Speech Intelligibility Measures for Speech Enhancement under Nonstationary Acoustic Noise

R. Tavares, L. Zão and R. Coelho
Laboratory of Acoustic Signal Processing, Electrical Engineering Department
Military Institute of Engineering (IME)
Rio de Janeiro, Brazil
email:{rtavares,zao,coelho}@ime.eb.br

***Abstract*—This paper presents the study of four objective intelligibility measures to evaluate the performance of speech enhancement techniques. The objective measures have high correlation with the speech intelligibility rates obtained in subjective listening tests. Four noise reduction algorithms are applied to enhance the speech signals corrupted by real highly nonstationary acoustic noises. The results show that, for most of the noise conditions, the best intelligibility results are obtained by the speech enhancement technique based on the empirical mode decomposition. Moreover, the Wiener filtering based technique also achieves interesting intelligibility gain for the highly nonstationary noises.**

***Index Terms*—speech enhancement, objective intelligibility measures, nonstationary noise.**

I. Introduction

The degradation of speech signals due to the presence of acoustic noises is still a major problem in the speech processing area. For many decades, speech enhancement techniques have been proposed to compensate or reduce the effects of acoustic noises [1], [2], [3]. Most of them apply the short-time Fourier transform (STFT) to obtain an estimation of the noise power spectrum. The noise components are then suppressed from the noisy speech signal spectrum before the enhanced version of the speech signal is reconstructed in the time domain.

The classical noise power estimators are based on voice activity detectors (VAD) [1]. The power spectrum of the noise components is then computed at each time frame as a smoothed adaptation of its past values obtained during the speech pauses. Although such procedures show reasonable accuracy for stationary background noises, they cannot precisely estimate time-varying spectra. Other algorithms, such as the minimum statistics (MS) [4], the improved minima controlled recursive averaging (IMCRA) [5] and the unbiased minimum mean-square error (UMMSE) [6], have been proposed to deal with these nonstationary noise conditions. Thus, the estimation of the noise power spectrum is applied to each time frame even during speech activity.

Alternative speech enhancement procedures have also been proposed based on time-frequency (TF) analysis, such as the wavelets [7], [8] and the empirical mode decomposition [9], [10]. In such proposals, the TF analysis is applied to decompose the noisy speech signal, and a decision criteria identifies the least corrupted components before the reconstruction of the enhanced version of the speech signal. Different from the STFT-based methods, the TF-based ones do not require an explicit estimation of the noise statistics.

In the literature, the speech enhancement approaches are generally evaluated in terms of speech quality improvement. The segmental signal-to-noise ratio (SegSNR) is the most commonly used objective quality measure. However, the comparative study presented in [11] showed that, besides the quality improvement, the STFT-based noise-reduction algorithms are not capable of increasing the speech intelligibility. This situation becomes more challenging in nonstationary noisy scenarios due to the inaccurate noise statistics tracking [12].

This paper compares the performance of STFT-based and TF-based speech enhancement techniques in terms of speech intelligibility. The noise-reduction algorithms are applied to noisy speech signals in highly nonstationary environments. Two STFT-based noise reduction algorithms are considered in the experiments: the spectral subtraction (SS) [1] and the UMMSE [6] noise estimator followed by the Wiener filtering approach [2]. The wavelets denoising [8] and the EMD-Hurst-based [10] TF-based algorithms are also considered in the speech enhancement experiments. Four objective intelligibility measures are used in the experiments: the frequency-weighted segmental SNR (fwSegSNR) [13], the coherence speech intelligibility index (CSII) [14], the fractional articulation index (fAI) [15] and short-time objective intelligibility (STOI) measure [16]. The main focus is to evaluate the objective measures in predicting the speech intelligibility scores in highly nonstationary noise.

II. Speech Enhancement Techniques

This Section describes the four speech enhancement techniques adopted in this work. The SS and the Wiener/UMMSE solutions apply the short-time Fourier transform (STFT) to firstly obtain an estimate of the noise power spectrum. Then, the identified noise components are subtracted or compensated from the STFT of the noisy signal to improve the speech quality. In the Wavelet Denoising and EMDH techniques, the noisy speech signal is decomposed using time-frequency analysis based on wavelets and EMD, respectively. Following, the speech signal is reconstructed using only the components that are not affected by the noise.

A. Spectral Subtraction

Let $y(t)$ be a speech utterance corrupted by an additive noise $\eta(t)$. Thus, it can be written $y(t) = x(t) + \eta(t)$, where $x(t)$

978-1-4799-3743-1/14/$31.00 ©2014 IEEE

represents the clean speech signal. By applying the STFT to the above relation, it can be written

$$Y(\kappa,\tau) = X(\kappa,\tau) + \mathcal{N}(\kappa,\tau)\,, \tag{1}$$

where κ and τ are the frequency bin and the time frame indexes, respectively. The first step of the spectral subtraction [1] is to estimate the noise power spectrum $\hat{\mathcal{N}}(\kappa,\tau)$. For this purpose, a VAD is applied to the input signal and the noise spectral components are estimated by averaging the STFT of frames identified with no speech presence.

After estimating $\hat{\mathcal{N}}(\kappa,\tau)$, it is subtracted from the noisy speech power spectrum,

$$|\bar{X}(\kappa,\tau)|^2 = |Y(\kappa,\tau)|^2 - \alpha(\kappa,\tau)\cdot|\hat{\mathcal{N}}(\kappa,\tau)|^2\,, \tag{2}$$

where $\alpha(\kappa,\tau) \geq 1$ is the oversubtraction factor. The enhanced speech signal $\hat{x}(t)$ is reconstructed by overlapping and adding the inverse Fourier transform of the clean speech power spectrum.

B. Wiener/UMMSE

In the second speech enhancement procedure, the unbiased minimum mean-square error (UMMSE) noise power estimation [6] is adopted to track the noise spectrum. As in the speech enhancement experiments presented in [6], the Wiener filtering approach proposed in [2] is used to suppress the noise components from the power spectrum of the speech signal.

In the UMMSE proposal, the authors combined speech presence uncertainty to the estimator originally proposed in [17], and found that the estimation of the noise power spectrum can be updated every time frame using a recursive smoothing. Different from MS [4] and IMCRA [5], the UMMSE does not consider the minimum statistics of several past frames to estimate the noise power spectrum. It means that the UMMSE is able to track the changes in the noise power spectrum with shorter delays, which is particularly important in nonstationary noise environments.

The adoption of the Wiener filtering speech enhancement approach is due to the fact that, as the UMMSE noise estimator, it also considers that the speech and noise spectral coefficients are complex Gaussian distributed. In this work, the Wiener/UMMSE approach was implemented with the same parameters defined in [6].

C. Wavelet Denoising

In the wavelet denoising [7] adopted as the third speech enhancement approach, the noisy speech signal is firstly decomposed into a series of approximation and detail coefficients. Then, a thresholding operation is applied to the detail coefficients to shrink the wavelets components of the noisy speech signal. The idea of the wavelet thresholding is to set to zero all the components that are attributed only to noise. Finally, the enhanced speech signal is reconstructed using the high-amplitude components only.

In the wavelets denoising implementation, the soft-thresholding operator is considered with the universal threshold $T = \sigma\sqrt{2\ln N}$ [7], where N is the total number of coefficients and σ is a rough estimate of the noise level. In this work, the noise level is estimated based on the median of the absolute deviation of the detail coefficients of each decomposition level, i.e., it is a level-dependent estimation.

D. EMDH

The EMDH technique [10] adopts the empirical mode decomposition [18] and the Hurst exponent (H) [19] to enhance the noisy speech signal. The EMD is a nonlinear time-domain adaptive method for decomposing a signal $x(t)$ into a series of oscillatory intrinsic mode functions (IMF) and a residual. The EMD is adopted due to two main advantages over the wavelet decomposition. Firstly, the wavelets-based approach is based on a set of pre-defined basis functions, which does not necessarily fits well to all kinds of signals. Moreover, the wavelet decomposition is not adaptable to local or temporary variations in the input signal. On the other hand, the EMD analyzes the speech signal in an entirely adaptive way, and it is completely based on the local properties of the input signal. It makes the EMD suitable for nonstationary signal analysis.

In the EMDH technique, after decomposing the noisy speech signal with the EMD, the Hurst exponent [19] is computed from each frame of the resulting IMFs to determine which of them are mainly composed by noise. Following the procedure in [10], the Hurst exponent is estimated and the threshold for the Hurst value is set to $H_{th} = 0.9$. It means that each frame of the speech signal is reconstructed using only the IMFs whose Hurst value follows $H < 0.9$. Finally, all the reconstructed frames are concatenated to obtain the enhanced version of the speech signal.

III. Speech Objective Intelligibility Measures

This Section briefly introduces the four objective measures adopted in this work to evaluate the speech enhancement algorithms in terms of speech intelligibility. Although a subjective listening test is considered to be the best approach to evaluate the intelligibility of speech signals, such tests are costly and time consuming [16]. Thus, the fwSegSNR [13], the CSII [14], the fAI [15] and the STOI [16] are selected in this work since they present high correlation to the intelligibility scores (% of correct words or sentences) obtained in subjective listening tests.

A. fwSegSNR

The first step to compute the frequency-weighted segmental SNR [13] is to obtain the spectra of the clean ($|X(j,\tau)|$) and enhanced ($|\hat{X}(j,\tau)|$) speech signals. It is achieved by the division of their entire bandwidth into $K = 25$ frequency bands using Gaussian-shaped filters. Then, the fwSegSNR is given by

$$\text{fwSegSNR} = \frac{10}{Q}\sum_{\tau=0}^{Q-1}\frac{\sum_{j=1}^{K} W(j,\tau)\log\frac{|X(j,\tau)|^2}{\left(|X(j,\tau)|-|\hat{X}(j,\tau)|\right)^2}}{\sum_{j=1}^{K} W(j,\tau)}\,, \tag{3}$$

where τ and j are the frame and frequency band indexes, respectively. In this work, the weights $W(j,\tau)$ in (3) are defined by $W(j,\tau) = |X(j,\tau)|^{(0.2)}$, since it leads to the highest correlation between the fwSegSNR and the intelligibility scores in subjective listening tests [12].

B. CSII

The CSII measure [14] was proposed as an extension to the speech intelligibility index (SII), standard ANSI S3.5-1997. The SII evaluates the SNR in each frequency band

978-1-4799-3743-1/14/$31.00 ©2014 IEEE

of the enhanced speech signal. The frequency bands are defined according to a critical band formulation that models the auditory periphery. The intelligibility estimate is computed as a weighted sum of the SNR values across all the frequency bands. Since in the SII the power spectrum of speech and noise are computed using long-term averages, it presents good accuracy for stationary noises, but it is inaccurate for nonstationary noises.

In the CSII proposal [14], the standard speech SNR estimate is replaced by the signal-to-distortion ratio (SDR) computed from the magnitude-squared coherence (MSC). In order to achieve a higher correlation to the intelligibility scores obtained in subjective listening tests, the authors in [14] computed the CSII separately for low-, medium- and high-level segments of each sentence. The highest correlation was obtained with a weighted sum of the corresponding measures (CSII_{Low}, CSII_{Med} and $\text{CSII}_{\text{High}}$) considering the following weights

$$\text{CSII} = 0.155\text{CSII}_{\text{Low}} + 0.845\text{CSII}_{\text{Med}} + 0.0\text{CSII}_{\text{High}}\,. \quad (4)$$

In this work, the CSII_{Low}, CSII_{Med} and $\text{CSII}_{\text{High}}$ are obtained according to the description in [14].

C. fAI

The articulation index (AI) [20] is the most commonly used measure to predict speech intelligibility. It is based on the principle that the intelligibility depends on the proportion of spectral information of the speech that is audible to the listener. To compute the AI, the speech spectrum is divided into 20 bands that are considered to equally contribute to the intelligibility. Then, the SNRs in each band are averaged using weighting functions defined by band-importance functions.

The fAI was proposed in [15] to overcome some limitations of the articulation index, mainly the fact that it cannot be used to handle nonlinear processing with additive noise, such as the spectral subtraction [1]. Moreover, the AI cannot be applied when the corrupting noise is nonstationary. In the fAI proposal [15], a new definition of the output SNR (fraction SNR - fSNR) is presented to handle with nonlinear noise-reduction techniques and derive a new intelligibility measure. The new SNR definition is particularly important when the nonlinear processing affects predominantly the speech signal rather than the noise. Finally, the fAI is computed as the weighted sum of fSNR values computed across all bands, considering the same weights defined by the band-importance functions as in the AI.

D. STOI

The short-time objective intelligibility measure [16] was proposed as a correlation-based method to evaluate the speech intelligibility degradation caused by the speech enhancement procedures. The STOI results presented in [16] showed its high and closest correlation with the subjective intelligibility rates obtained with speech signals enhanced by noise-reduction algorithms.

The first step of the STOI is to obtain the STFT of the clean and the noisy versions of the speech signal, i.e., $X(\kappa,\tau)$ and $Y(\kappa,\tau)$, respectively. Then, $X(\kappa,\tau)$ and $Y(\kappa,\tau)$ are grouped in 15 one-third octave bands. Denoting $\|X_{(j,\tau)}\|$ as the ℓ^2 norm of the vector of the STFT components that belongs to the j^{th} band ($j = 1, 2, \ldots, 15$), the temporal envelope vector $\mathbf{x}_{(j,\tau)}$ of the clean speech is given by

$$\mathbf{x}_{(j,\tau)} = \left[\|X_{(j,\tau-29)}\|, \|X_{(j,\tau-28)}\|, \ldots, \|X_{(j,\tau)}\|\right]. \quad (5)$$

The temporal envelope vector $\mathbf{y}_{(j,\tau)}$ of the noisy speech is obtained in analogy to (5). The intermediate intelligibility measure, $\text{STOI}_{(j,\tau)}$, is defined as the correlation coefficient between $\mathbf{x}_{(j,\tau)}$ and the normalized version of $\mathbf{y}_{(j,\tau)}$. Finally, the STOI measure is given by averaging the intermediate values over the 15 one-third octave bands and all Q speech frames.

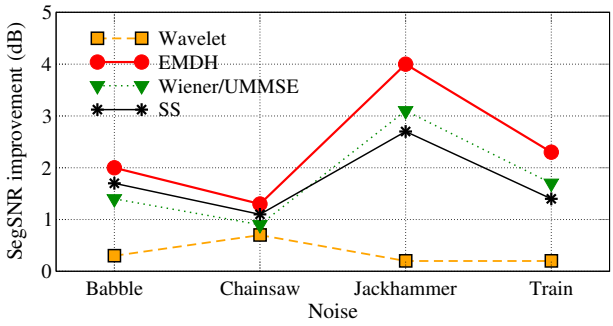

Fig. 1. The SegSNR improvement obtained for each of the acoustic noise sources averaged over the SNR values: -10 dB, -5 dB, 0 dB, 5 dB and 10 dB.

IV. Speech Enhancement Intelligibility Experiments

The speech enhancement experiments are conducted with a subset of 24 speakers (16 male and 8 female) randomly selected from the TIMIT speech database [21]. It leads to a total of 240 speech segments, 10 per speaker, with sampling rate of 16 kHz and average time duration of 3 seconds. Four highly nonstationary acoustic noises (Babble, Chainsaw, Jackhammer and Train) are used to corrupt the speech signals considering five SNR values: -10 dB, -5 dB, 0 dB, 5 dB and 10 dB. The noises are collected from the NOISEX-92 [22] (Babble), the Freesound.org[1] (Train) and the Freesfx.co.uk[2] (Chainsaw and Jackhammer) databases.

The speech enhancement techniques are firstly evaluated in terms of segmental SNR (SegSNR), to serve as a reference to the objective intelligibility measures. The SegSNR is the most commonly used measure to assess the quality of enhanced speech signals. Fig. 1 presents the SegSNR improvement, in dB, averaged across the five values of SNR. The improvement is computed as the difference between the SegSNR result obtained from the enhanced signal and that obtained from the noisy (unprocessed) speech signal. Note that the EMDH achieves the highest SegSNR gain for the four noise sources. Regarding the STFT-based techniques, the SS outperforms the Wiener/UMMSE for the Babble and Chainsaw noises. On the other hand, the Wiener/UMMSE leads to better SegSNR gain for the Jackhammer and Train noises. Finally, the wavelet denoising achieves the lowest improvement for all the noise sources.

[1] http://www.freesound.org.
[2] http://www.freesfx.co.uk.

978-1-4799-3743-1/14/$31.00 ©2014 IEEE

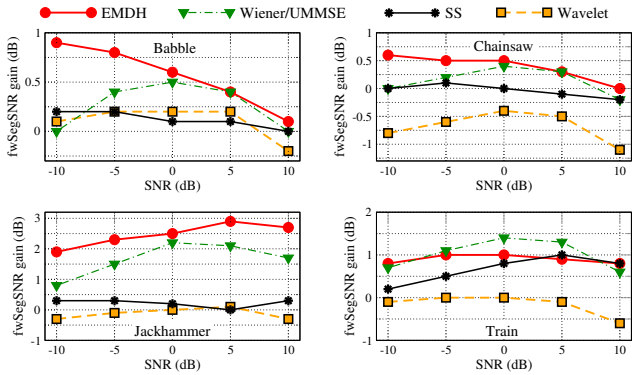

Fig. 2. fwSegSNR improvement (dB) obtained with the four speech enhancement techniques considering the four acoustic noise sources.

A. *fwSegSNR Results*

The fwSegSNR results obtained from the speech enhancement procedures are depicted in Fig. 2. The fwSegSNR showed high correlation with the sentence recognition scores in subjective listening tests in the experiments presented in [12]. The results are presented in terms of fwSegSNR gain, computed in the same manner as the SegSNR improvement (Fig. 1).

It can be observed from Fig. 2 that the EMDH leads to the best results for most of the noise conditions. The exceptions occur for the Train noise, for which the Wiener/UMMSE achieves the highest fwSegSNR gain for three SNR values: -5 dB, 0 dB and 5 dB. It is important to note that, opposed to the SegSNR results (refer to Fig. 1), the Wiener/UMMSE technique outperforms the SS in terms of fwSegSNR for most of the noise conditions, even for the Babble and Chainsaw noises. This conclusion is in line with the results presented in [12], that showed that the largest improvement in speech quality, as obtained with SS, does not imply the speech intelligibility results. Once more, the wavelet denoising presents the worst fwSegSNR results.

B. *Intelligibility Prediction Results*

Different from the fwSegSNR, ehe CSII, fAI and STOI measures are used to predict the intelligibility scores in subjective listening tests. For this purpose, logistic functions are used to transform the objective measure results into predicted percentage of correcly recognized words.

The CSII and STOI results are transformed via following logistic function

$$f(d) = 100/(1 + \exp(a\,d + b))\,, \qquad (6)$$

where d represents the corresponding objective measure. In this work, the coefficients a and b are determined to fit the objective measures to the intelligibility scores obtained in subjective listening tests presented in [11].

The prediction intelligibility results (percentage of correctly recognized words) obtained with the CSII and the STOI are presented in Tabs. I and II. Both CSII and STOI measures indicate that the EMDH technique leads to the highest average results for the Babble and Chainsaw noises. For the other noise sources, the Wiener/UMMSE achieve the best intelligibility prediction results. It is interesting to note that the predicted scores obtained with either the Wiener/UMMSE or the EMDH techniques are always higher than those due to the SS or the wavelets denoising. This can be explained by the fact that both the EMDH and the Wiener/UMMSE consider the occurrence of nonstationary noises in their formulation, while the SS and the wavelet denoising do not. Moreover, the intelligibility prediction scores with SS are generally higher than those obtained with the wavelet denoising. However, different from the fwSegSNR results, the wavelet denoising

TABLE I
INTELLIGIBILITY RATE PREDICTION (%) OBTAINED WITH THE COHERENCE SPEECH INTELLIGIBILITY INDEX (CSII).

Noise	SNR	SS	Wiener	Wavelet	EMDH
Babble	10	93.0	92.2	84.9	92.2
	5	68.1	71.4	57.4	71.5
	0	29.1	34.6	22.2	36.4
	-5	9.7	10.0	6.7	12.6
	-10	3.0	2.8	2.0	4.0
	Aver.	40.6	42.2	34.6	43.4
Chainsaw	10	82.5	80.6	80.1	82.3
	5	44.0	45.6	45.9	49.1
	0	14.7	15.3	15.0	18.8
	-5	4.3	4.2	3.9	5.8
	-10	1.7	1.6	1.5	2.1
	Aver.	29.4	29.5	29.3	31.6
Jackhammer	10	96.1	97.5	96.0	97.4
	5	76.6	91.6	87.8	90.5
	0	50.4	74.3	61.7	66.9
	-5	23.5	40.3	24.6	30.1
	-10	9.1	14.5	6.9	10.4
	Aver.	51.1	63.7	55.4	59.1
Train	10	97.9	97.7	95.2	97.6
	5	92.3	91.9	84.4	90.8
	0	69.0	72.0	57.8	67.8
	-5	28.8	36.9	22.8	32.8
	-10	8.6	11.8	6.5	11.4
	Aver.	59.3	62.1	53.3	60.1

TABLE II
INTELLIGIBILITY RATE PREDICTION (%) OBTAINED WITH THE SHORT-TIME OBJECTIVE INTELLIGIBILITY (STOI) MEASURE.

Noise	SNR	SS	Wiener	Wavelet	EMDH
Babble	10	89.6	88.8	82.6	89.0
	5	69.8	72.0	62.3	73.4
	0	28.8	37.6	25.4	42.0
	-5	5.9	9.2	5.5	12.5
	-10	1.1	1.6	1.2	2.6
	Aver.	39.0	41.8	35.4	43.9
Chainsaw	10	86.6	85.7	86.9	88.2
	5	55.1	57.0	61.3	61.8
	0	16.7	19.3	23.5	25.1
	-5	2.8	3.5	3.8	5.0
	-10	0.7	1.1	1.0	1.5
	Aver.	32.4	33.3	35.3	36.3
Jackhammer	10	90.7	92.9	92.4	92.7
	5	71.0	85.9	84.7	86.9
	0	37.9	72.2	68.2	72.4
	-5	13.9	44.1	39.2	42.4
	-10	5.2	17.4	13.4	16.5
	Aver.	43.7	62.5	59.6	62.2
Train	10	90.8	90.2	89.0	90.0
	5	81.6	80.8	76.3	81.0
	0	60.1	63.5	53.9	63.6
	-5	23.4	35.9	25.7	35.3
	-10	4.8	11.2	6.6	10.8
	Aver.	52.1	56.3	50.3	56.1

978-1-4799-3743-1/14/$31.00 ©2014 IEEE

TABLE III
INTELLIGIBILITY RATE PREDICTION (%) OBTAINED WITH THE FRACTIONAL ARTICULATION INDEX (FAI).

Noise	SNR	SS	Wiener	Wavelet	EMDH
Babble	10	94.5	94.3	93.5	95.3
	5	83.3	86.0	85.0	88.7
	0	44.0	60.6	58.8	65.9
	-5	9.6	23.4	23.0	28.3
	-10	0.6	3.9	3.6	5.8
	Aver.	46.4	53.6	52.8	56.8
Chainsaw	10	96.1	96.4	93.9	97.4
	5	88.5	92.3	86.5	94.1
	0	64.3	80.5	69.3	86.3
	-5	36.3	61.8	41.3	70.6
	-10	6.3	27.6	13.4	40.3
	Aver.	58.3	71.7	60.9	77.8
Jackhammer	10	96.9	97.8	96.5	98.5
	5	87.3	96.1	94.2	97.4
	0	57.1	92.2	88.2	94.8
	-5	18.6	81.0	73.5	87.8
	-10	4.2	53.2	43.2	65.6
	Aver.	52.8	84.1	79.1	88.8
Train	10	97.8	97.5	96.4	97.9
	5	95.9	95.4	93.7	96.2
	0	89.6	89.8	86.9	92.0
	-5	62.0	73.0	68.2	78.6
	-10	20.1	39.5	34.2	47.7
	Aver.	73.1	79.1	75.9	82.5

technique outperforms the SS for the Jackhammer noise.

For the fAI, the logistic function adopted to predict the intelligibility scores is the same as proposed in [15],

$$f(d) = (1 - 10^{-dP/Q})^2 \,, \tag{7}$$

with $P = 27.5$ and $Q = 8.4$. The predicted intelligibility scores obtained with the fAI measures are presented in Tab. III. For the fAI measure, the EMDH outperforms the other baseline techniques for all the noise conditions. Once again, the Wiener/UMMSE technique outperforms the SS approach. On the other hand, the spectral subtraction achieves the lowest average predicted intelligibility scores for all the noise sources.

V. CONCLUSION

This paper evaluated the performance of different speech enhancement techniques with four objective intelligibility measures. The idea of using the objective measures is to avoid the need for subjective listening tests, which are costly and time consuming. The experiments are conducted in highly nonstationary noise scenarios. Two of the speech enhancement techniques are based on the use of the short-time Fourier transform, and the other two apply time-frequency analysis to the noisy speech signals. All the four objective measures agree in the sense that the best intelligibility results are achieved with the Wiener/UMMSE and the EMDH techniques, which were originally proposed under the hypothesis that noises are nonstationary. The fwSegSNR results showed that the EMDH technique achieved the best improvement for most of the noise conditions. The intelligibility rate prediction based on the CSII, fAI and STOI measures reinforced the superior performance of the EMDH technique, especially for the Babble and Chainsaw noises. Finally, the CSII and the STOI indicated that the Wiener/UMMSE technique is very promising on achieving good intelligibility scores for the Jackhammer and Train noises.

REFERENCES

[1] S. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 27, pp. 113–120, April 1979.

[2] P. Scalart and J. Filho, "Speech enhancement based on a priori signal to noise estimation," *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, vol. 32, pp. 629–632, December 1996.

[3] I. Cohen and B. Berdugo, "Speech enhancement for non-stationary noise environments," *Signal Processing*, vol. 81, pp. 2403 – 2418, November 2001.

[4] R. Martin, "Noise power spectral density estimation based on optimal smoothing and minimum statistics," *IEEE Transactions on Speech and Audio Processing*, vol. 9, pp. 504–512, July 2001.

[5] I. Cohen, "Noise spectrum estimation in adverse environments: improved minima controlled recursive averaging," *IEEE Transactions on Speech and Audio Processing*, vol. 11, pp. 466–475, September 2003.

[6] T. Gerkmann and R. Hendriks, "Unbiased MMSE-based noise power estimation with low complexity and low tracking delay," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, pp. 1383–1393, May 2012.

[7] D. Donoho and I. Johnstone, "Threshold selection for wavelet shrinkage of noisy data," *Proceedings of the 16th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC'94)*, vol. 1, pp. A24–A25, November 1994.

[8] D. Donoho, "De-noising by soft-thresholding," *IEEE Transactions on Information Theory*, vol. 41, no. 3, pp. 613–627, 1995.

[9] K. Khaldi, A. Boudraa, A. Bouchikhi, and M. Alouane, "Speech enhancement via EMD," *EURASIP Journal on Advances in Signal Processing*, vol. 2008, p. 873204, May 2008.

[10] L. Zão, R. Coelho, and P. Flandrin, "Speech enhancement with emd and hurst-based mode selection," *IEEE/ACM Transactions on Audio, Speech and Language Processing*, vol. 22, pp. 897–909, May 2014.

[11] P. Loizou and Y. Hu, "A comparative intelligibility study of single-microphone noise reduction algorithms," *J. Acoust. Soc. Am.*, vol. 22, pp. 1777–1786, September 2007.

[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, pp. 3387–3405, May 2009.

[13] Y. Hu and P. Loizou, "Evaluation of objective quality measures for speech enhancement," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 16, pp. 229–238, January 2008.

[14] J. Kates and K. Arehart, "Coherence and the speech intelligibility index," *J. Acoust. Soc. Amer.*, vol. 117, pp. 2224–2237, April 2005.

[15] P. Loizou and J. Ma, "Extending the articulation index to account for non-linear distortions introduced by noise-suppression algorithms," *J. Acoust. Soc. Am.*, vol. 130, pp. 986–995, August 2011.

[16] C. Taal, R. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of time-frequency weighted noisy speech," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 19, pp. 2125–2136, September 2011.

[17] R. Hendriks, R. Heusdens, and J. Jensen, "MMSE based noise PSD tracking with low complexity," in *IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP 2010)*, pp. 4266–4269, March 2010.

[18] N. Huang, Z. Shen, S. Long, M. Wu, H. Shih, Q. Zheng, N. Yen, C. Tung, and H. Liu, "The empirical mode decomposition and the hilbert spectrum for nonlinear and non-stationary time series analysis," *Proceedings of the Royal Society of London. Series A: Mathematical, Physical and Engineering Sciences*, vol. 454, pp. 903–995, March 1998.

[19] E. Hurst, "Long-term storage capacity of reservoirs," *American Society of Civil Engineers Trans.*, pp. 770–799, April 1951.

[20] K. Kryter, "Methods for the calculation and use of the articulation index," *The Journal of the Acoustical Society of America*, vol. 34, pp. 1689–1697, July 1999.

[21] J. Garofolo, L. Lamel, W. Fisher, J. Fiscus, D. Pallett, N. Dahlgren, and V. Zue, "TIMIT acoustic-phonetic continuous speech corpus," *Linguistic Data Consortium, Philadelphia*, 1993.

[22] A. Varga and H. Steeneken, "Assessment for automatic speech recognition: II. NOISEX-92: a database and an experiment to study the effect of additive noise on speech recognition systems," *Speech Communications*, vol. 12, pp. 247–251, July 1993.

978-1-4799-3743-1/14/$31.00 ©2014 IEEE