Soc. Am.*, vol. 126, pp. 269–280, 2009.