Objective Intelligibility Prediction of Speech by Combining Correlation and Distortion based Techniques

Angel M. Gomez[1,2], *Belinda Schwerin*[1], *Kuldip Paliwal* [1]

[1]Signal Processing Laboratory, School of Engineering, Griffith University, Australia
[2]Dpt. of Signal Theory, Networking and Communications, University of Granada, Spain
amgg@ugr.es, b.schwerin@griffith.edu.au, k.paliwal@griffith.edu.au

Abstract

A number of techniques based on correlation measurements have recently been proposed to provide an objective measure of intelligibility. These techniques are able to detect nonlinear distortions and provide intelligibility scores highly correlated with those given by human listeners. However, the performance of these techniques has not been found satisfactory for measuring the speech intelligibility of speech enhancement algorithms. In this paper we first investigate the different correlation-based methods, in the context of speech enhancement. We then propose to combine these correlation-based techniques with spectral distance-based ones. Results presented show that objective intelligibility prediction is significantly improved by this combination.

Index Terms: objective measure, speech intelligibility assessment, speech enhancement, correlation-based techniques, spectral distance-based techniques.

1. Introduction

In speech enhancement, algorithms aim to improve the quality and/or intelligibility of corrupted speech while minimizing the speech distortion introduced by the enhancement process. Here, speech quality is often characterized by the level of audible distortion which is present while intelligibility may be characterized by the amount of speech which can be correctly recognized. To evaluate the effectiveness of an enhancement algorithm at achieving this goal, subjective experiments where listeners rate the quality of stimuli or identify words can be used, but are timely and expensive. Consequently, objective measures which are fast and cheap to implement, are often relied on as an initial gauge of effectiveness. In this paper, we therefore investigate measures for predicting the intelligibility of speech processed using speech enhancement algorithms, with the aim of improving their correlation to subjective test scores.

Many of the objective measures used to evaluate speech intelligibility are based on the distance between the clean and corrupted speech, in either the time or spectral domain. Of the more known and used measures are the articulation index (AI) [1] and the speech transmission index (STI) method [2]. These spectral distance measures are based on the assumption proposed by French [1] that the intelligibility of a speech signal is given by the sum of the contributions to intelligibility within individual frequency bands. The AI method applies a function of the signal-to-noise ratio (SNR) in a set of bands then averages across these bands to predict intelligibility. This method was later extended to the SII method [3], which correlates well with subjective intelligibility for stimuli corrupted with additive noise. The STI method uses a modulation transfer function (MTF) instead of SNR in each subband, improving its correlation to subjective scores for stimuli distorted by linear filtering, reverberation, as well as additive noise. Many variations upon these methods have been reported in the literature, but generally still suffer the problem of being poorly correlated to the subjectively measured intelligibility of stimuli subjected to nonlinear processing [4, 5].

In speech enhancement, algorithms generally operate in the frequency domain, applying a suppression function to the noisy magnitude spectra in order to enhance speech. This results in nonlinear distortions and the intelligibility of the resulting stimuli is not accurately predicted by spectral distance based methods. For example, speech processed using spectral subtraction is predicted to improve intelligibility but subjective scores say otherwise.

Recently a new set of techniques, such as the short-time objective intelligibility (STOI) [6] measure and the excitation spectra correlation (ESC) method [7], have been proposed to take into account nonlinear distortion. These techniques maintain the main ideas from SII and STI, such as speech frequency band spanning and weighting but, instead of using an instantaneous distortion measure such as SNR, a correlation-based approach is followed. Correlation can be computed along the frequency or time dimension, and for the latter, in short-time or long-time segments, leading to different techniques and performance. These approaches are better able to evaluate some nonlinear distortions, providing intelligibility predictions that are well correlated with those from human listeners. However, as we will show, there are other types of distortions that are neglected by these correlation-based approaches.

In this paper, we investigate these correlation-based techniques and explore the types of distortions evaluated. We then propose their joint application with spectral distance techniques, and we show a significant improvement in intelligibility prediction accuracy can be achieved in this way.

2. Experimental framework

The experiments presented in this work make use of subjective intelligibility scores from a sentence intelligibility evaluation study reported in [8]. In [8], 20 IEEE sentences from [9], downsampled to 8kHz and additively corrupted with 4 real-world recorded noises from the AURORA database (babble, car, street, and train) [10] at SNRs of 0 and 5 dB, were processed by 8 noise-suppression algorithms to produce a total (including unprocessed noisy stimuli) of 72 treatments. Using these sentences as the corpus, the study in [8] conducted subjective intelligibility measuring experiments involving 40 native American English speaking participants, finding the mean subjective intelligibility scores associated with each treatment type.

In this work, each objective intelligibility measure is applied to each of the treatments and sentences of the above corpus. Objective scores for each sentence are then averaged to find a mean score for each treatment type. A logistic function, is then used to map this mean score to a value between 0 and 1. This function is given by

$$f(x) = \frac{1}{1 + e^{a+bx}} \tag{1}$$

where a and b are free parameters which are computed through a nonlinear least squares procedure which fits the objective scores with the subjective intelligibility ones.

Finally, Pearson's correlation coefficient, r, is used to assess the performance of each technique as a predictor of the intelligibility of corrupted speech. In addition, the standard deviation of the prediction error σ_e is also computed as $\sigma_e = \sigma_d\sqrt{1-r^2}$, where σ_d is the standard deviation of the speech intelligibility scores for a given treatment type.

3. Correlation based intelligibility techniques

A number of techniques have recently been proposed for objectively evaluating speech intelligibility that are based on Pearson's correlation or its squared counterpart. Pearson's correlation (or simply correlation) gives an indication of the linear relationship between two random variables. While correlation provides values between -1 and 1, squared correlation produces values between 0 and 1, being preferred when the sign of the linear relationship is not relevant. When a processed or noisy signal is compared through squared correlation to its corresponding clean version, a value close to 1 indicates the signals are linearly related, suggesting that only a weak nonlinear distortion is present. On the contrary, a value close to 0 could be interpreted as the presence of a strong nonlinear distortion.

In general, objective intelligibility techniques exploiting correlation share the principles of SII (speech spectra spanning and weighted averaging of each band measure). However, these techniques can consider different sampling frequencies, frame segmentation, number of frequency bands and band filter shape. In this work we start from a common framework where the same processing parameters are applied. This can lead to techniques slightly different from those proposed by their respective authors, but allows a better comparison of the different approaches.

The magnitude spectrum of the signal $x(\tau)$ is first computed through a short-time Fourier transform (STFT) on a frame by frame basis as follows,

$$X(n,k) = |\sum_{\tau=-\infty}^{\infty} x(\tau)w(n-\tau)e^{-j2\pi k\tau/N}| \tag{2}$$

where n refers to the discrete-time (frame) index, k is the discrete frequency bin, N is the frame duration and $w(n)$ is the analysis window. A Hamming window with frame duration of 32 ms is used for STFT analysis. This frame duration is justified through intelligibility studies, and is commonly used by many speech processing and intelligibility assessment techniques (e.g. [11, 12, 7]). Additionally, for a sample frequency of 8 kHz, the selected frame duration provides an adequate resolution (256 samples) for Fourier analysis. Frames are obtained every 8 ms so that STFT spectra are obtained with a frequency resolution of 125 Hz. By considering the trajectories of each acoustic frequency bin as an independent signal across time, modulation domain components up to a frequency of 62.5 Hz are retained. This ensures that no speech information is lost, with important speech information assumed to be at frequencies less than 16 Hz [13]. Applying 25 overlapping Gaussian-shaped filters spaced nonuniformly across frequency (in proportion to the ear's critical bands) over the STFT magnitude spectra [14], a final representation of the clean and processed speech, $X_j(n)$ and $Y_j(n)$, is then obtained as a function of both the frame number n and the band number j.

Correlation between these signals can now be computed. This can be performed along either the frequency or time dimension, for the complete signal or in short-time segments, leading to the following different schemes.

3.1. Correlation along frequency

Correlation between clean and processed signals can be computed along frequency as

$$r(n)^2 = \frac{(\sum_{j=0}^{J-1} X_j(n) - \hat{X}(n)) \cdot (\sum_{j=0}^{J-1} Y_j(n) - \hat{Y}(n))}{\sum_{j=0}^{J-1}(X_j(n) - \hat{X}(n))^2 \cdot \sum_{j=0}^{J-1}(Y_j(n) - \hat{Y}(n))^2} \tag{3}$$

where $\hat{X}(n)$ and $\hat{Y}(n)$ are the mean values across frequency for frame n of the clean and the processed signal, respectively. The intelligibility score is then obtained as an average of $r^2(n)$ over all the frames:

$$C_{freq} = \frac{1}{N}\sum_{n=0}^{N-1} r^2(n)\cdot \tag{4}$$

The resulting scheme is exactly the same as the excitation spectra correlation (ESC) method proposed in [7]. As shown in that work, this measure is related to the signal to residual noise ratio, whose time-domain counterpart is the segmental SNR. This technique performs relatively well when predicting intelligibility. As Table 1 shows, C_{freq} or the ESC technique yields a $r = 0.82$ correlation with the subjective intelligibility scores.

3.2. Correlation along time

Alternatively, correlation between clean and processed signals can be computed along the time dimension. Here, equation (3) is modified as

$$r_j^2 = \frac{(\sum_{n=0}^{N-1} X_j(n) - \hat{X}_j) \cdot (\sum_{n=0}^{N-1} Y_j(n) - \hat{Y}_j)}{\sum_{n=0}^{N-1}(X_j(n) - \hat{X}_j)^2 \cdot \sum_{n=0}^{N-1}(Y_j(n) - \hat{Y}_j)^2} \tag{5}$$

where now $\hat{X}_j$ and $\hat{Y}_j$ are the mean values along time for frequency band j of the clean and the processed signal, respectively. As before, an intelligibility score can be obtained by averaging r_j^2 along bands,

$$C_{time} = \frac{1}{J}\sum_{j=0}^{J-1} r_j^2\cdot \tag{6}$$

Table 1 shows the correlation and standard deviation for this approach. As can be seen, C_{time} yields a lower correlation than the scheme based on correlation along frequency (C_{freq}). However, when correlation is computed over time, the method can be modified to further improve results.

A very related approach to the C_{time} technique is the Normalized Covariance Metric (NCM) [4]. In this method, signals are first bandpass filtered and spanned into several bands. A Hilbert transform is then used to obtain the envelopes of each band. After being downsampled to 25 Hz, clean and processed envelopes are compared through correlation along time.

NCM and C_{time} scheme share the same basis with the following difference. In NCM, a filterbank is applied in time domain and a low pass filtering (in the downsampler) forces the maximum modulation component to 12.5 Hz. In the previous C_{time} scheme, the filterbank is applied in the spectral domain and the maximum modulation component is 65.5 Hz. Therefore reducing the maximum modulation component to the more speech related frequency of 12.5 Hz, results are slightly improved, as can be seen in Table 1 (C_{time} @ 12.5 Hz).

The final step performed in NCM transforms the (squared) correlation values r_j^2 into an SNR:

$$SNR_j = 10\log_{10}\left(\frac{r_j^2}{1-r_j^2}\right). \tag{7}$$

As in the SII technique, SNR values are limited in the range of [-15,15] dB and mapped linearly between 0 and 1. A weighted average is then performed along bands to compute the intelligibility score. This transformation into and limitation of SNR values, results in a better correlation of $r = 0.81$ (Table 1, *NCM*).

3.3. Correlation along short-time segments

Instead of considering all the frames, correlation can be computed in a short segment from the current frame. In such a way, non-stationary distortions can be better accounted for. Thus, a correlation value can be obtained as

$$r_j^2(n) = \frac{(\sum_{m=0}^{M-1} X_j(n-m) - \hat{X}_{j,n})(\sum_{m=0}^{M-1} Y_j(n-m) - \hat{Y}_{j,n})}{\sum_{m=0}^{M-1}(X_j(n-m) - \hat{X}_{j,n})^2 \sum_{m=0}^{M-1}(Y_j(n-m) - \hat{Y}_{j,n})^2} \tag{8}$$

where $\hat{X}_{j,n}$ and $\hat{Y}_{j,n}$ now represent the mean values of the M-frame block ending at frame n for clean and processed signals at band j, respectively. Then, an intelligibility score can be given as the average of these correlation values as,

$$C_{short-time} = \frac{1}{NJ} \sum_{n=0}^{N-1} \sum_{j=0}^{J-1} r_j^2(n) \cdot \tag{9}$$

Results obtained by this method depend on the number of frames considered in the block. Preliminary tests showed the best intelligibility prediction ($r = 0.81$, $C_{short-time}$ in Table 1) is achieved with $M = 192$.

As in the previous subsection, we can apply a 12.5 Hz low pass filter over $X_j(n)$ and $Y_j(n)$. As a result, the correlation with the subjective intelligibility scores is slightly improved ($r = 0.82$).

This same scheme is applied by the short-time objective intelligibility (STOI) method proposed in [6]. Instead of the quadratic value, STOI uses $r_j(n)$. Also, a voice activity detector preprocesses the speech signals to remove silence segments. An interesting addition in STOI is the modification performed over the processed signal, by which $Y_j(n)$ is modified in order to not exceed a maximum allowed signal to distortion ratio (SDR) [6]. This can be seen as the usual limitation imposed by many intelligibility techniques on perceived SNR per band. As a reference, Table 1 shows the results obtained through this technique (*STOI*).

4. Combination with spectral distance based techniques

As has been commented, correlation applied along time or along frequency, can reveal the presence of nonlinear distortions. In this way, techniques based on correlation achieve a reasonably good intelligibility prediction. However, correlation is completely unable to detect some other kinds of distortions. For example, a uniformly attenuated band along time (i.e. filtered) will be undetected by a C_{time}, or $C_{short-time}$ based technique (or C_{freq} for a uniformly attenuated frame). These distortions can affect intelligibility but are not taken into account by the correlation based techniques.

An intelligibility prediction technique based on correlation can benefit from a joint use with a non-correlation based one. Use with a technique based on a spectral distance measurement, for example, is enough to improve the intelligibility prediction. To show this, we have tested three different distance based (and not correlation based) techniques described in the literature: the PESQ algorithm [15], the frequency weighted segmental SNR [11], and the SNRloss method described in [7]. As can be noted, the first two of these are intended for speech quality rather than intelligibility. However, these techniques have in common that they compute a distortion metric based solely on instantaneous information within the frame.

The PESQ algorithm transforms clean and processed signals to the sone domain, where it computes the distance measure between them [15]. The frequency weighted segmental SNR works in a similar way to SII, but the speech spectra are first normalized to have an area of one, and then spanned into bands. An $SNR_j(n)$ is obtained for each band in each frame as,

$$SNR_j(n) = 10 \log_{10} \frac{X_j(n)^2}{(X_j(n) - Y_j(n))^2} \cdot \tag{10}$$

These values are limited in the range of [-15, 15] dB and linearly mapped to values between 0 and 1. A single score is finally obtained by a weighted average along frequency bands and a simple mean along time.

A similar scheme is followed in the SNRloss method with the following exceptions. Spectra normalization is avoided, and SNR computation is substituted by a simple distortion measure given by,

$$SD_j(n) = 10 \log_{10} \frac{X_j(n)^2}{Y_j(n)^2} \cdot \tag{11}$$

A single score is finally obtained in the same way as before, but with the SNR $SD_j(n)$ range limited to [-3, 3] dB.

In addition to testing combinations of correlation and non-correlation based techniques, we have also tested the combination of the C_{freq} technique with all other correlation based techniques, as it can be considered orthogonal to them (as well as related to the residual distortion ratio).

In order to join the intelligibility measures provided by these non-correlation based techniques with the correlation based ones, a linear combination is proposed. This has the advantage that the logistic function can be easily extended from that given in eqn. (1) to,

$$y = \frac{1}{1 + e^{a+bx_1+cx_2}}, \tag{12}$$

where y is the final adjusted score, x_1 and x_2 are the scores provided by each technique, and a, b and c are the linear parameters to be adjusted by the least square procedure.

Table 1 shows the results obtained by the techniques of Section 3 when linearly combined with the PESQ algorithm (columns 4 and 5), the frequency weighted segmental SNR (columns 6 and 7), the C_{freq} technique (columns 8 and 9), and the SNRloss method (columns 10 and 11). As can be seen, intelligibility prediction is significantly improved by this joint application, achieving particularly good results when the SNRloss method is applied. In such a case, correlation with subjective intelligibility scores can improve up to a $r = 0.90$ value when STOI is used.

As a reference, Table 2 shows the results obtained by the PESQ algorithm, the frequency weighted segmental SNR and the SNRloss method individually. In addition, this table also show the results achieved when PESQ is combined with SNRloss, and C_{time} with STOI, as these are the best pairs found when techniques of the same class are combined.

5. Discussion

In general, the objective intelligibility prediction techniques based on correlation perform reasonably well (results between $r = 0.78$ and $r = 0.86$). By comparing the basic C_{freq} and C_{time} techniques one can derive that frequency dimension is initially the best one in which to compute the correlation between clean and processed signals. However, with some additional processing, such as short-time computation and 12.5 Hz filtering, the difference

Objective measure	No combination		PESQ comb.		fwSNRseg comb.		C_{freq} comb.		SNRloss comb.	
	r	σ_e	r	σ_e	r	σ_e	r	σ_e	r	σ_e
C_{freq} (*ESC*)	0.82	0.10	0.86	0.09	0.84	0.09	–	–	0.86	0.09
C_{time}	0.78	0.11	0.83	0.10	0.86	0.09	0.86	0.09	0.87	0.09
C_{time} @ 12.5 Hz	0.79	0.11	0.84	0.09	0.88	0.08	0.87	0.09	0.88	0.08
NCM	0.81	0.10	0.85	0.09	0.88	0.08	0.88	0.08	0.88	0.08
$C_{short-time}$	0.81	0.10	0.84	0.09	0.87	0.09	0.86	0.09	0.87	0.09
$C_{short-time}$ @ 12.5 Hz	0.82	0.10	0.85	0.09	0.88	0.08	0.88	0.08	0.89	0.08
STOI	0.86	0.09	0.88	0.08	0.89	0.08	0.89	0.08	0.90	0.08

Table 1: *Results obtained for the described correlation-based techniques by themselves and when jointly used with different distortion-based techniques.*

Objective measure	r	σ_e
PESQ	0.79	0.11
fwSNRseg	0.78	0.11
SNRloss	0.82	0.10
PESQ + SNRloss	0.84	0.09
C_{time} + STOI	0.86	0.09

Table 2: *Reference results obtained by PESQ algorithm (PESQ), the frequency weighted segmental SNR (fwSNRseg) and the SNRloss method (SNRloss) individually, and by the best same-class combination found.*

is not longer significant. This can be explained first by the fact that C_{freq} is somewhat related to a per band SNR measure. Indeed, the correlation achieved by this technique is the same as the SNRloss method and it is in the range of other signal to distortion based techniques [12]. On the other hand, C_{freq} allows a frame-by-frame analysis which is not available in the C_{time} technique. This frame-by-frame analysis can be included in the form of a short-time correlation computation as in $C_{short-time}$ and its related variations (including STOI) and, as can be seen, better results are obtained.

When techniques are combined, those based on correlation along time can clearly benefit from combination with non-correlation based techniques. C_{freq} has a particularly interesting behavior. Its union with non-correlation based techniques can improve results, but the best ones are obtained when a technique based on correlation along time is applied. In fact, the joint application of these techniques implies that a correlation measure is computed both in time and frequency dimensions and then linearly mixed. In that way, uniform distortions as proposed in Section 4 will be detected.

Finally, it is worth noting that $C_{short-time}$ with 12.5 Hz filtering and STOI provides almost the same results when combined with the SNRloss method. This would suggest that the additional processing performed in STOI, particularly the processed signal modification so as to not to exceed certain SDR, is in fact oriented towards alleviating the inability of the correlation-based techniques to detect some kinds of distortions. An alternative way of solving this would be by using a combination with a spectral distance based technique.

6. Conclusions

In this paper we summarize the different schemes in which correlation measures between a clean and a processed signals are used to predict speech intelligibility. These have in common that, although some nonlinear distortions are well detected, other distortions can be neglected. Due to this, a joint combination with spectral distance based measures, such as PESQ, the frequency weighted segmental SNR and the SNRloss method, is proposed. This allows a better intelligibility prediction which is highly correlated with the intelligibility scores provided by real listeners.

7. Acknowledgments

The authors would like to thank professor P. Loizou for his invaluable help sharing and explaining the intelligibility database.

This work has been supported by the Spanish government grant JC2010-0194 and project CEI BioTIC GENIL (CEB09-0010).

8. References

[1] N. French and J. Steinberg, "Factors governing the intelligibility of speech sounds," *J. Acoust. Soc. Amer.*, vol. 19, no. 1, pp. 90–119, 1947.

[2] H. Steeneken and T. Houtgast, "A physical method for measuring speech-transmission quality," *J. Acoust. Soc. Amer.*, vol. 67, no. 1, pp. 318–326, Jan 1980.

[3] A. N. S. Institute, "Methods for calculation of the speech intelligibility index," *Technical Report S3.5-1997*, 1997.

[4] R. Goldsworthy and J. Greenberg, "Analysis of speech-based speech transmission index methods with implications for nonlinear operations," *J. Acoust. Soc. Amer.*, vol. 116, no. 6, pp. 3679–3689, Dec 2004.

[5] C. Ludvigsen, C. Elberling, and G. Keidser, "Evaluation of a noise reduction method–comparison of observed scores and scores predicted from sti," *Scand. Audiol. Suppl.*, vol. 38, pp. 50–55, 1993.

[6] C. Taal, R. Hendriks, R. Heusdens, and J. Jensen, "A short-time objective intelligibility measure for time-frequency weighted noisy speech," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Process. (ICASSP)*, Dallas, Texas, USA, Mar 2010, pp. 4214–4217.

[7] J. Ma and P. Loizou, "Snr loss: A new objective measure for predicting the intelligibility of noise-suppressed speech," *Speech Communication*, vol. 53, no. 3, pp. 340–354, Mar 2011.

[8] Y. Hu and P. C. Loizou, "Subjective comparison and evaluation of speech enhancement algorithms," *Speech Communication*, vol. 49, no. 7–8, pp. 588–601, Jul-Aug 2007.

[9] E. Rothauser, "Ieee recommended practice for speech quality measurements," *IEEE Trans. on Audio Electroacoust.*, vol. 17, no. 3, pp. 225–246, Sep 1969.

[10] D. Pearce and H. Hirsch, "The aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions," in *Proc. of International Conference on Spoken Language Processing (ICSLP)*, China, Oct 2000, pp. 29–32.

[11] Y. Hu and P. Loizou, "Evaluation of objective quality measures for speech enhancement," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 16, no. 1, pp. 229–238, Jan 2008.

[12] J. Ma, Y. Hu, and P. Loizou, "Objective measures for predicting speech intelligibility in noisy conditions based on new band-importance functions," *J. Acoust. Soc. Amer.*, vol. 125, no. 5, pp. 3387–3405, May 2009.

[13] R. Drullman, J. Festen, and R. Plomp, "Effect of temporal envelope smearing on speech reception," *J. Acoust. Soc. Amer.*, vol. 95, no. 2, pp. 1053–1064, Feb 1994.

[14] P. Loizou, *Speech Enhancement: Theory and Practice*. Boca Raton, FL: Taylor and Francis, 2007.

[15] "Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assesment of narrow-band telephone networks and speech codecs," *ITU-T P.862 Recommendation*, 2001.