

A SIMPLE CORRELATION-BASED MODEL OF INTELLIGIBILITY FOR NONLINEAR SPEECH ENHANCEMENT AND SEPARATION

Jesper B. Boldt¹, Daniel P. W. Ellis²

¹Oticon A/S, Kongebakken 9, DK-2765 Smørum, Denmark

²LabROSA, Dept. of Electrical Engineering, Columbia University
New York, NY 10027, USA
jeb@oticon.dk, dpwe@ee.columbia.edu

ABSTRACT

Applying a binary mask to a pure noise signal can result in speech that is highly intelligible, despite the absence of any of the target speech signal. Therefore, to estimate the intelligibility benefit of highly nonlinear speech enhancement techniques, we contend that SNR is not useful; instead we propose a measure based on the similarity between the time-varying spectral envelopes of target speech and system output, as measured by correlation. As with previous correlation-based intelligibility measures, our system can broadly match subjective intelligibility for a range of enhanced signals. Our system, however, is notably simpler and we explain the practical motivation behind each stage. This measure, freely available as a small Matlab implementation, can provide a more meaningful evaluation measure for nonlinear speech enhancement systems, as well as providing a transparent objective function for the optimization of such systems.

1. INTRODUCTION

Speech enhancement concerns taking a target speech signal that has been corrupted, by the addition of interfering sources and transmission through an acoustic channel, and mitigating the impact of these corruptions. Enhancement can have two, distinct goals: improving *quality*, which relates to how “clear” or “natural” the enhanced speech sounds, and improving *intelligibility*, which focuses on the more practical problem of whether a listener can understand the message in the original target speech. Although we might expect that quality and intelligibility are strongly correlated, there are ample situations in which speech of relatively low quality can nonetheless achieve high intelligibility [17, 22], and where improving quality does not necessarily improve intelligibility [10].

In this paper we ignore quality (and related effects such as listener fatigue) and concentrate on intelligibility. We focus specifically on time-frequency masking algorithms, which have been widely used in automatic speech recognition [6], computational auditory scene analysis [21], noise reduction [15, 2], and source separation [23, 16]. In this type of algorithm, a time-varying and frequency-dependent gain is applied across a number of frequency channels. In some variants, the gains are quantized to zero or one, giving a *binary masking* algorithm where the pattern of gains is referred to as the binary mask. One type of binary mask – the ideal binary mask (IBM) – has shown to be able to increase speech intelligibility significantly [3, 2, 14]. This mask is ‘ideal’ in that it relies on perfect knowledge of both clean target and interference prior to mixing, and is constructed to pass only those time-frequency cells in which the target energy exceeds the interference. An intriguing property of the IBM is that applying such a mask to a sound consisting only of noise results in high intelligibility for the speech upon which the mask was based [22, 13], even though the perceived

quality of the reconstructed speech is very poor: depending on the resolution of the time-frequency distribution, it will have no pitch or other fine structure, and fine nuances of energy modulation are lost. Similar characteristics are found to those of noise-excited channel-vocoded speech [17]. An attempt to measure the signal to noise ratio (SNR) in such signals would find no trace of the original target in the final output, so SNR-based measures will not be a useful basis for accounting for this intelligibility. What is preserved, however, is the broad envelope in time and frequency. This suggests that an intelligibility estimate could be developed based on the similarity of this envelope between target speech and system output.

In this paper, we use correlation as a measure of similarity between time-frequency envelopes of target and enhanced speech. Given this basic principle, we make a number of design choices and system enhancements with a view to matching the general properties of observed subjective intelligibility of nonlinearly-enhanced signals. At each stage, we strive for the simplest and most transparent processing that can effectively match the subjective results. Our outcome is a simple correlation-based measure that can predict intelligibility with approximately the same fidelity as more complex models based on far more detailed models of auditory processing [4]. We feel this simplicity and transparency is a considerable advantage as a guide for developing enhancement systems.

2. NORMALIZED SUBBAND ENVELOPE CORRELATION

To estimate intelligibility, the correlation between the time-frequency representations of the target (reference) and the output of the time-frequency masking algorithm is calculated:

$$\sum_{\tau} \sum_k \mathbf{T}(\tau, k) \cdot \mathbf{Y}(\tau, k), \quad (1)$$

where τ the time index, k the frequency index, $\mathbf{T}(\tau, k)$ is the energy envelope of the target signal, and $\mathbf{Y}(\tau, k)$ is the energy envelope of the output. This correlation will not have an upper bound, and in low energy regions of $\mathbf{T}(\tau, k)$ the inclusion of potential unwanted energy in $\mathbf{Y}(\tau, k)$ will have a very small impact on the correlation. To improve this behavior, we normalize with the Frobenius norm of $\mathbf{T}(\tau, k)$ and $\mathbf{Y}(\tau, k)$ and refer to this measure as the normalized subband envelope correlation (**nSec**):

$$\mathbf{nSec} = \sum_{\tau} \sum_k \frac{\mathbf{T}(\tau, k) \cdot \mathbf{Y}(\tau, k)}{\|\mathbf{T}(\tau, k)\| \|\mathbf{Y}(\tau, k)\|} \quad (2)$$

The **nSec** is bounded between zero and one. The lower bound is reached if no energy is found in the same regions of $\mathbf{T}(\tau, k)$ and $\mathbf{Y}(\tau, k)$. The upper bound is reached if the two signals are identical or only differ by a scale factor. Geometrically interpreted, **nSec** is the angle between $\mathbf{T}(\tau, k)$ and $\mathbf{Y}(\tau, k)$ if calculated using a single time or frequency index.

3. EXPERIMENTAL DATA

To verify that **nSec** is a useful measure of speech intelligibility, we use the results from Kjems et al. [13], where speech intelligibility of

This work was supported by Oticon A/S, The Danish Agency for Science, Technology and Innovation, and by the NSF under grant no. IIS-0535168. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the sponsors.

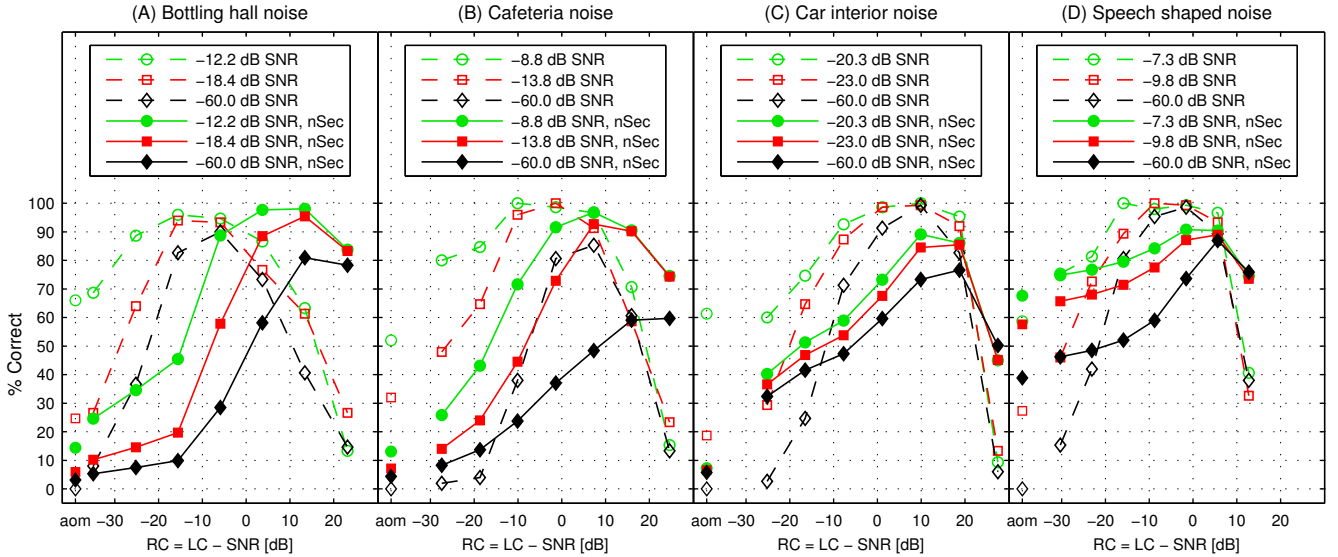


Figure 1: Estimated intelligibility by **nSec** compared to subjective listening tests in four different noise conditions and three SNR levels. **nSec** is shown with solid lines/filled symbols, and subjective listening tests are shown with dotted lines/hollow symbols. The results are plotted as a function of the RC value which determines the sparseness of the binary mask - higher RC values imply fewer ones in the binary mask. The all-one mask (aom) is the unprocessed condition and does not correspond to a specific RC value.

IBM-masked noisy speech is measured using normal hearing subjects, three SNR levels, four noise types, and different local SNR criteria (LC). LC is the threshold used to construct the IBM; a larger LC results in an IBM with proportionally fewer nonzero values:

$$\text{IBM}(\tau, k) = \begin{cases} 1, & \text{if } \frac{\mathbf{T}(\tau, k)}{\mathbf{N}(\tau, k)} > \text{LC} \\ 0, & \text{otherwise} \end{cases}, \quad (3)$$

where $\mathbf{N}(\tau, k)$ is the energy envelope of the noise signal.

Two of the three SNR levels used in the experiments by Kjems et al. were set to 20% and 50% intelligibility of the speech and noise mixtures with no binary masking. The third SNR level was fixed at -60 dB to examine the effect of applying the IBM to pure noise. Four different noise conditions were used: speech shaped noise (SSN), cafeteria noise, car interior noise with mainly low-frequency energy, and noise from a bottling hall with mainly high-frequency energy. The LC values resulted in IBMs consisting of between 1.5% and 80% of nonzero cells, and an all-one mask (aom) was used to measure the intelligibility of the unprocessed mixture with no binary masking. A 64 channel Gammatone filterbank with centerfrequencies from 55 Hz to 7742 Hz equally spaced on the ERB (equivalent rectangular bandwidth) scale was used, and the output was divided into 20 ms frames with 10 ms overlap. The results are shown with dotted lines and hollow symbols in Figure 1 (and are repeated in subsequent figures). To align the results, they are plotted as a function of the RC value defined as $\text{RC} = \text{LC} - \text{SNR}$ in units of dB. Using this x -coordinate, the binary masks will be identical at the same RC value and independent of the SNR levels.

To compare the **nSec** with the results by Kjems et al., we use 10 sentences from their experiment which have been mixed with noise and processed with the IBM. Silence between the sentences are removed from the waveforms, and $\mathbf{T}(\tau, k)$ and $\mathbf{Y}(\tau, k)$ are calculated using a 16 channel Gammatone filterbank with center frequencies from 80 Hz to 8000 Hz equally spaced on the ERB scale. The energy from each frequency channel in the filterbank is divided into segments of 80 ms with 40 ms overlap. All processing is done at 20 kHz. The calculated time-frequency representations $\mathbf{T}(\tau, k)$ and $\mathbf{Y}(\tau, k)$ are inserted in Equation 2, and the **nSec** scaled by a factor of 100 is shown with solid lines and filled symbols in Figure 1.

4. MODIFICATIONS TO THE **nSec**

Looking at Figure 1, it can be seen that even though the **nSec** is not aligned with the subjective listening tests, the overall shape and behavior is encouraging: Increasing SNR gives a better or similar **nSec**, and a distinct peak in correlation as a function of RC value is seen at all curves except for the -60 dB SNR cafeteria noise (Fig.1.B). If this curve had been continued to higher RC values, it would have made a peak at some point, because higher RC values makes the binary mask more sparse with fewer ones, and, ultimately, $\mathbf{Y}(\tau, k)$ will be zero. At the other extreme, at low RC values, the **nSec** levels off which is most evident from Figure 1.A and 1.D. The reason is that at some RC value, the time-frequency units added to $\mathbf{Y}(\tau, k)$ by lowering the RC value will not change the numerator of Equation 2 because no energy is found at these time-frequency units in $\mathbf{T}(\tau, k)$. At the same time, the denominator will continue to increase as the RC value decreases, due only to the added energy in $\|\mathbf{Y}(\tau, k)\|$; $\|\mathbf{T}(\tau, k)\|$ is a fixed value independent of SNR and RC value.

Comparing the three SNR levels, it can be seen that the peak of the **nSec** shifts towards lower RC values for higher SNRs – a reasonable property, if we recognize that the IBM for a certain target and noise sound is a function of the RC value only, and that increasing SNR level implies that the RC value can be lowered without increasing the number of noise-dominated time-frequency units in the binary masked mixture. At increasing SNR levels, the RC value is lowered by increasing the LC value with less than the increase in SNR level.

The **nSec** for the speech shaped noise (Fig.1.D) with an all-one mask is considerable higher at all three SNR levels compared to other noise types. The **nSec** of the -60 dB SNR mixture with an all-one mask is approximately 0.4, despite the fact that practically no target sound is found in the mixture. Two random signals always will give a positive correlation as long as they contain energy in some of the same time-frequency regions, and the speech shaped noise do, since it was made by superimposing 30 sequences of the speech from the corpus with random silence durations and starting times [20]

The last observation we make of the unmodified **nSec** is that the location of the peaks are at higher RC values compared to the subjective listening tests. This property is caused by the fact that

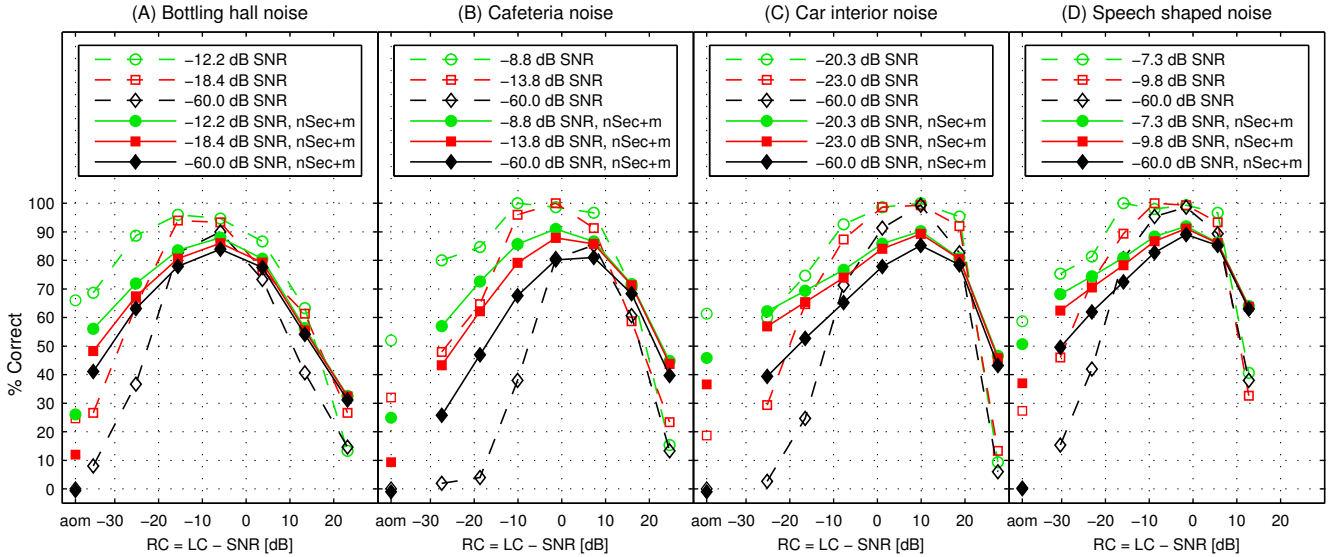


Figure 2: The modified $nSec$ with frequency normalization, compression, and DC removal compared to the subjective listening tests.

$nSec$ is decreased when $\mathbf{Y}(\tau, k)$ contains energy in the low energy regions of $\mathbf{T}(\tau, k)$; to get a high $nSec$, the binary mask should only present the high energy regions of $\mathbf{T}(\tau, k)$.

To improve the alignment between the subjective listening tests and the $nSec$, the following modifications are introduced:

4.1 Frequency Normalization

In speech signals, high frequencies have less energy than low frequencies, but this difference does not reflect the frequencies' importance to intelligibility. Using the $nSec$ without any frequency normalization will make the low frequencies dominate the result. Furthermore, the auditory system can to some degree adapt to the listening situation and a minor fixed coloration of the speech spectrum is not expected to affect intelligibility. To compensate for the difference in energy and any fixed colorations, we normalize the frequency channels to equal energy. This normalization has the drawback that at increasing RC values, when the binary mask becomes more sparse, some frequency channels will contain few non-zero elements, which would become very large because of the normalization. To avoid these high level time-frequency units, amplitude compression should follow the normalization (although in frequency channels with no non-zero elements, no normalization should be applied).

4.2 Compression

To decrease the relative importance of high level time-frequency units mainly produced by the frequency normalization, compression can be applied to $\mathbf{T}(\tau, k)$ and $\mathbf{Y}(\tau, k)$. Compression will move the peaks of the $nSec$ curves towards lower RC values, but also reduces the difference between the three SNR levels. To align the $nSec$ peaks with the subjective listening tests, $\mathbf{T}(\tau, k)$ and $\mathbf{Y}(\tau, k)$ are raised to the power of 0.15.

4.3 DC removal

As previously stated, the $nSec$ will be positive even if two random signals are used because their energy is always positive. To reduce this offset in the time-frequency representations, each frequency channel should be high-pass filtered. This high-pass filtering will push the values down to zero in the case where we have flat, but nonzero, energy and emphasize changes in energy instead of absolute levels. The used high-pass filter has a single zero at 1 and a single pole at 0.95.

5. RESULT

As seen in Figure 2, the modifications improve the correspondence between the subjective listening test and the $nSec$. The differences are most pronounced at low and high RC values where the slope of the modified $nSec$ is too shallow, and in the unprocessed condition (aom) the results are too low and too closely placed in the bottling hall and cafeteria noise condition. Ideally, the three SNR levels should give a intelligibility of 50%, 20% and 0%, but the compression, which was introduced to shift the peaks of the $nSec$ towards lower RC values, also compresses the results at low RC values, making them more equal. At high RC values the shallow slope of the modified $nSec$ is also a outcome from using compression. Compression increases the impact of low-amplitude time-frequency units and a more sparse mask is needed to reduce the $nSec$.

To allow some nonlinearity in the relationship between the $nSec$ and speech intelligibility, a logistic function can be applied:

$$p(c) = \frac{1}{1 + e^{(o-c)/s}}, \quad (4)$$

where o is the offset, and s is the slope of the logistic function [4]. To find the offset and the slope we use the unconstrained nonlinear minimization function `fminsearch` in Matlab to minimize the squared error between $nSec$ and the results from the subjective listening test using speech shaped noise. The found offset and slope of $o = 0.62$ and $s = 0.09$ are used to transform the $nSec$ results from Figure 2 into the results seen in Figure 3. The overall performance is improved: a better correspondence between the subjective listening tests and the $nSec$ is seen, but this is achieved at the expense of the match in the situation with no binary masking (aom).

6. DISCUSSION

Our proposed method uses a different approach compared to intelligibility measures as AI, SII, and STI [1, 8, 19] by using the correlation as the fundamental function for measuring intelligibility. In the AI, SII, and STI, the intelligibility is measured as a sum and weighting of SNR in a number of frequency channel. A more similar approach to ours is used in [11] for measuring speech quality and in [4] to measure intelligibility. In both works, the cross-correlation coefficient is used to measure the similarity between *internal representations* of the target and test signal. The internal representations are the expected patterns of neural activity from the auditory periphery calculated using the model by Dau et. al. [7]. In [4] the modu-

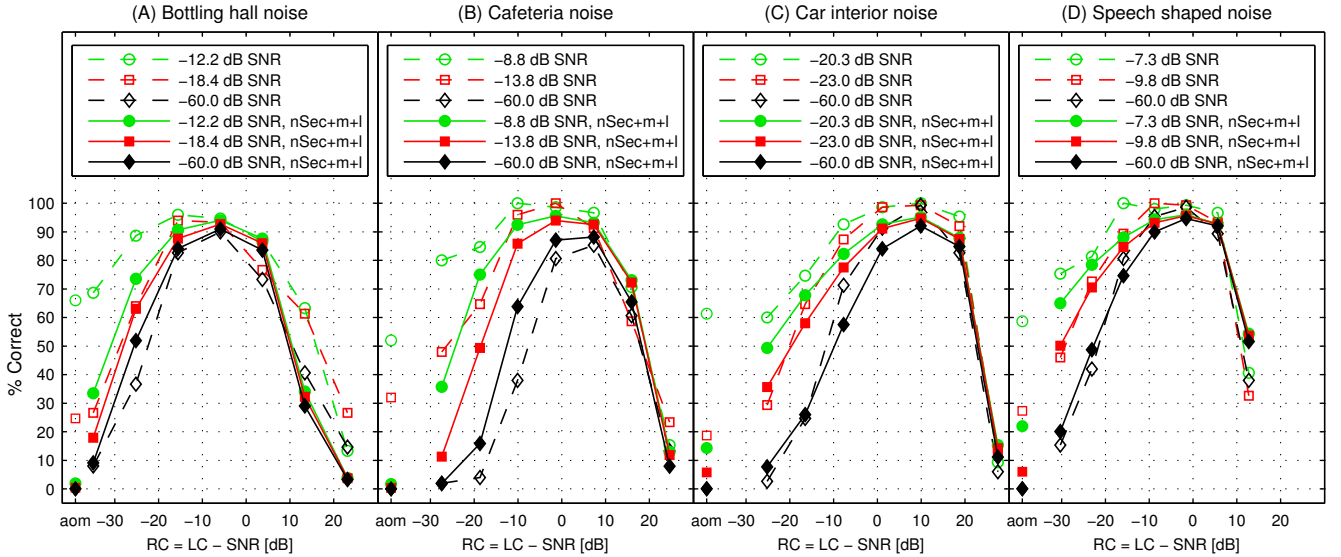


Figure 3: The modified $nSec$ transformed with a logistic function (Eq. 4) and compared to the subjective listening tests.

lation filterbank is replaced by a modulation low-pass filter, and the cross-correlation coefficient is calculated using short time frames of 20 ms with 50% overlap. The cross-correlation coefficients are grouped into low, medium, and high level correlation frames as in [12], but only the average of the high level correlation frames is used in the model output which is mapped to intelligibility using a logistic function equal to Equation 4.

The model by Christiansen et al. shows significant improvements compared to the speech-based coherence SII method [12] and the speech-based STI method [9], when using the same subjective results as in this study. The predicted intelligibility, using the same 10 sentences as for the $nSec$, is shown in Figure 4, and a substantial reason for the promising results is, as explained by the authors, the use of 20 ms frames, which is an interesting difference from the $nSec$. The main deviation between the model and the subjective results is found using the bottling hall noise (Fig.4.A), which is explained by Christiansen et al. to be caused by a too high influence from the low frequencies on the final result.

It is of interest to compare our approach and results with the model by Christiansen et al., but concluding which one is better is not appropriate from the results shown in Figure 3 and 4. Mainly because the logistic function used in Figure 3 was fitted directly to the subjective results using the speech shaped noise condition, whereas the logistic function used by Christiansen et al. was fitted to the psychometric curves from subjective listening tests of unprocessed mixtures at different SNR levels. The consequence of this difference is evident using the all-one mask, where the results from the $nSec$ are too close and too low, which is not the case for the model by Christiansen et al. An interesting difference between the two methods is the bottling hall noise, where the $nSec$, although very similar at the three SNR levels, has a better alignment of the peaks, which is caused by the frequency normalization as explained in section 4.1.

We might question whether the proposed modifications of the $nSec$ are the correct ones to use, and if they appear in the correct order. The modifications could be compared to processing steps in the auditory system, but in this case we have selected and ordered them purely to adjust the $nSec$ to the subjective results and not to simulate specific aspects of the auditory system. Similarly, the use of the correlation as underlying basis was supported by the preliminary results seen in Figure 1, and not by assumptions about correlation being used at some level in human perception. Introducing additional steps – simple or complicated – could potentially improve the precision of the method, but would also introduce ad-

ditional processing and parameters that would make the system less transparent for the user.

Another approach to measure intelligibility is the use of automatic speech recognition systems, where the number of correctly identified words or phonemes are used as a measure of speech intelligibility. This method has shown promising results [18, 5], but it is vulnerable to peculiarities of speech recognition systems that can make them differ widely from the perception of listeners. Trivial mismatch between the processed signals and the training data used by the recognizer can result in misleading low results.

A straightforward approach to evaluate time-frequency masking algorithms is to count the number of errors in the binary mask. Although we believe that the binary mask itself can explain a large amount of the intelligibility, this approach has various drawbacks e.g. the type of errors can have widely differing impact [15], the location of errors is important, and it is not certain which type of binary mask should be used as reference. Furthermore, this approach will not show the difference between applying the same binary mask to mixtures at different SNR levels.

The $nSec$ has shown a fine agreement with subjective listening test of the IBM applied to different mixtures and SNR levels, but this is only one of many methods of time-frequency masking. In the present work, we have not examined how the $nSec$ will behave using e.g. non-binary masks – the general case of applying a time-varying gain in a number of frequency bands – but we are hopeful that it will continue to agree with human performance. We note that the $nSec$ can fail if the target and system output become misaligned e.g. if the processed mixture is delayed compared to the target, however this could be accommodated by searching over a timing skew parameter (full cross-correlation).

7. CONCLUSION

By focusing on the correlation between the broad spectral envelope of target and system output, while completely ignoring the fine structure, we arrive at an intelligibility measure able to match a range of subjective results that would be very difficult to explain by SNR measures. We therefore suggest that future work on nonlinear speech enhancement, if it is concerned with intelligibility, should use measures based on correlation in place of SNR. To this end, we have released a simple drop-in implementation of our measure, written in Matlab¹.

¹See <http://labrosa.ee.columbia.edu/projects/intelligibility/>

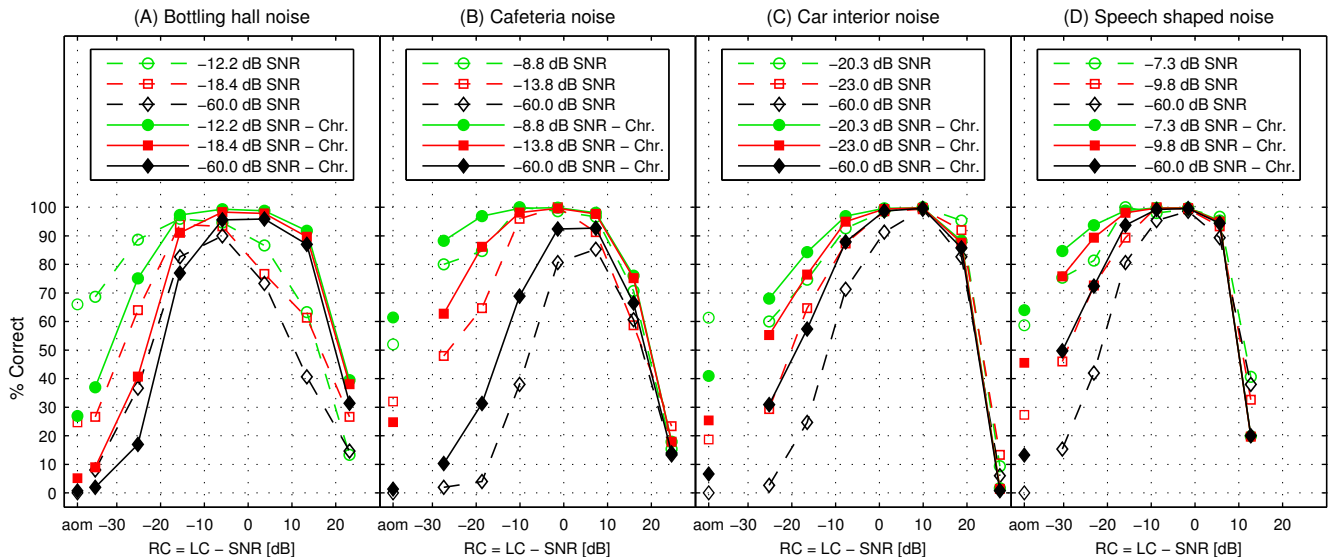


Figure 4: The predicted intelligibility using the model by Christiansen et al. [4] compared to the subjective listening test.

Although there are other, existing intelligibility measures that are able to match subjective data as closely as ours, our measure is constructed to be as simple as possible, with a consequent benefits in terms of transparency and diagnosis: when a system performs poorly under this measure, it is relatively easy to look at the processed envelopes going into the final correlation to see in which regions they are most different, thereby suggesting where to look for improvements. We hope that measures of this kind can help to focus and promote progress in speech intelligibility enhancement systems.

REFERENCES

- [1] ANSI S3.5-1997. American national standard: Methods for the calculation of the speech intelligibility index, 1997.
- [2] M. Anzalone, L. Calandruccio, K. Doherty, and L. Carney. Determination of the potential benefit of time-frequency gain manipulation. *Ear and Hearing*, 27(5):480–492, 2006.
- [3] D. Brungart, P. Chang, B. Simpson, and D. Wang. Isolating the energetic component of speech-on-speech masking with ideal time-frequency segregation. *JASA*, 120(6):4007–4018, 2006.
- [4] C. Christiansen, T. Dau, and M. S. Pedersen. Prediction of speech intelligibility based on an auditory model. *In preparation*.
- [5] M. Cooke. A glimpsing model of speech perception in noise. *JASA*, 119(3):1562–1573, 2006.
- [6] M. Cooke, P. Green, L. Josifovski, and A. Vizinho. Robust automatic speech recognition with missing and unreliable acoustic data. *Speech Comm.*, 34(3):267–285, 2001.
- [7] T. Dau, B. Kollmeier, and A. Kohlrausch. Modeling auditory processing of amplitude modulation. I. Modulation detection and masking with narrowband carriers. *J. Acoust. Soc. Am.*, 102:2892–2905, 1997.
- [8] N. R. French and J. C. Steinberg. Factors governing the intelligibility of speech sounds. *JASA*, 19(1):90–119, 1947.
- [9] I. Holube and B. Kollmeier. Speech intelligibility prediction in hearing-impaired listeners based on a psychoacoustically motivated perception model. *JASA*, 100(3):1703–1716, 1996.
- [10] Y. Hu and P. C. Loizou. A comparative intelligibility study of speech enhancement algorithms. In *Proc. ICASSP*, pages IV–561–564, Hawaii, 2007.
- [11] R. Huber and B. Kollmeier. PEMO-Q - a new method for objective audio quality assessment using a model of auditory perception. *IEEE Transactions on Audio, Speech and Language Processing*, 14(6):1902–1911, 2006.
- [12] J. M. Kates and K. H. Arehart. Coherence and the speech intelligibility index. *JASA*, 117(4 I):2224–2237, 2005.
- [13] U. Kjems, J. B. Boldt, M. S. Pedersen, T. Lunner, and D. Wang. Role of mask pattern in intelligibility of ideal binary-masked noisy speech. *Submitted to JASA*.
- [14] N. Li and P. C. Loizou. Effect of spectral resolution on the intelligibility of ideal binary masked speech. *JASA*, 123(4):EL59–EL64, 2008.
- [15] N. Li and P. C. Loizou. Factors influencing intelligibility of ideal binary-masked speech: Implications for noise reduction. *JASA*, 123(3):1673–1682, 2008.
- [16] M. S. Pedersen, D. Wang, J. Larsen, and U. Kjems. Two-microphone separation of speech mixtures. *IEEE Trans. on Neural Networks*, 19(3):475–492, 2008.
- [17] R. Shannon, F.-G. Zeng, V. Kamath, J. Wygonski, and M. Ekelid. Speech recognition with primarily temporal cues. *Science*, 270(5234):303–4, 1995.
- [18] S. Srinivasan and D. Wang. A model for multitalker speech perception. *JASA*, 124(5):3213–3224, 2008.
- [19] H. Steeneken and T. Houtgast. A physical method for measuring speech-transmission quality. *JASA*, 67(1):318–326, 1980.
- [20] K. Wagener, J. L. Jøsvassen, and R. Ardenkjaer. Design, optimization and evaluation of a danish sentence test in noise. *International Journal of Audiology*, 42(1):10–17, 2003.
- [21] D. Wang and G. J. Brown, editors. *Computational Auditory Scene Analysis*. Wiley & IEEE Press, Hoboken, New Jersey, 2006.
- [22] D. Wang, U. Kjems, M. S. Pedersen, J. B. Boldt, T. Lunner, and T. Lunner. Speech perception of noise with binary gains. *JASA*, 124(4):2303–2307, 2008.
- [23] O. Yilmaz and S. Rickard. Blind separation of speech mixtures via time-frequency masking. *IEEE Trans. on Signal Processing*, 52(7):1830–1847, 2004.