Available online at www.sciencedirect.com

SciVerse ScienceDirect

Speech Communication 54 (2012) 503–515

SPEECH COMMUNICATION

www.elsevier.com/locate/specom

Improving objective intelligibility prediction by combining correlation and coherence based methods with a measure based on the negative distortion ratio

Angel M. Gómez [a,b,*], Belinda Schwerin [b], Kuldip Paliwal [b]

[a] Dept. Teoría de la Señal, Telemática y Comunicaciones, University of Granada, Facultad de Ciencias, Campus de Fuentenueva S/N, Granada 18071, Spain
[b] Signal Processing Laboratory, School of Engineering, Griffith University, Australia

Received 22 August 2011; received in revised form 7 October 2011; accepted 3 November 2011
Available online 12 November 2011

Abstract

In this paper we propose a novel objective method for intelligibility prediction of enhanced speech which is based on the negative distortion ratio (NDR) – that is, the amount of power spectra that has been removed in comparison to the original clean speech signal, likely due to a bad noise estimate during the speech enhancement procedure. While negative spectral distortions can have a significant importance in subjective intelligibility assessment of processed speech, most of the objective measures in the literature do not well account for this type of distortion. The proposed method focuses on a very specific type of noise, so it is not intended to be used alone but in combination with other techniques, to jointly achieve a better intelligibility prediction. In order to find an appropriate technique to be combined with, in this paper we also review a number of recently proposed methods based on correlation and coherence measures. These methods have already shown a high correlation with human recognition scores, as they effectively detect the presence of nonlinearities, frequently found in noise-suppressed speech. However, when these techniques are jointly applied with the proposed method, significantly higher correlations (above $r = 0.9$) are shown to be achieved.

Keywords: Speech intelligibility; Objective measures; Speech enhancement; Distance-based methods; Correlation-based methods; Coherence-based methods; Negative spectral distortion

1. Introduction

Speech enhancement algorithms aim to improve the quality and/or intelligibility of corrupted speech signals. Normally, this is done by reducing the noise such that the residual noise is not perceptually annoying to the listener, while minimizing the distortion introduced by the enhancement process. The quality of the resulting speech can generally be characterized by the level of audible distortion, while the intelligibility can be characterized by the amount of speech that can be correctly recognized. In applications where humans are the end user of enhanced speech signals, subjective tests, where listeners rate the quality of stimuli or identify words, are the most reliable method for quantifying the perceived quality or intelligibility of speech processed by different enhancement algorithms (Falk and Chan, 2008). However, these tests are time consuming and expensive. For this reason, there is an increasing interest in developing objective measures that accurately predict the quality and/or intelligibility of a speech signal. In this work, we investigate measures for improved prediction of the intelligibility of speech processed using enhancement algorithms, with the aim of improving their correlation to subjective intelligibility scores.

* Corresponding author at: Dept. Teoría de la Señal, Telemática y Comunicaciones, University of Granada, Facultad de Ciencias, Campus de Fuentenueva S/N, Granada 18071, Spain. Tel.: +34 958243271; fax: +34 958243230.

E-mail addresses: amgg@ugr.es (A.M. Gómez), b.schwerin@griffith.edu.au (B. Schwerin), k.paliwal@griffith.edu.au (K. Paliwal).

0167-6393/$ - see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.specom.2011.11.001

Early attempts to predict speech intelligibility led to the development of the articulation index (AI) (French and Steinberg, 1947), which correlates well with subjective intelligibility for stimuli corrupted with additive noise. This method accounts for the contribution of different regions of the spectrum to intelligibility, applying a function of the signal-to-noise ratio (SNR) in a set of bands and performing a weighted average across them. The AI method was later extended to the speech-intelligibility index (SII), and finally standardized by ANSI (ANSI, 1997). Most of the objective measures proposed after SII share the assumption that the intelligibility of a speech signal is given by the sum of the contributions to intelligibility within individual frequency bands (French and Steinberg, 1947). The speech transmission index (STI) (Steeneken and Houtgast, 1980), an objective method widely used for room-acoustics assessment, applies the same bandwidth spanning as SII but, instead of computing an SNR in each subband, uses a modulation transfer function (MTF). The MTF function allows STI to detect reductions in temporal envelope modulation, thereby improving its correlation to subjective scores for stimuli distorted by reverberation, linear filtering, as well as additive noise.

Many variations upon the above 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 (Goldsworthy and Greenberg, 2004; Ludvigsen et al., 1993). This makes these methods unsuitable for the intelligibility assessment of noise-suppressed signals, as nonlinear distortions are frequently introduced by speech enhancement algorithms. As an example, speech processed using spectral subtraction is predicted to improve intelligibility while subjective scores say otherwise.

In this paper we propose a novel objective method for intelligibility assessment of noise-suppressed speech. This method relies on an idea supported by some authors (e.g., Ma and Loizou, 2011; Loizou and Kim, 2011; Kim and Loizou, 2010) – namely, the usefulness of distinguishing between two types of distortions, according to the sign of the difference between the corrupted and clean spectral components. While positive distortions commonly appear in noise corrupted signals, negative ones are only expected after speech enhancement processing. This is because speech enhancement algorithms generally rely on an estimate of the noise to achieve noise reduction, and consequently can also remove some of the clean spectra. With the exception of some methods, such as the SNRloss measure (Ma and Loizou, 2011), most of the intelligibility evaluation techniques lump these positive and negative distortions together, paying no attenuation to the sign. However, while positive distortions can be concealed by the ear, negative ones could imply some loss of information from the original speech spectra. Therefore, the perceptual effects of these two distortions on speech intelligibility should not be assumed to be equivalent (Kim and Loizou, 2010).

The method proposed in this work provides a score based on the negative distortion ratio measured between clean and processed signals. This score is not intended to be used alone but in combination with another intelligibility prediction technique as, otherwise, the rest of distortions introduced during the enhancement procedure would be neglected. In order to find an appropriate technique for the proposed approach to be combined with, we also review a number of recently proposed methods based on correlation and coherence measures, such as the short-time objective intelligibility (STOI) (Taal et al., 2011) measure and the coherence SII (CSII) method (Kates and Arehart, 2005), among others. These methods have shown a high correlation with subjective test scores for enhanced speech signals, as they effectively detect the presence of nonlinearities. Nevertheless, a significant improvement in intelligibility prediction accuracy can be achieved when these approaches are combined with our method.

The rest of the paper is organized as follows. First, the proposed technique is detailed in Section 2, while in Section 3, we provide a brief review of different correlation and coherence based methods that will be combined with it. Then, these methods are tested under the experimental framework described in Section 4 and individual and combined results are presented in Section 5. Finally, Section 6 summarizes the conclusions of this work.

2. Negative distortion ratio

There is some evidence that positive and negative differences between the processed and clean spectra have different perceptual effects over intelligibility (Ma and Loizou, 2011; Loizou and Kim, 2011; Kim and Loizou, 2010). A positive spectral distortion appears when a spectral component in the enhanced signal is greater in magnitude than the corresponding clean one (positive difference), and can be interpreted as residual noise which has not been completely removed by the enhancement algorithm. On the contrary, a negative spectral distortion occurs when this difference is negative, as a result of an excessive removal of energy from a component, likely due to a bad noise estimate used in the suppression function.

Of particular interest for evaluating speech enhancement methods are these negative spectral distortions, as they are predominantly introduced by the enhancement procedure, and many of the measures in the literature do not well account for this type of distortion. Therefore, in this work we propose a new intelligibility prediction measure that focuses on this type of distortion – that is, the negative difference between the enhanced and clean spectra. However, since the resulting method neglects other distortions which also reduce the intelligibility of speech, the method is not intended to be used by itself. Instead, it is used in combination with other techniques to achieve an improved intelligibility prediction.

The proposed measure is obtained from a critical-band spectral representation of the clean and processed signals. Here we assume that both signals are time-aligned and have a sampling rate of 8000 Hz. Initially, a spectral representation of both signals is obtained through a short-time Fourier transform (STFT) as:

$$X(n,k)=\sum_{l=-\infty}^{\infty}x(l)w(n-l)e^{-j2\pi kl/N}, \tag{1}$$

where n refers to the discrete-time index, k is the discrete frequency bin, N is the frame duration in samples, and $w(n)$ is the analysis window. A Hamming window is used as the analysis window, and signals are segmented into frames of 32 ms in duration with a 75% overlap (i.e., an 8 ms shift). This frame duration is widely used by other intelligibility assessment techniques (e.g., Hu and Loizou, 2008; Ma et al., 2009; Ma and Loizou, 2011), it is within the range of 20–40 ms typically used in speech processing, and given the sampling frequency (8 kHz), provides an adequate STFT resolution.

Critical-band analysis is then performed by means of a filterbank consisting of 25 overlapping Gaussian-shaped windows (Loizou, 2007). Magnitude values in the STFT-bins are weighted and summed according to each Gaussian-shaped window as:

$$X_j(m)=\sum_{k=0}^{N-1}|X(mT,k)|\cdot W_j(k), \tag{2}$$

where T is the frame shift (8 ms) and $W_j(k)$ represents the jth filter window from the filterbank. Filters (starting with a center frequency of 50 Hz as shown in Fig. 1) are spaced in proportion to the ear's critical bands.

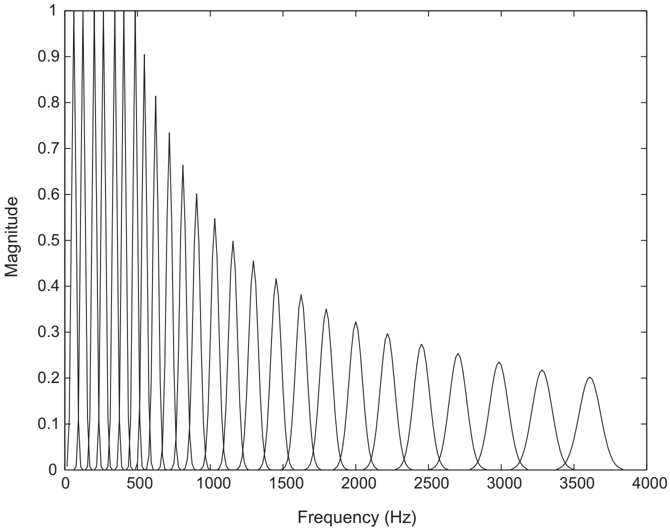


Fig. 1. Filterbank with Gaussian-shaped windows applied during critical-band analysis.

For clean signal $x(n)$ and processed signal $y(n)$, the relationship between the spectral representations of both signals can be expressed as:

$$\begin{aligned}Y_j(m)=X_j(m)+D_j(m),\quad & j=0,1,2,\ldots,J-1,\\ & m=0,1,2,\ldots,M-1,\end{aligned} \tag{3}$$

where J is the number of bands; M is the number of frames; $X_j(m)$ and $Y_j(m)$ are the filterbank output for band j at frame m for the clean and processed signals, respectively; and $D_j(m)$ represents the difference between them (distortion).

A simple signal-to-noise ratio (SNR) for each band j and frame m could be obtained from this difference as:

$$SNR_j(m)=10\log_{10}\frac{X_j(m)^2}{D_j(m)^2}. \tag{4}$$

In order to reduce the effect from a particularly high or low SNR value over the final score, ratio values can be restricted to the range $[-SNR_L, SNR_U]$ and then linearly mapped into the $[0,1]$ range. Finally, an overall intelligibility prediction score for the complete signal can be calculated by averaging across bands and time as:

$$SNR=\frac{1}{JM}\sum_{j=0}^{J-1}\sum_{m=0}^{M-1}Q_j(m), \tag{5}$$

where $Q_j(m)$ is the restricted and linearly mapped SNR for band j at frame m.

The metric described above is similar to the AI method, except that while the AI method applies a weighting to account for the relative importance (for intelligibility) of each band, no (or equivalently uniform) weighting was applied here. There is further similarity to the critical band version of SII, where uniform weights were applied from 450 to 4000 Hz. In each case however, no distinction is made between spectral positive and negative differences since, as can be seen, the square operation is applied over $D_j(m)$.

In this work, instead of combining both positive and negative distortions into a single value, we propose to focus on subtractive distortions only, thereby deriving a unique measure from it. Thus, instead of an SNR, we define the negative distortion ratio (NDR) for each band j and frame m as:

$$NDR_j(m)=\begin{cases}20\log_{10}\left(-\frac{X_j(m)}{D_j(m)}\right), & \text{if } D_j(m)<0,\\ SNR_U, & \text{if } D_j(m)\geqslant 0.\end{cases} \tag{6}$$

As can be seen, only subtractive distortions are taken into account while additive ones are neglected. When an additive distortion is found, a fixed SNR_U value (which later maps to a value of 1) is returned. Also, it must be noted that $|D_j(m)|\leqslant X_j(m)$ when $D_j(m)<0$, otherwise filter output $Y_j(m)$ would be lower than zero.[1] Therefore, $NDR_j(m)$ is always a positive value, and mapping can be simplified to:

[1] Implicitly, this also means that no negative distortions can be found in silence and pause segments. In these segments, $X_j(m)$ approaches zero.

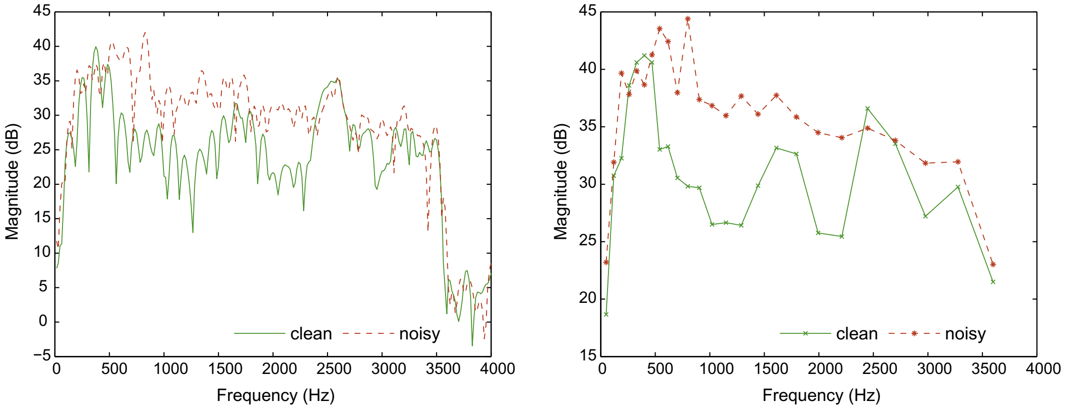


Fig. 2. Magnitude spectra (left) and critical filterbank outputs (right) obtained for a speech frame in clean condition and corrupted additively by car noise using a single windowed frame.

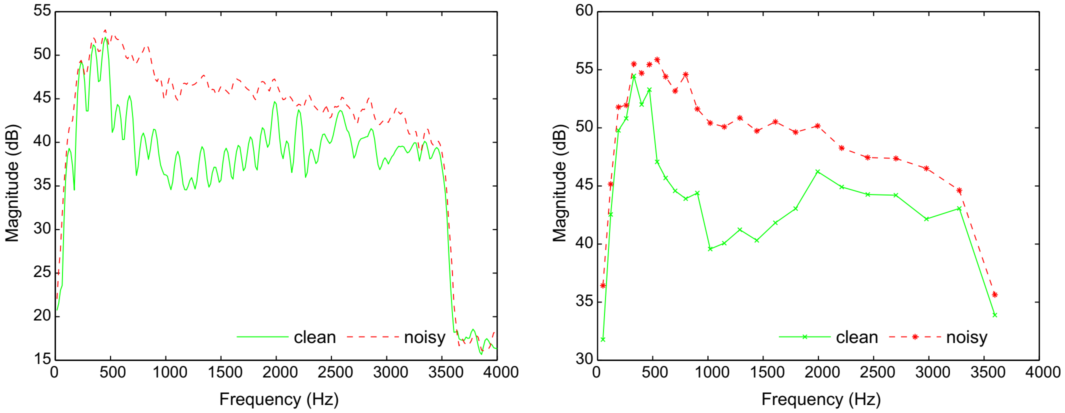


Fig. 3. Magnitude spectra (left) and critical filterbank outputs (right) obtained for a speech frame in clean condition and corrupted additively by car noise, using an averaged periodogram over 20 previous frames.

$$Q_j(m) = \frac{\min\left(NDR_j(m), SNR_U\right)}{SNR_U}, \tag{7}$$

where SNR_U now behaves like a threshold up to which a negative distortion ratio is considered, and is a tunable parameter that can be determined experimentally. As before, an overall intelligibility prediction score NDR is obtained by averaging the $Q_j(m)$ values across time and bands.

It is noteworthy to recall that the negative distortion ratio is obtained from spectral estimators which inherently present some variability. This variability can cause pernicious effects on the metric. When speech is contaminated with additive noise, only positive differences between the noisy and clean spectra are expected to be found. However, in practice, negative differences can frequently appear due to the variability of the spectral estimator. This is demonstrated by Fig. 2 (left) which shows an example of a magnitude spectrum of a clean speech frame (shown as a green line) compared with its corresponding noisy version (shown as red dashes). As can be seen, there are regions where the noisy magnitude spectra is under the clean one. This also affects the critical band representation, shown in Fig. 2 (right), so that some filter outputs from corrupted speech appear under the clean ones.

We can reduce this variability on the spectra by considering an averaged periodogram instead of a simple DFT. By averaging K consecutive frames, estimator variance is reduced by a factor of $\sqrt{K}$. Fig. 3 shows an example of this. As can be seen, after an averaging of 20 previous frames, no regions of clean spectra or filter outputs are under the noisy ones.

Using this approach, spectra variability can be controlled, so that negative differences can be avoided when speech signals are only distorted by additive noise. This approach still, however, preserves those differences caused by the enhancement algorithm itself. This is demonstrated by Figs. 4 and 5, where the averaged magnitude spectra and critical filterbank outputs for the same clean and corrupted speech frames are compared after the noisy one has been enhanced by traditional Wiener filtering (Scalart and Filho, 1996).[2] As can be seen in Fig. 5, despite 20-frame averaging, negative distortions introduced by the

[2] Stimuli used to generate Figs. 2–5 are from the corpus of (Hu and Loizou, 2007) as described in Section 4. Wiener filtered stimuli (from the corpus) were constructed using a reference implementation (Loizou, 2007) of Wiener filtering based on *a priori* SNR estimation (Scalart and Filho, 1996).

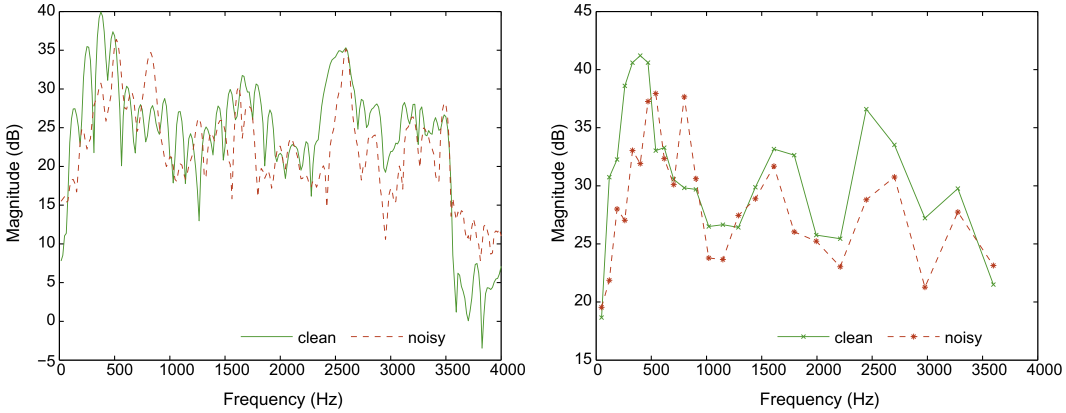


Fig. 4. Magnitude spectra (left) and critical filterbank outputs (right) obtained for a speech frame in clean condition and corrupted additively by car noise and enhanced by Wiener filtering, using a single windowed frame.

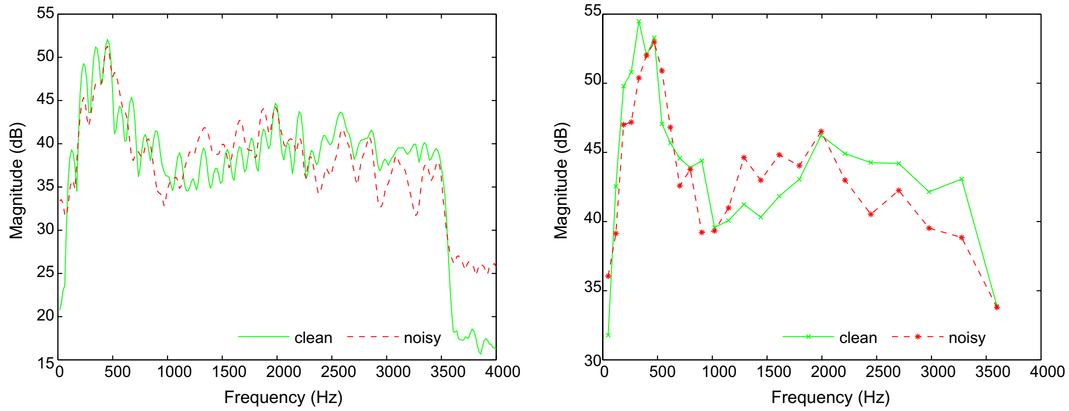


Fig. 5. Magnitude spectra (left) and critical filterbank outputs (right) obtained for a speech frame in clean condition and corrupted additively by car noise and enhanced by Wiener filtering, using an averaged periodogram over 20 previous frames.

enhancement procedure are still present in the magnitude spectra. While it could be argued that averaging has negative effects when we analyse speech signals, as stationarity is not assured over such long periods, here we are interested in the distortion caused by the enhancement process, not in the speech signal itself. Thus, as long as this distortion does not change rapidly, we can afford the averaging of several frames in order to provide a better estimation of the spectral distortion.

We can easily incorporate the above idea in our method by applying a moving average (in time) over the filterbank outputs as:

$$\overline{X}_j(m) = \frac{1}{K}\sum_{k=0}^{K} X_j(m-k), \tag{8}$$

$$\overline{D}_j(m) = \frac{1}{K}\sum_{k=0}^{K} D_j(m-k), \tag{9}$$

and by replacing $X_j(m)$ and $D_j(m)$ in Eq. (6) by $\overline{X}_j(m)$ and $\overline{D}_j(m)$, respectively.

3. Correlation and coherence based intelligibility measures

The proposed NDR method provides a distance measure focused on a very specific type of distortion. As such, it neglects any other distortions caused by the enhancement procedure which also reduce the intelligibility of speech. Nonlinear distortions, which are well detected by correlation and coherence based methods, are a good example of this. In Gomez et al. (2011) we showed that by combining distance-based measures (of which NDR is an example), with correlation-based techniques, we can achieve better intelligibility predictions than by using either measure alone. Therefore in this section we provide a brief review of different correlation and coherence based methods that will be applied in combination with the proposed NDR method with the aim of improving intelligibility predictions.

Methods used for measuring nonlinear distortions generally make use of one of two metrics which are particularly good in revealing nonlinearities: the Pearson's correlation (or its squared counterpart) (Gibbons, 1985), and the magnitude-squared coherence (MSC) function (Carter et al., 1973). Pearson's correlation (or correlation coefficient) gives an indication of the linear relationship between two random variables. While the 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. Squared correlation can be used to compare critical bands from clean and processed signals. While a value close to 1 indicates a strong linear relation between them, a value close to 0 could be interpreted as the presence of a strong nonlinear distortion. On the other hand, the MSC function is a real function between zero and one which gives the fraction of signal power linearly related at each frequency between two signals. Similarly, MSC can be used to reveal the amount of the enhanced signal power that is linearly dependent on the clean one, while its complementary value can be related to the unrelated fraction or non-linear distortion.

The remainder of this section provides details of different correlation and coherence based metrics, as they will be applied in the experiments of later sections. Although these share the same principles as SII, that is, speech spectra spanning and weighted averaging of each band measure, they often apply different sampling frequencies, frame segmentation, number of frequency bands and band filter shapes. Here, we will use a unified framework in order to provide a better comparison of the different techniques. This can lead to techniques slightly different from those proposed by their respective authors, but it will allow us to better analyze the effect of the several refinements included in them. To this end, we will consider the same signal sampling (8000 Hz), time-alignment and segmentation (32 ms, 75% overlapping) proposed in Section 2, as well as an identical critical-band analysis (25 overlapping Gaussian-shaped windows). When required by the method, band frequency weighting will be performed according to the band-importance functions described in (ANSI, 1997) for sentence stimuli.

3.1. Correlation-based methods

A simple way to measure the similarity between clean and noisy speech signals is by frame computation of the correlation coefficient (Silverman and Dixon, 1976). Thus, given a frame index m, the squared correlation over the filterbank outputs can be obtained as (Ma and Loizou, 2011):

$$r^2(m) = \frac{\left(\sum_{j=0}^{J-1}\left(\left(X_j(m)-\overline{X}_m\right)\cdot\left(Y_j(m)-\overline{Y}_m\right)\right)\right)^2}{\sum_{j=0}^{J-1}\left(X_j(m)-\overline{X}_m\right)^2\cdot\sum_{j=0}^{J-1}\left(Y_j(m)-\overline{Y}_m\right)^2}, \tag{10}$$

where J is the number of bands, and $\overline{X}_m$ and $\overline{Y}_m$ are the mean values across frequency for frame m of the clean and the processed signal, respectively. It is worth mentioning that $r^2(m)$ is related to the signal to residual noise ratio (SNR_{ES}) (Ma and Loizou, 2011), for which the time-domain counterpart is the segmental SNR. Correlation computed along frequency is used in the excitation spectra correlation (ESC) method presented in (Ma and Loizou, 2011), which proposes an intelligibility score obtained as an average of $r^2(m)$ over all frames:

$$\mathrm{ESC} = \frac{1}{M}\sum_{m=0}^{M-1} r^2(m). \tag{11}$$

Alternatively, we can compute the correlation between clean and processed filterbank outputs along the time dimension. In such a case, a squared correlation per filter band can be obtained as:

$$r_j^2 = \frac{\left(\sum_{m=0}^{M-1}\left(\left(X_j(m)-\overline{X}_j\right)\cdot\left(Y_j(m)-\overline{Y}_j\right)\right)\right)^2}{\sum_{m=0}^{M-1}\left(X_j(m)-\overline{X}_j\right)^2\cdot\sum_{m=0}^{M-1}\left(Y_j(m)-\overline{Y}_j\right)^2}, \tag{12}$$

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

$$C_{time} = \frac{1}{J}\sum_{j=0}^{J-1} r_j^2. \tag{13}$$

However, this measure can be enhanced by applying some considerations like those included in the normalized covariance metric (NCM) (Goldsworthy and Greenberg, 2004). NCM computes r_j^2 coefficients in the same way as defined above, except that the filterbank output trajectories are low-pass filtered with a 12.5 Hz cutoff frequency.[3] This filtering is performed because important speech information is usually assumed to be at frequencies less than 16 Hz (Drullman et al., 1994). We can extend this low-pass filtering to the C_{time} scheme, and we refer to the resulting method as C_{time}(12.5 Hz).

In addition, NCM transforms the squared correlation values r_j^2 into an SNR per band:

$$SNR_j = 10\log_{10}\left(\frac{r_j^2}{1-r_j^2}\right). \tag{14}$$

As was done in SII, SNR values are limited to the range of $[-15, 15]$ dB, to prohibit excessively high or low values from disrupting the metric, then mapped linearly between 0 and 1. A weighted average is then performed across bands to compute the intelligibility score as:

$$NCM = \frac{\sum_{j=0}^{J-1} w_j\cdot Q_j}{\sum_{j=0}^{J-1} w_j}, \tag{15}$$

[3] In NCM, the filterbank is implemented in the time domain, that is, signals are bandpass filtered and spanned into several bands, and a Hilbert transform is used to obtain the envelopes of each band. Then, envelopes are low-pass filtered and downsampled to 25 Hz before being compared through correlation along time.

where Q_j and w_j are, respectively, the SNR mapped value and the band-importance weight for frequency band j.

Finally, instead of considering all frames, correlation along time can also be computed for a short segment up to the current frame. In such a way, non-stationary distortions can be better accounted for. In order to do so, Eq. (12) can be modified as:

$$r_j^2(m) = \frac{\left(\sum_{l=0}^{L-1}\left(\left(X_j(m-l) - \overline{X}_{j,m}\right) \cdot \left(Y_j(m-l) - \overline{Y}_{j,m}\right)\right)\right)^2}{\sum_{l=0}^{L-1}\left(X_j(m-l) - \overline{X}_{j,m}\right)^2 \sum_{l=0}^{L-1}\left(Y_j(m-l) - \overline{Y}_{j,m}\right)^2}, \tag{16}$$

where $\overline{X}_{j,m}$ and $\overline{Y}_{j,m}$ now represent the mean values of the L-frame block ending at frame m for clean and processed signals at band j, respectively. Again, a simple intelligibility score can be given as the average of these correlation values as:

$$C_{short\text{-}time} = \frac{1}{JM}\sum_{m=0}^{M-1}\sum_{j=0}^{J-1} r_j^2(m)\cdot \tag{17}$$

In addition, for comparison purposes, we can also derive a $C_{short\text{-}time}$ (12.5 Hz) measure by applying a 12.5 Hz low pass filter over $X_j(m)$ and $Y_j(m)$ as before.

The short-time objective intelligibility (STOI) method (Taal et al., 2011) computes an objective score very similarly to $C_{short\text{-}time}$ but uses $r_j(m)$ instead of the quadratic value. By design (signal sample rate, window length and overlapping), spectral components above 40 Hz from filter-output trajectories are discarded. Contrary to other intelligibility techniques, here correlation is not transformed into an SNR and then limited to a range (as in NCM). Instead, a clipping is performed over the processed signal, by which $Y_j(m)$ is modified such that it does not exceed a maximum allowed distortion (Taal et al., 2011). Also, a voice activity detector pre-processes the speech signals to remove silence segments.

3.2. Coherence-based techniques

Formally, the magnitude-squared coherence (MSC) between two signals is defined as (Carter et al., 1973):

$$|\gamma(\omega)|^2 = \frac{|S_{xy}|^2}{S_{xx}(\omega)S_{yy}(\omega)}, \tag{18}$$

where S_{xy}, S_{xx} and S_{yy} are the cross-spectral and the power spectral densities of $x(n)$ and $y(n)$, respectively. The MSC coherence represents the fraction of power linearly related between the clean and the enhanced signals along frequency. Although a similar interpretation can be given to the correlation when computed across time for each filter band, it must be noted that only magnitude spectra are used there and thus, any phase information is neglected. On the contrary, thanks to the use of the cross-spectral density, MSC can account for not only the in-phase spectra (cospectrum) but also the out-of-phase ones (quadspectrum).

For finite signals, the MSC function can be estimated by computing the cross and power spectra through a number of M overlapping windowed segments as:

$$|\gamma(k)|^2 = \frac{\left|\sum_{m=0}^{M-1} X(mT,k)Y^*(mT,k)\right|^2}{\sum_{m=0}^{M-1}|X(mT,k)|^2 \sum_{m=0}^{M-1}|Y(mT,k)|^2}, \tag{19}$$

where the asterisk denotes the complex conjugate, T is the frame shift, and $X(n,k)$ and $Y(n,k)$ are the short-time Fourier transform of the clean and the processed speech, respectively. Filter windows from critical-band analysis, $W_j(k)$, can then be applied over the MSC, providing a coherence measure per band as:

$$MSC_j = \sum_{k=0}^{N-1} |\gamma(k)|^2 \cdot W_j(k). \tag{20}$$

Thus, a simple coherence-based score can be given by simply averaging the MSC_j across critical-bands as:

$$MSC = \frac{1}{J}\sum_{j=0}^{J-1} MSC_j. \tag{21}$$

It must be noted that bias and variance effects are present on MSC due to the finite number of segments used in the estimation procedure. Although these can be alleviated using large overlaps (>50%) (Kates, 1992), as was used in this work, they make computing a coherence-based measure over short-time segments cumbersome. A way to border this limitation is proposed in the coherence SII (CSII) method (Kates and Arehart, 2005) where a speech to (non-linear) distortion ratio (SDR) is obtained for each frame as:

$$SDR_j(m) = 10log_{10}\frac{\sum_{k=0}^{N-1}\widehat{P}(m,k)\cdot W_j(k)}{\sum_{k=0}^{N-1}\widehat{N}(m,k)\cdot W_j(k)}, \tag{22}$$

where $\widehat{P}(m,k)$ and $\widehat{N}(m,k)$ are, respectively, estimations of the speech and noise power spectra, obtained as:

$$\widehat{P}(m,k) = |\gamma(k)|^2 \cdot |Y(mT,k)|^2 \tag{23}$$

$$\widehat{N}(m,k) = \left(1 - |\gamma(k)|^2\right) \cdot |Y(mT,k)|^2, \tag{24}$$

that is, using the power spectra of the processed speech, $|Y(n,k)|^2$, and $|\gamma(k)|^2$ and its complementary value. While MSC represent the fraction of the output signal power which is linearly dependent on the input at frequency bin k (i.e. speech), the complementary fraction, $1 - |\gamma(k)|^2$ gives the output power that is unrelated, that is, the nonlinear distortion and noise.

Finally, SDR values are limited within the range of $[-15,15]$ dB (consistent with the limitation applied in SII and a number of other measures), and mapped linearly between 0 and 1. An average is performed first across frames to compute the intelligibility score per band. Then, per band averages are weighted taking into account the band importance and combined to provide an intelligibility score. Originally, the simplified ro-ex filters suggested by

Moore and Glasberg (1983) are proposed for the CSII method. However, in this paper these have been replaced by the overlapping Gaussian-shaped filterbank described in Section 2 as discussed at the beginning of this section.

4. Experimental framework

In order to evaluate the above described methods and, in particular, the proposed NDR measure, the corpus and the subjective scores from the sentence intelligibility evaluation study reported in (Hu and Loizou, 2007) have been used in this work. In the cited study, the recordings available in (Loizou, 2007), consisting of all the sentences from the IEEE sentence database (Rothauser, 1969)[4] recited by a male speaker, were downsampled to 8 kHz and additively corrupted with 4 real-world recorded noises from the AURORA database (babble, car, street, and train) (Pearce and Hirsch, 2000) at SNRs of 0 and 5 dB. Then, 8 noise-suppression algorithms were applied to produce a total of 72 treatments, including unprocessed noisy stimuli. Using these sentences as the corpus, the study in (Hu and Loizou, 2007) conducted subjective intelligibility measuring experiments involving 40 native American English speaking participants. Each listener assessed a total of 18 different treatments, each one consisting of 20 sentences, ensuring no subject listened to the same sentence twice. Finally, mean subjective intelligibility scores were found for each treatment type from the percentage of words correctly identified (where all words were considered in the scoring).

In this work, each objective intelligibility measure is applied to each of the sentences of the above corpus. No pause/silence removal is considered in any method except for STOI, which follows the reference implementation in (Taal et al., 2011). A mean score for each treatment type is then obtained by averaging objective scores for each sentence. In order to compare the objective and subjective values, a mapping function is applied. This is used as objective measures do not directly predict an absolute intelligibility value, but a monotonic relationship is present between the objective scores and the results from listening experiments. For this purpose, and as done by many works in the literature (e.g., Ma and Loizou, 2011; Taal et al., 2011; Kates and Arehart, 2005; Boldt and Ellis, 2009; Christiansen et al., 2010), we use a logistic function such as:

$$l(x) = \frac{1}{1 + e^{\beta_0 + \beta_1 x}}, \tag{25}$$

where x is the objective score, and parameters β_0 and β_1 are known as regression coefficients. Values for these coefficients are computed through a logistic regression (Balakrishnan, 1992), as those which best fit the objective scores to the subjective intelligibility scores.

Logistic regression is applied in this work not only for the mapping of scores, but also to facilitate the combination of different intelligibility prediction techniques with the proposed NDR method. The logistic function always takes on values between zero and one, and can be interpreted as a way of describing the relationship between one or more independent variables and a probability, in our case the probability of correctly identifying words (i.e., subjective intelligibility score). Thus, the *logit*, or the total linear combination of all the independent variables used in the model, can be extended to $\beta_0 + \beta_1 x_1 + \beta_2 x_2$, modifying the logistic function as:

$$l(x_1, x_2) = \frac{1}{1 + e^{\beta_0 + \beta_1 x_1 + \beta_2 x_2}}, \tag{26}$$

to include the scores from two different objective methods (x_1 and x_2). In this way, the regression coefficients describe the size of the contribution of each method, which is automatically obtained during the logistic regression (as those values which provide the best fit).

The leave-one-out cross-validation procedure is applied to parameter fitting and logistic function mapping in order to ensure use of mutually exclusive training and testing sets, and thereby prevent overfitting. In this procedure, regression parameters are determined using the entire data set excepting one treatment. Intelligibility predictions for the excluded treatment are then evaluated via the above logistic function with these determined parameters, providing a mapped score. The procedure is repeated for every treatment in the corpus.

Finally, Pearson's correlation coefficient, r, between subjective and objective scores is used to assess the performance of each technique as a predictor of corrupted speech intelligibility. 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. Absolute values of r nearer to one and smaller values of σ_e indicate a better speech intelligibility prediction.

5. Results

As discussed in previous sections, the proposed negative distortion ratio (NDR) is intended to be used in combination with other intelligibility prediction methods. In particular, findings reported in (Gomez et al., 2011) suggest that best results would be achieved by combining it with a correlation-based technique. For comparison, we therefore begin by evaluating each of the correlation and coherence based methods described in Section 3.

Table 1 (columns labeled *Not combined*) summarizes the correlation coefficients (r) and the standard deviation of the prediction errors (σ_e) between the objective scores from these correlation and coherence based methods and the recognition scores from real listeners along the entire testing database. Here, $C_{short\text{-}time}$ and $C_{short\text{-}time}$ (12.5 Hz) have been applied with a block length of $L = 192$ frames

[4] IEEE database contains phonetically balanced sentences with low word-context predictability.

Table 1
Correlation coefficients and standard deviation of the prediction errors between the predictions from correlation and coherence based methods and the recognition scores from real listeners, when applied individually or jointly with the NDR method ($SNR_U = 4$ dB and $K = 192$).

Objective measure	Not combined		NDR combined	
	r	σ_e	r	σ_e
PESQ	0.779	0.109	0.788	0.107
fwSNRseg	0.769	0.111	0.762	0.113
SNRloss	0.806	0.103	0.799	0.105
ESC	0.811	0.102	0.834	0.096
C_{time}	0.766	0.112	0.858	0.089
C_{time} (12.5 Hz)	0.775	0.110	0.862	0.088
NCM	0.796	0.106	0.878	0.083
$C_{short\text{-}time}$	0.796	0.105	0.883	0.082
$C_{short\text{-}time}$ (12.5 Hz)	0.814	0.101	0.894	0.078
STOI	0.854	0.091	0.904	0.074
MSC	0.811	0.102	0.860	0.089
CSII	0.870	0.086	0.912	0.071

(selected based on the best predictions achieved in preliminary experiments), while STOI has been used with the parameters proposed by its authors (15 dB distortion clipping and 400 ms block). In addition, the correlation and the standard deviation from the PESQ algorithm (ITU-T P.862, 2001), the frequency weighted segmental SNR (fwSNRseg) (Hu and Loizou, 2008) and the SNRloss method described in (Ma and Loizou, 2011) (with $SNR_{Lim} = 3$ dB and $C_- = C_+ = 1$) have also been included. These last three techniques are well known methods often referenced in the literature which compute a distance-based measure. We could also include in this group the ESC measure, which computes the correlation along the filterbank outputs within a frame and it is closely related to the residual distortion ratio (Ma and Loizou, 2011). As can be observed, all the evaluated methods show an acceptable correlation with the recognition scores obtained from real listeners. However, the best performance ($r = 0.87$) is achieved by the CSII technique, followed by STOI ($r = 0.85$), both based on non-linearity measures. From the results, we can briefly remark that filtering the filterbank output trajectories improves the correlation-based predictions (C_{time} vs C_{time} (12.5 Hz), $C_{short\text{-}time}$ vs $C_{short\text{-}time}$ (12.5 Hz)). Also, working on a short-time basis is indicated to be beneficial (C_{time} vs $C_{short\text{-}time}$, MSC vs CSII), as is converting the correlation and coherence metrics into SNRs and/or limiting the effect of extreme values over the total score (C_{time} (12.5 Hz) vs NCM, $C_{short\text{-}time}$ (12.5 Hz) vs STOI, MSC vs CSII).

Table 1 also shows the correlation and standard deviation achieved when the aforementioned techniques are combined with the proposed NDR method (column labeled *NDR combined*). To perform this combination, the extended logistic function of Eq. (26) is used. As can be observed, although distance-based methods scarcely benefit from this joint application, correlation and coherence based techniques significantly improve, yielding higher correlations and lower error standard deviations. Again, best prediction correlations are achieved by STOI and CSII which, after combining with the proposed NDR method, improve to $r = 0.90$ and $r = 0.91$, respectively.

Fig. 6 shows the correlation coefficients obtained for all the techniques, individually and by combining with the NDR method, after restricting the test corpus by type of noise (babble, car, street and train). Again, objective intelligibility predictions improve after the combination for all the correlation and coherence based techniques, while distance-based ones remain practically the same. In particular, the NDR combination is shown to be significantly beneficial when predicting intelligibility from noise-suppression algorithms applied to street noise corrupted speech. In this specific case, it is worth mentioning that the objective methods based on non-linearity measures significantly reduce their accuracy when this kind of noise is present. However, when jointly applied with the NDR method, their high correlation ($r > 0.9$) is recovered.

As mentioned in Section 2, the proposed method depends on two parameters, the SNR_U threshold or the maximum SNR up to which a negative distortion is considered, and the number of frames, K, averaged to avoid the effects of spectra variability. In the previous results (Table 1 and Fig. 6), $SNR_U = 4$ dB and $K = 192$ were considered. These values were selected after a set of tests were carried out for each combination of SNR_U and K taking values from 0.25 to 15 dB and from 1 to 256 frames, respectively. Fig. 7 summarizes the results from these tests for each of the non-linearity based measures. As can be observed, despite the variation in results due to the technique with which the NDR methods was jointly applied, consistent behavior is present in all results. In general, sustained improvements in correlation with real listeners are obtained for a wide range of SNR_U and K values, forming a kind of plateau which decreases for excessively low SNR_U values and shows a slight crest at $K = 192$. The shapes obtained when the NDR method is combined with techniques based on long and short-time correlation (C_{time} and $C_{short\text{-}time}$ with and without a 12.5 Hz low-pass filtering on the filterbank trajectories), as well as with the NCM method, are practically identical. When the STOI, MSC and CSII methods are considered instead, along with the above mentioned plateau, a prominence can be found around $SNR_U = 4$ dB, making the crest at $K = 192$ more noticeable. This suggests that, around these values, some beneficial interaction additionally appears between the NDR method and these techniques. In the case of combinations with STOI and CSII methods, this might be explained by the additional pre- and post-processing operations that are included.

As we showed in (Gomez et al., 2011), intelligibility predictions provided by correlation-based techniques can be improved by combining them with distance-based ones. This is explained by the fact that correlation is completely unable to detect some types of distortions otherwise easily detectable. As an example, a seriously but uniformly

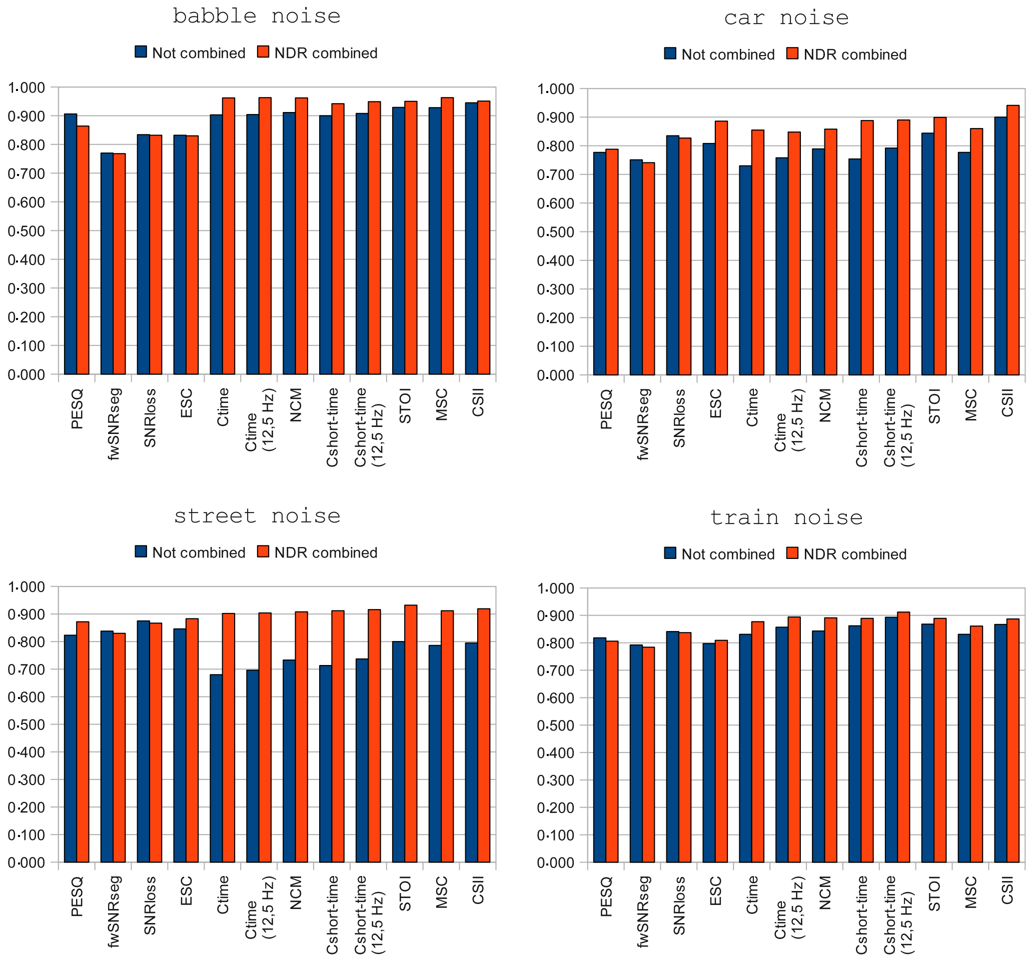


Fig. 6. Correlation achieved by correlation and coherence based methods, when applied individually or jointly with the NDR method, grouped by noise type.

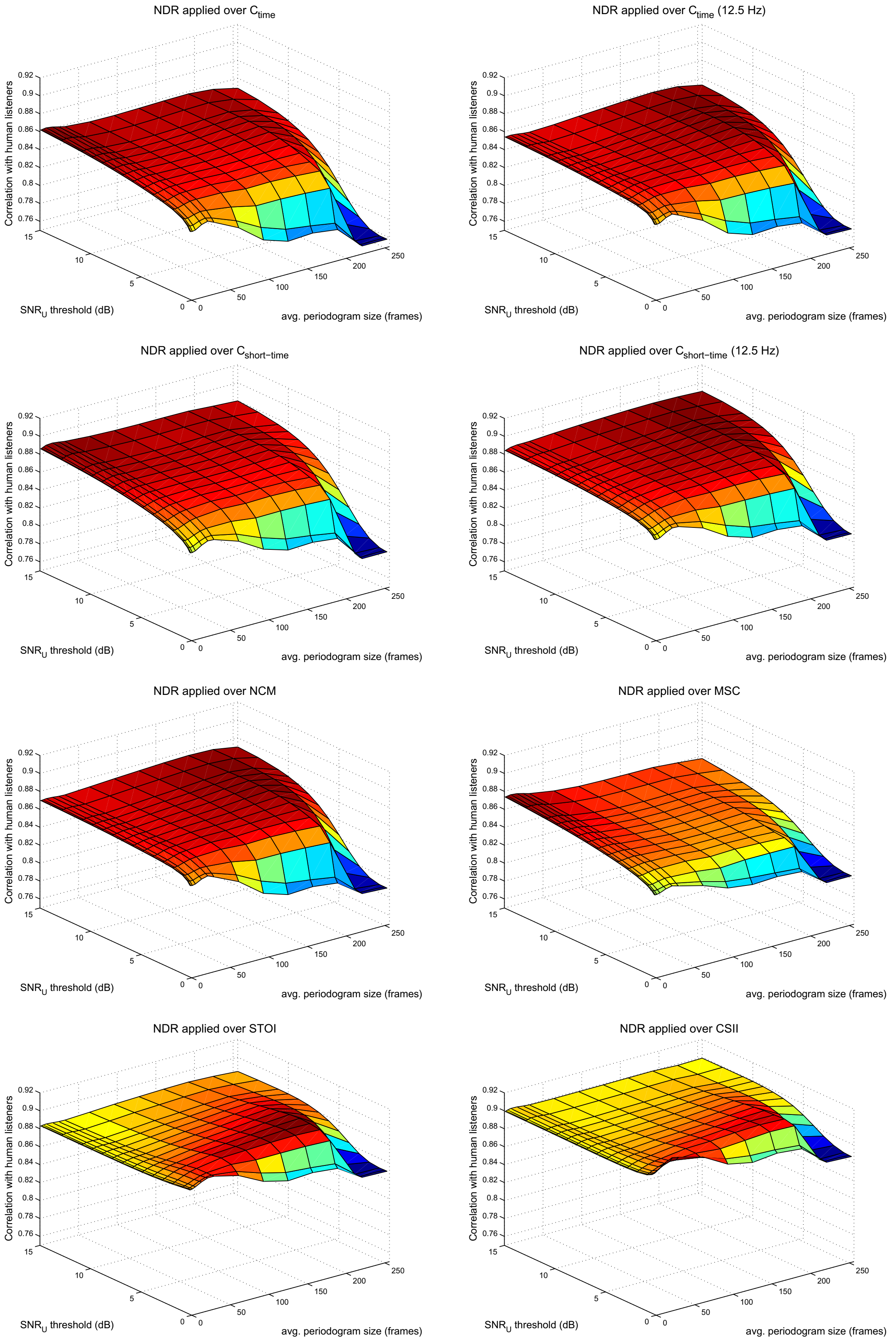
attenuated band along time (i.e., filtered) will affect the speech intelligibility but is undetected by a C_{time} or $C_{short\text{-}time}$ based technique (or ESC for a uniformly attenuated frame). The same reasoning can be extended to coherence-based techniques. Table 2 shows the correlation coefficient achieved when several techniques based on non-linearity measures (C_{time} and $C_{short\text{-}time}$ with and without the 12.5 Hz trajectory filtering, NCM, MSC, STOI and CSII) are combined with different distance-based measures (PESQ, ESC, frequency weighted segmental SNR and SNRloss), as well as the NDR method. As expected, significant improvements in the predictions are obtained by this joint application.

As can be observed in Table 2, when combined, NDR achieves similar or better correlation with human scores than achieved using any of the other distance-based measures. When long-time based methods are considered, combinations with SNRloss, fwSNRseg, and NDR result in quite similar correlations. However, for combinations with short-time based ones, NDR outperforms the other distance-based measures investigated. Introducing a measure focused on the speech attenuation suffered in the enhanced signal seems particularly beneficial for these techniques. Combinations with others metrics which implicitly take into account the energy removed from speech, either in terms of SNR, as in the frequency weighted segmental SNR, or in terms spectral distortion, as in the SNRloss metric, also yield significant improvements (rows 5 and 6 in Table 2). However, as the results for NDR show, best intelligibility predictions are achieved when this information is separated from other information such as positive spectral distortions.

Finally, Fig. 8 shows the correlation coefficients achieved by $C_{short\text{-}time}$ (without and with 12.5 Hz low-pass trajectory filtering), STOI and CSII when combined with distance-based methods, after restricting the test corpus by noise type (babble, car, street and train). As can be observed, the highest correlations are achieved when the NDR method is jointly applied. This is especially true in the case of street noise suppressed speech (as before), but also with car noise suppressed speech.

6. Conclusions

In this paper, we have proposed and evaluated a novel objective method based on the negative distortion ratio



NDR applied over C_{time}

NDR applied over C_{time} (12.5 Hz)

NDR applied over $C_{short-time}$

NDR applied over $C_{short-time}$ (12.5 Hz)

NDR applied over NCM

NDR applied over MSC

NDR applied over STOI

NDR applied over CSII

Fig. 7. Correlation achieved by several techniques based on non-linearity measures when combined with the NDR method considering different SNR_U thresholds and averaged periodogram sizes (K).

Table 2
Correlation coefficients obtained by the techniques based on non-linearity measures when combined with PESQ, ESC, frequency weighted segmental SNR (fwSNRseg), SNRloss method and the proposed NDR method.

	C_{time}	C_{time} (12.5 Hz)	*NCM*	*MSC*	$C_{short\text{-}time}$	$C_{short\text{-}time}(12.5Hz)$	*STOI*	*CSII*
No comb.	0.766	0.775	0.796	0.811	0.796	0.814	0.854	0.870
PESQ	0.817	0.826	0.833	0.836	0.825	0.838	0.862	0.867
ESC	0.847	0.860	0.866	0.853	0.853	0.870	0.878	0.875
fwSNRseg	0.852	0.866	0.872	0.851	0.856	0.874	0.880	0.879
SNRloss	0.857	0.870	0.874	0.853	0.857	0.875	0.887	0.882
NDR	0.858	0.862	0.878	0.860	0.883	0.894	0.904	0.912

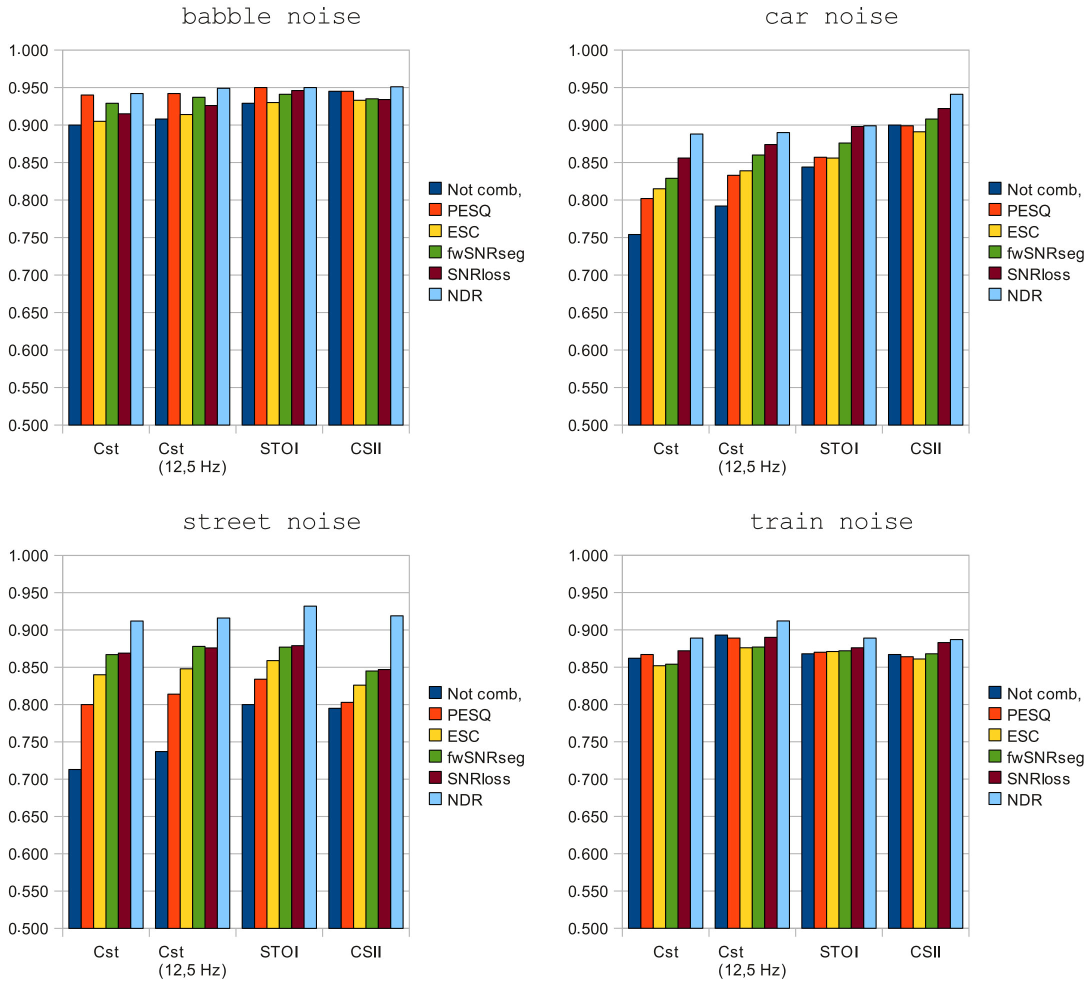


Fig. 8. Correlation obtained by $C_{short\text{-}time}$ (Cst), $C_{short\text{-}time}$ with 12.5 Hz low-pass filterbank output trajectory filtering (*Cst* (12.5 Hz)), STOI and CSII when combined with distortion-based techniques grouped by noise type.

for intelligibility prediction of noise-suppressed speech signals. This method obtains a critical-band representation from clean and processed speech and computes a distance-based metric fixed only on the negative distortion, that is, the negative difference between the enhanced and clean spectra. Negative spectral distortion is predominantly introduced by enhancement algorithms and there exists evidence which supports its different perceptual effect over intelligibility. As in many other methods, the proposed measure is bounded to avoid excessively high values disrupting the metric. In addition, averaged periodograms are considered during critical-band analysis to minimize the pernicious effects of spectral estimation variability.

As the presented method focuses only on a specific type of distortion, it is not intended to be used alone but in combination with another intelligibility assessment technique. Recently, a number of novel methods based on correlation and coherence measures have been successfully applied to the intelligibility evaluation of enhanced speech. In this paper, we investigate them and propose their joint application with our method. As a result, a better intelligibility prediction, highly correlated with the recognition scores provided by real listeners, is achieved.

Although combining correlation and distance based methods is not a novel idea, our measure significantly improves the predictions of these methods (and also coherence-based ones) in comparison to that achieved by combining them with other distance-based methods, such as PESQ, the frequency weighted segmental SNR or the SNRloss method. This is particularly true in the case of correlation and coherence based methods which operate on a short-time basis, such as $C_{short\text{-}time}$, STOI and CSII. In these cases, introducing a measure such as NDR, which focuses on the speech attenuation suffered in the enhanced signal, significantly reduces the deviation of the prediction error and improves the correlation to human intelligibility scores.

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).

References

ANSI, 1997. Methods for Calculation of the Speech Intelligibility Index. Technical Report S3.5-1997.

Balakrishnan, N., 1992. Handbook of the Logistic Distribution. Dekker.

Boldt, J., Ellis, D. 2009. A simple correlation-based model of intelligibility for nonlinear speech enhancement and separation. In: Proc. EUPSIPCO 2009, Glasgow, Scotland, pp. 1849–1853.

Christiansen, C., Pedersen, M.S., Dau, T., 2010. Prediction of speech intelligibility based on an auditory preprocessing model. Speech Commun. 52 (7-8), 678–692.

Carter, G., Knapp, C., Nuttall, A., 1973. Estimation of the magnitude-squared coherence function via overlapped fast fourier transform processing. IEEE Trans. Audio Electroacoust. 21, 337–344.

Drullman, R., Festen, J., Plomp, R., 1994. Effect of temporal envelope smearing on speech reception. J. Acoust. Soc. Am. 95 (2), 1053–1064.

Falk, T.H., Chan, W.-Y., 2008. A non-intrusive quality measure of dereverberated speech. In: Proc. Internat. Workshop on Acoustic Echo and Noise Control.

French, N., Steinberg, J., 1947. Factors governing the intelligibility of speech sounds. J. Acoust. Soc. Am. 19 (1), 90–119.

Gibbons, J., 1985. Nonparametric Statistical Inference, second ed. Dekker.

Goldsworthy, R., Greenberg, J., 2004. Analysis of speech-based speech transmission index methods with implications for nonlinear operations. J. Acoust. Soc. Am. 116 (6), 3679–3689.

Gomez, A., Schwerin, B., Paliwal, K., 2011. Objective intelligibility prediction of speech by combining correlation and distortion based techniques. In: Proc. ISCA European Conf. on Speech Communication and Technology (EUROSPEECH), Florence, Italy.

Hu, Y., Loizou, P., 2007. A comparative intelligibility study of single-microphone noise reduction algorithms. J. Acoust. Soc. Am. 122 (3), 1777–1786.

Hu, Y., Loizou, P., 2008. Evaluation of objective quality measures for speech enhancement. IEEE Trans. Audio Speech Lang. Process. 16 (1), 229–238.

ITU-T P.862, 2001. Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs. ITU-T P.862 recommendation.

Kates, J.M., 1992. On using coherence to measure distortion in hearing aids. J. Acoust. Soc. Am. 91 (4), 2236–2244.

Kates, J.M., Arehart, K.H., 2005. Coherence and the speech intelligibility index. J. Acoust. Soc. Am. 117 (4), 2224–2237.

Kim, G., Loizou, P., 2010. Why do speech-enhancement algorithms not improve speech intelligibility. In: Proc. IEEE Internat. Conf. on Acoustics, Speech, and Signal Process (ICASSP), Dallas, Texas, USA, vol. 1, pp. 4738–4741.

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

Loizou, P., Kim, G., 2011. Reasons why current speech-enhancement algorithms do not improve speech intelligibility and suggested solutions. IEEE Trans. Speech Audio Process. 19 (1), 47–56.

Ludvigsen, C., Elberling, C., Keidser, G., 1993. Evaluation of a noise reduction method–comparison of observed scores and scores predicted from STI. Scand. Audiol. Suppl. 38, 50–55.

Ma, J., Hu, Y., Loizou, P., 2009. Objective measures for predicting speech intelligibility in noisy conditions based on new band-importance functions. J. Acoust. Soc. Am. 125 (5), 3387–3405.

Ma, J., Loizou, P., 2011. Snr loss: A new objective measure for predicting the intelligibility of noise-suppressed speech. Speech Commun. 53 (3), 340–354.

Moore, B., Glasberg, B., 1983. Suggested formulas for calculation auditory-filter bandwidths and excitation patterns. J. Acoust. Soc. Am. 74 (3), 750–753.

Pearce, D., Hirsch, H., 2000. The aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions. In: Proc. of Internat. Conf. on Spoken Language Processing (ICSLP), Beijing, China, pp. 29–32.

Rothauser, E., 1969. IEEE recommended practice for speech quality measurements. IEEE Trans. Audio Electroacoust. 17 (3), 225–246.

Scalart, P., Filho, J., 1996. Speech enhancement based on a priori signal to noise estimation. In: Proc. IEEE Internat. Conf. on Acoustics, Speech, and Signal Process (ICASSP), Atlanta, Georgia, USA, vol. 2, pp. 629–632.

Silverman, H., Dixon, N., 1976. A comparison of several speech-spectra classification methods. IEEE Trans. Speech Audio Process. 24, 289–295.

Steeneken, H., Houtgast, T., 1980. A physical method for measuring speech-transmission quality. J. Acoust. Soc. Am. 67 (1), 318–326.

Taal, C., Hendriks, R., Heusdens, R., Jensen, J., 2011. An algorithm for intelligibility prediction of time-frequency weighted noisy speech. IEEE Trans. Audio Speech Lang. Process. 19 (7), 2125–2136.