Speech enhancement based on stationary bionic wavelet transform and maximum a posterior estimator of magnitude-squared spectrum

Talbi Mourad[1]

Received: 27 August 2016 / Accepted: 18 October 2016 / Published online: 5 November 2016

Abstract Numerous efforts have focused on the problem of reducing the impact of noise on the performance of various speech systems such as speech coding, speech recognition and speaker recognition. These approaches consider alternative speech features, improved speech modeling, or alternative training for acoustic speech models. In this paper, we propose a new speech enhancement technique, which integrates a new proposed wavelet transform which we call stationary bionic wavelet transform (SBWT) and the maximum a posterior estimator of magnitude-squared spectrum (MSS-MAP). The SBWT is introduced in order to solve the problem of the perfect reconstruction associated with the bionic wavelet transform. The MSS-MAP estimation was used for estimation of speech in the SBWT domain. The experiments were conducted for various noise types and different speech signals. The results of the proposed technique were compared with those of other popular methods such as Wiener filtering and MSS-MAP estimation in frequency domain. To test the performance of the proposed speech enhancement system, four objective quality measurement tests [signal to noise ratio (SNR), segmental SNR, Itakura–Saito distance and perceptual evaluation of speech quality] were conducted for various noise types and SNRs. Experimental results and objective quality measurement test results proved the performance of the proposed speech enhancement technique. It provided sufficient noise reduction and good intelligibility and perceptual quality, without causing considerable signal distortion and musical background noise.

Keywords Bionic wavelet transform · Stationary bionic wavelet transform · Maximum a posterior estimator of magnitude-squared spectrum · Speech enhancement

1 Introduction

Enhancing speech signal corrupted by uncorrelated additive noise is still remaining as a challenging task for researchers due to shortcoming of existing speech enhancement techniques in real world noise conditions. The noise presence affects the performance of speech processing systems. These systems include voice coders, speech recognition, hearing aids and mobile phones. The speech enhancement objective is to improve the intelligibility and perceptual quality of speech by minimizing the effect of noise. Existing techniques for this task include Wiener filtering (Deller et al. 2000; Haykin 1996), spectral subtraction (Deller et al. 2000; Boll 1979), wavelet transform (WT) (Seok and Bae 1997; Bahoura and Rouat 2001, 2006; Cohen 2001; Lu and Wang 2003; Chen and Wang 2004; Hu and Loizou 2004), etc.

An emerging tendency in the speech enhancement domain consists of using a filter bank based on a specific psychoacoustic model of human auditory system (critical bands). The principle behind this is based on the fact that embedding the model of psychoacoustic of human auditory system in filter bank can improve the perceptual quality and the intelligibility of speech. Moreover, it is well known that the human auditory system can roughly be described as a non-uniform band-pass filter bank and humans are capable to detect the original speech signal in noisy environments without noise prior knowledge (Taşmaz and Erçelebi 2008). Different frequency transformations (scales) are proposed for considering the hearing perceptive aspect (Mel, Bark, ERB, and so on). It is worth

✉ Talbi Mourad
mouradtalbi196@yahoo.fr

[1] Center of Researches and Technologies of Energy of Borj Cedria, Tunis, Tunisia

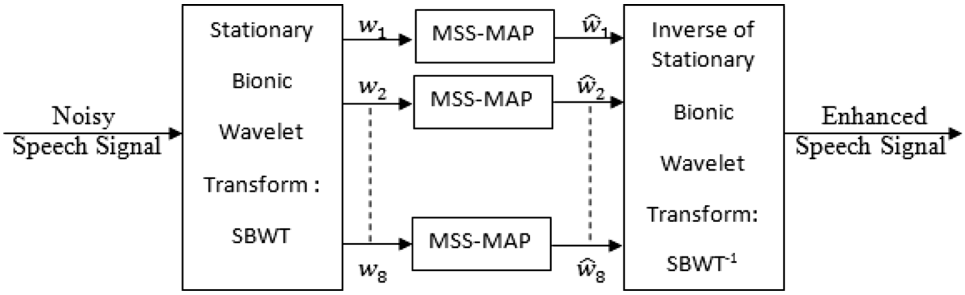


Fig. 1 The block diagram of the proposed technique

mentioning that the majority of the perceptual speech enhancement approaches are based on the wavelet packet transform (Johnson et al. 2007). Moreover, the wavelet packet transforms were efficiently combined with others denoising methods for improving the performance of speech enhancement techniques based on wavelets. Therefore, many hybrid speech enhancement systems used both WT and others tools such as Wiener filtering (Mahmoudi 1997), spectral subtraction (Shao and Chang 2007) and Ephraim and Malah approach (Taşmaz and Erçelebi 2008). Daqrouq et al. (2010) have investigated the utilization of wavelet filters via multistage convolution by reverse biorthogonal wavelets in high and low pass band frequency parts of speech signal. Speech signal is decomposed into two pass bands of frequency; high and low, and then the noise is removed in each band individually in different stages via wavelet filters. This approach provides better outcomes because it does not cut the speech information, which occurs when utilizing conventional thresholding (Daqrouq et al. 2010). In Vaz et al. (2013) was proposed a method for speech enhancement of data collected in extremely noisy environments, such as those found during magnetic resonance imaging (MRI) scans. Vaz et al. (2013) have proposed a two-step algorithm to perform this noise suppression. First, they used probabilistic latent component analysis in order to learn dictionaries of the noise and (speech + noise) portions of the data and used these to factor the noisy spectrum into estimated speech and noise components. Second, they applied a wavelet packet analysis in conjunction with a wavelet threshold that minimizes the KL divergence between the estimated speech and noise to achieve further noise suppression (Vaz et al. 2013).

In this paper, we propose a new technique of noise reduction and speech enhancement. This technique integrates a new proposed WT which we call stationary bionic wavelet transform (SBWT) and the maximum a posterior estimator of magnitude-squared spectrum (MSS-MAP) (Yang and Loizou 2011). According to (Yang and Loizou 2011), statistical estimators of the magnitude-squared spectrum (MSS) are derived based on the assumption that the MSS of the noisy speech signal can be computed as the sum of the (clean) signal and noise magnitude-squared spectra. maximum a posterior (MAP) and minimum mean square error (MMSE) estimators are derived based on a Gaussian statistical model. The gain function of the MAP estimator was found to be the same as the gain function used in the ideal binary mask that is extensively used in computational auditory scene analysis. As such, it was binary and assumed the value of 1 if the local signal-to-noise ratio (SNR) exceeded 0 dB, and assumed the value of 0 otherwise. By modeling the local instantaneous SNR as an F-distributed random variable, soft masking techniques were derived integrating SNR uncertainty. The soft masking technique, in particular, which weighted the noisy magnitude-squared spectrum by the a priori probability that the local SNR exceeds 0 dB. The obtained results in (Yang and Loizou 2011) was shown to be identical to the Wiener gain function. The obtained results in (Yang and Loizou 2011) indicated that the estimators proposed in (Yang and Loizou 2011) yielded significantly better speech quality than the conventional minimum mean square error spectral power estimators, in terms of yielding lower residual noise and lower speech degradation. Concerning the SBWT (Talbi and Aicha 2014), it is introduced in order to solve the problem of the perfect reconstruction associated with the bionic wavelet transform (BWT). The MSS-MAP estimation (Yang and Loizou 2011) was used for estimation of speech in the SBWT domain.

The rest of this paper is organized as follows: Sect. 2 describes the proposed speech enhancement technique by giving a detailed overview of the SBWT and the different steps followed in this technique. In Sect. 3, we will deal with MSS-MAP in SBWT domain. Section 4 is devoted to the evaluation metrics. In Sect. 5 are presented results and discussions. Finally, the conclusion is given in Sect. 6.

2 The proposed technique

In this work, we propose a new speech enhancement technique, which integrates a new proposed wavelet transform which we call SBWT and the MSS-MAP. The SBWT is introduced in order to solve the problem of the perfect reconstruction associated with the BWT. The MSS-MAP estimation was used for speech estimation in the SBWT domain. The block diagram of the proposed technique is presented in Fig. 1.

As shown in Fig. 1, this proposed technique consists at first step in applying the SBWT to the noisy speech signal. Then each of the obtained noisy stationary bionic wavelet coefficients, $\mathrm{w_i}, 1 \leq \mathrm{i} \leq 8$, is denoised separately in order to obtain eight denoised stationary bionic wavelet coefficients, $\hat{\mathrm{w}}_\mathrm{i}, 1 \leq \mathrm{i} \leq 8$. The denoising of each coefficient $\mathrm{w_i}, 1 \leq \mathrm{i} \leq 8$ is perfomed by using the technique based on the MSS-MAP estimation (Yang and Loizou 2011). Finally, the denoised speech signal is obtained from the application of the inverse of SBWT, SBWT^{-1}, to the denoised coefficients, $\widehat{\mathrm{w}}_\mathrm{i}, 1 \leq \mathrm{i} \leq 8$.

2.1 The bionic wavelet transform

Yao and Zhang (2001) have proposed the *BWT* as an adaptive wavelet transform designed specifically to model the human auditory system. The term 'bionic' means that the *BWT* is rooted in an active biological mechanism (Johnson et al. 2007). In addition, the *BWT* decomposition is both perceptually scaled and adaptive (Johnson et al. 2007). The initial perceptual aspect of this transform comes from the logarithmic spacing of the baseline scale variables, which are designed to match basilar membrane spacing (Johnson et al. 2007). Then, two adaptation factors control the time-support used at each scale, based on a non-linear perceptual model of the auditory system (Johnson et al. 2007). The basis of this transform is the Giguerre–Woodland non-linear transmission line model of the auditory system (Giguere 1993; Giguere and Woodland 1994), an active-feedback electro-acoustic model incorporating the auditory canal, middle ear and cochlea (Johnson et al. 2007). The model yields estimates of resistance and the time-varying acoustic compliance along the displaced basilar membrane, as a physiological acoustic mass function, cochlear frequency-position mapping, and feedback factors representing the active mechanisms of outer hair cells. The net result can be seen as a technique for estimating the time-varying quality factor Q_{eq} of the cochlear filter banks as a function of the input sound waveform. Giguere and Woodland (1994), Zheng et al. (1999), and Yao and Zhang (2002) have given all details on the elements of this model. The adaptive nature of the Bionic Wavelet Transform is insured by a time-varying linear factor $T(a,\tau)$ which represents the scaling of the cochlear filter bank quality factor Q_{eq} at each scale over time. Incorporating this directly into the scale factor of a Morlet mother wavelet, the following formula is obtained:

$$X_{BWT}(a,\tau) = \frac{1}{T(a,\tau)\sqrt{a}} \int x(t)\tilde{\varphi}^{*}\left(\frac{t-\tau}{a \cdot T(a,\tau)}\right) e^{-jw_0\left(\frac{t-\tau}{a}\right)} dt \tag{1}$$

Where a and τ represent respectively scale and time shift variables and $\tilde{\varphi}$ is expressed as follow:

$$\tilde{\varphi}(t) = e^{-\left(\frac{t}{T_0}\right)^2} \tag{2}$$

The function $\tilde{\varphi}(t)$ is the amplitude envelope of the Morlet mother wavelet and the factor w_0 represents the base fundamental frequency of the unscaled mother wavelet. Here this parameter is taken as $w_0 = 15,165.4\ Hz$ for the human auditory system, per the original work of Yao and Zhang (2002). The factor T_0 represents the initial time-support. The discretization of the scale variable a is performed using pre-determined logarithmic spacing across the desired frequency range, in order that the center frequency at each scale is expressed as follow (Johnson et al. 2007):

$$w_m = w_0/(1.1623)^m, m = 0, 1, 2, \ldots \tag{3}$$

For this implementation, based on original work of Yao and Zhang for cochlear implant coding (Yao and Zhang 2002), coefficients at 22 scales, $m = 7, \ldots, 28$, are computed using numerical integration of the continuous wavelet transform. These 22 scales correspond to center frequencies logarithmically spaced from 225 to 5300 Hz. The adaptation factor $T(a,\tau)$ for each time and scale is calculated using the following formula (Johnson et al. 2007):

$$T(a,\tau+\Delta\tau) = \frac{1}{\left(1 - G_1 \frac{C_s}{C_s + |X_{BWT}(a,\tau)|}\right) \cdot \left(1 + G_2 \left|\frac{\partial}{\partial\tau} X_{BWT}(a,\tau)\right|\right)} \tag{4}$$

where G_1 is the active gain factor representing the outer hair cell active resistance function, G_2 is the active gain factor representing the time-varying compliance of the Basilar membrane, and $C_s = 0.8$ is a constant that represents non linear saturation effects in the cochlear model (Johnson et al. 2007). Practically speaking, the partial derivative of Eq. (4) is approximated using the first difference of the previous points of the BWT at that scale (Johnson et al. 2007). From the Eq. (1), we can see that the duration of the amplitude envelope of the wavelet is affected by the factor $T(a,\tau)$ which does not affect the frequency of the associated complex exponential. Therefore, one useful manner for thinking of the BWT is as a mechanism for adapting the time support of the underlying wavelet according to the quality factor Q_{eq} of the corresponding cochlear filter model at each scale. Yao and Zhang (2002) have proved that the bionic coefficients, $X_{BWT}(a,\tau)$ can be computed as a product of the original *WT* coefficients $X_{WT}(a,\tau)$ and a constant $K(a,\tau)$ which is a function of the adaptation factor $T(a,\tau)$. For the Morlet mother wavelet, this adaptive multiplying factor can be formulated as follow:

$$X_{BWT}(a,\tau) = K(a,\tau)X_{WT}(a,\tau) \tag{5}$$

with

$$K(a,\tau) = \frac{\sqrt{\pi}}{C}\frac{T_0}{\sqrt{1+T^2(a,\tau)}} \tag{6}$$

where C is a normalizing constant calculated from the integral of the squared mother wavelet. This representation yields an efficient computational technique for calculating BWT coefficients directly from the original WT coefficients without requiring at each scale and time, to compute numerical integration of Eq. (1) (Johnson et al. 2007). There are diverse key differences between a filterbank based wavelet packet transform (WPT) using an orthonormal wavelet such as the Daubechies family, as used for the comparative baseline technique and the discretized continuous wavelet transform (CWT) using the Morlet mother wavelet, used for the BWT. One is that the WPT is perfectly reconstructable, while the discretized CWT is an approximation whose exactness depends on the placement and number of frequency bands selected. Another difference is that the frequency support of the orthonormal wavelet families used for WPTs and DWTs covers a broader bandwidth while the Morlet wavelet consists of a single frequency with an exponentially decaying time support. The Morlet mother wavelet is thus more “frequency focused” along each scale, which is what allows the direct adaptation of the time support, the central mechanism of the adaptation of the BWT.

2.2 Stationary bionic wavelet transform (SBWT)

As previously mentioned, in this work, we have used in our speech enhancement system, a new wavelet transform which we call SBWT. This new transform is obtained by replacing the discretized CWT used in the BWT computation, by the stationary wavelet transform (SWT). In Fig. 2, are given the different steps of the SBWT computation and also the steps of the computation of its inverse, SBWT^{-1}. According to this figure, the stationary bionic wavelet coefficients are obtained by multiplying the stationary wavelet coefficients by the K factor (Eq. (6)). These stationary wavelet coefficients are obtained from the application of the SWT to the input signal. The steps of the SBWT computation are the same steps followed in the BWT computation but the unique difference consists in replacing the discretized CWT by SWT. The reconstructed signal is obtained by multiplying at first step, the stationary bionic wavelet coefficients by 1/K and then applying the SWT^{-1} to the resulting coefficients.

In the implementation of *SWT* and *SBWT*, we have used the Daubechies mother wavelet with ten vanishing moments (https://www.nag.co.uk/numeric/MB/manual_22_1/pdf/C09/c09aa.pdf).

In Tables 1 and 2, are listed the values of $max(|x-y|)$ between the original speech signal, x and the reconstructed

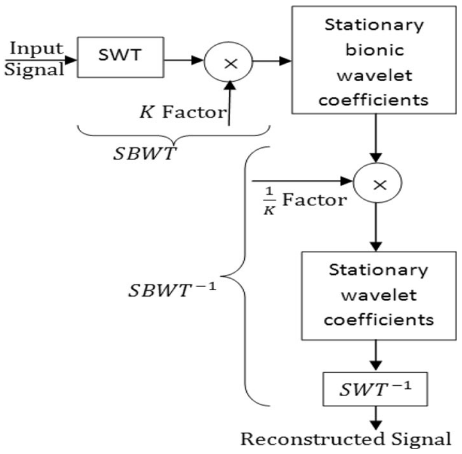


Fig. 2 The stationary bionic wavelet transform (*SBWT*) and its inverse, (*SBWT*$^{-1}$)

Table 1 Case of female voice

Speech signal	$max(\lvert x-y \rvert)$		
	SBWT	BWT	
Scale number	8	22	30 (Talbi et al. 2010)
Signal1	7.0342e-06	0.0694	0.0676
Signal2	9.7201e-06	0.1428	0.1429
Signal3	1.5658e-05	0.1877	0.0700
Signal4	1.4170e-05	0.2062	0.0705
Signal5	1.4137e-05	0.0527	0.0418
Signal6	1.1788e-05	0.1633	0.1614
Signal7	1.4955e-05	0.2305	0.2294
Signal8	1.0856e-05	0.1629	0.0636
Signal9	1.2150e-05	0.1585	0.1014
Signal10	2.1509e-05	0.0677	0.0623

Table 2 Case of male voice

Speech signal	$max(\lvert x-y \rvert)$		
	SBWT	BWT	
Scale number	8	22	30 (Talbi et al. 2010)
Signal1	1.7974e-05	0.1897	0.0667
Signal2	1.4011e-05	0.2449	0.1523
Signal3	1.1984e-05	0.1983	0.1205
Signal4	1.4847e-05	0.1893	0.0430
Signal5	1.1492e-05	0.3015	0.0730
Signal6	0.0068	0.2495	0.1389
Signal7	1.7819e-05	0.2730	0.1255
Signal8	1.4949e-05	0.1897	0.1340
Signal9	1.4087e-05	0.1550	0.0713
Signal10	1.2989e-05	0.1743	0.0875

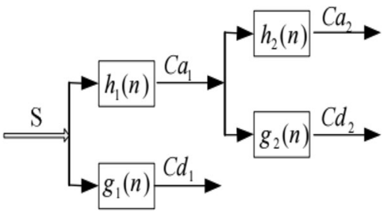


Fig. 3 Filter bank implementation of SWT

speech signal, y obtained after application of the *BWT* or the SBWT and its inverse. The original signal *x* is obtained by applying the MSS-MAP (Yang and Loizou 2011) to the noisy speech signal (Fig. 4). The Fig. 4 shows the different steps of the procedure followed in this paper to verify the perfect reconstruction of the transform, *BWT* or *SBWT*.

2.2.1 Stationary wavelet transform (SWT)

In both discrete wavelet transform (DWT) and WPT, after filtration the coefficients will down sampled, that prevents redundancy and allow using the same pair of filter in different levels. And so, these transforms will suffer from the lack of shift invariance, which means that small shifts in the input signal can cause major variations in the distribution of energy between coefficients at deferent levels and may causes some error in reconstruction (Mortazavi and Shahrtash 2008). This problem is carried out by eliminating the down sampling steps after filtration at each level in SWT. By eliminating down sampling, the number of coefficients at each level is as long as original signal. Figure 3 shows decomposition of a signal by SWT up two levels. In decomposition of a signal through a filter bank, if down sampling operators were eliminated, for the next level of decomposition the high and low pass filters must be modified. For this, the low pass and high pass filters at each level will be up sampled by putting zero between each filter's coefficients of previous level that called a trous algorithm (Mortazavi and Shahrtash 2008; Shensa 1992). Denoising a signal by SWT has the same three steps as DWT (Mortazavi and Shahrtash 2008).

It is worth mentioning that for computing the Error, max (|x − y|) and verifying the perfect reconstruction of the two transforms (BWT and SBWT), we first have enhanced the speech signal by MSS-MAP based technique (Yang and Loizou 2011). The application of this technique is performed because the clean speech signal is generally not available but we know only the noisy speech signal. So to compute the error between the original signal and the reconstructed signal, we first have to suppress the noise corrupting this original signal and we have chosen MSS-MAP (Yang and Loizou 2011) for this aim.

In Fig. 4, the noisy speech signal is obtained by corrupting the clean speech signal by the noise. which is selected to be the car noise with SNR = 10 dB. Hence the values listed in Tables 1 and 2, are obtained in that case. These values show clearly that the use of SBWT permits to have a lower Error between the original signal x and the reconstructed signal y, than that obtained in case of using BWT. The latter introduces some distortions on the reconstructed speech signals compared to the original speech signals and this especially when the number of scales is N = 22. For the BWT, the error between the original signal and the reconstructed signal (Table 1), is reduced when using N = 30 instead of N = 22.

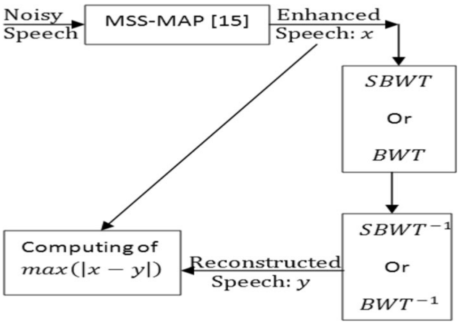

Noisy Speech
MSS-MAP [15]
Enhanced Speech: x
$SBWT$ Or BWT
$SBWT^{-1}$ Or BWT^{-1}
Reconstructed Speech: y
Computing of $max(|x-y|)$

Fig. 4 The procedure of verifying the perfect reconstruction of the wavelet transform (BWT or SBWT)

3 Maximum a posterior estimator of magnitude-squared spectrum in SBWT domain

Generally, conventional speech enhancement techniques based on thresholding in wavelet domain may introduce some degradation on the original speech signal. This especially occurs for unvoiced sounds. Therefore many speech enhancement systems based on wavelets use others tools such as Wiener filtering, spectral subtraction and MMSE-STSA estimation (Taşmaz and Erçelebi 2008; Ephraim and Malah 1984). The latter is used with the undecimated wavelet packet-perceptual filterbanks in the speech enhancement system proposed by Taşmaz and Erçelebi (2008). In that system (Taşmaz and Erçelebi 2008), is first performed the perceptual filterbank (CB-UWP) (critical bands–undecimated wavelet package) decomposition of the degraded speech signal by applying the undecimated wavelet packet perceptual transform to this signal. Seventeen critical sub-bands are obtained from this decomposition and this is done by referring to psychoacoustic model (Taşmaz and Erçelebi 2008). Each of these critical sub-bands is denoised by using the speech enhancement technique proposed by Ephraim and Malah (1984). The estimation of the clean speech signal is finally obtained by the CB-UWP reconstruction from the denoised subband signals. This speech enhancement principle proposed in (Taşmaz and Erçelebi 2008), is used in this work (Fig. 1) and the CB-UWP decomposition is replaced by the SBWT decomposition and the MMSE-STSA estimation is replaced by MSS-MAP estimation. Such as in the speech enhancement system proposed in (Taşmaz and Erçelebi 2008), each of stationary bionic wavelet coefficient, $\mathrm{w_i}, 1 \leq \mathrm{i} \leq 8$ (Fig. 1) obtained from the application of SBWT to the noisy speech signal, is processed as a noisy speech signal and is denoised using MSS-MAP introduced by Yang and Loizou (2011).

As previously mentioned the SBWT is introduced to solve the problem of the perfect reconstruction associated with BWT. Moreover, the SBWT among all wavelet transforms (Biswas et al. 2014; Singh and Mutawa 2016), tends to uncorrelated data (Bahoura and Rouat 2006) and simplifies noise cancellation. Moreover, the application of MSS-MAP in SBWT domain (Fig. 1) for denoising the noisy sub-bands, $\mathrm{w_i}, 1 \leq \mathrm{i} \leq 8$, introduces better adaptation for noise and speech estimations compared to the application of the MSS-MAP to the entire noisy speech signal. All these facts motivate us to propose this new speech enhancement technique (SBWT/MSS-MAP).

4 The evaluation metrics

To test the performance of the proposed speech enhancement technique, the objective quality measurement tests, SNR, segmental signal-to-noise ratio (SSNR), Itakura–Saito distance and perceptual evaluation of speech quality (PESQ), were used.

4.1 Signal-to-noise ratio

The following formula was used to calculate the SNR of enhanced speech signals:

$$SNR(dB) = 10 \cdot log_{10}\left(\frac{\sum_{n=0}^{N-1} x^2(n)}{\sum_{n=0}^{N-1} (\hat{x}(n) - x(n))^2}\right) \quad (7)$$

where x(n) and $\hat{\mathrm{x}}(\mathrm{n})$ are respectively, the original and the enhanced signals and N is the number of samples in the original signal.

4.2 Segmental signal to noise ratio

The frame based segmental SNR is an objective measure of speech quality. It is computed by averaging frame level estimates as follows:

$$SSNR(dB) = \frac{1}{M}\sum_{m=0}^{M-1} 10 \cdot log_{10}\left(\frac{\sum_{n=N_m}^{N_m+N-1} x^2(n)}{\sum_{n=N_m}^{N_m+N-1} (\hat{x}(n) - x(n))^2}\right) \quad (8)$$

where $x(n)$ and $\hat{x}(n)$ represent respectively the original and the enhanced signals, M is the number of frames, N is the number of samples in each short time frame and $\mathrm{N_m}$ is the beginning of the m-th frame. Since the SNR can become very small and negative during silence periods, the SSNR values are limited to the range of [−10, 35 dB].

4.3 Itakura–Saito distance

The Itakura–Saito distance measure, based on the dissimilarity between the clean and the enhanced speech, is calculated between sets of linear prediction coefficients (LPC) estimated over synchronous frames. This measure is greatly affected by spectral dissimilarity due to mismatch in formant locations, with little contribution from errors in matching spectral valleys. Such behavior is desirable since the auditory system is more sensitive to errors in formant position and bandwidth than to spectral valleys between peaks. In this work, the average Itakura–Saito measure (as defined by Eq. (9)) across all speech frames of a given sentence, was calculated for evaluating the speech enhancement technique.

$$\mathrm{ISd(a,b)} = \left((\mathrm{a}-\mathrm{b})^{\mathrm{T}}\mathrm{R}(\mathrm{a}-\mathrm{b})\right)/\left(\mathrm{a^T R a}\right) \quad (9)$$

where a and b represent respectively the LPC of the clean speech signal and the LPC of the enhanced speech signal $\hat{x}(\mathrm{n})$ and R represents the matrix of autocorrelation. The symbol T represents the transpose symbol.

4.4 Perceptual evaluation of speech quality

The perceptual evaluation of speech quality (PESQ) algorithm is an objective quality measure that is approved as the ITU-T recommendation P.862 (Rix et al. 2001). It is a tool of objective measurement introduced to predict the results of a subjective mean opinion score (MOS) test. It was proved (Hu and Loizou 2008; Zavarehei et al. 2006) that the PESQ correlated better with MOS than the traditional objective speech measures.

5 Results and discussions

In this section, ten Arabic speech sentences produced by a female speaker and ten others are produced by a male speaker. These sentences are artificially corrupted in additive manner with different noise types (white, F16 cockpit, Tank, Pink and Car noises) at different values of SNR. These noises were taken from the AURORA database (Hirsch and Pearce 2000). The used Arabic sentences (Table 3) are material phonetically balanced and they are sampled at 16 kHz.

Table 3 The list of the used Arabic speech sentences

Arabic speech sentences	
Female speaker	Male speaker
أحفظ من الأرض	يذيع الخبر لا لن
أين المسا فرين	أكمل بالإسلام رسالتك
لا لم يستمتع بثمرها	سقطت إبرة
سيؤذيهم زماننا	من لم ينتفع
كنت قدوة لهم	غفل عن ضحكاتها
ازار صائما	و لماذا نشف مالهم
كال و غبط الكبش	أين زوايانا و قانوننا
هل لذعته بقول	صاد الموروث مدلعا
عرف واليا و قائدا	نبه آبائكم
خالا بالنا منكما	أظهره و قم

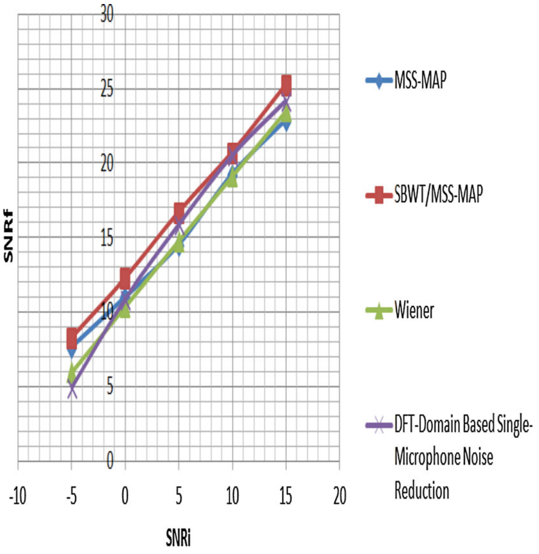


Fig. 5 Speech signal corrupted by volvo noise

The noisy speech signals were enhanced by using the proposed technique (SBWT/MSS-MAP), the technique based on MSS-MAP estimation (Yang and Loizou 2011), the Wiener Filtering (Loizou 2007) and the speech enhancement technique based on discrete fourier transform (DFT), proposed in (Hendriks et al. 2013).

Figures 5, 6, 7 and 8 show the curves obtained from the SNR, the SSNR, the Itakura–Saito distance (ISd) and PESQ computations for the different techniques: the technique based on MSS-MAP estimation (Yang and Loizou 2011), the proposed technique (SBWT/MSS-MAP), Wiener Filtering (Deller et al. 2000; Haykin 1996) and DFT-domain based single microphone noise reduction (Hendriks et al. 2013).

The results obtained from SNR computation and in case of Volvo noise corrupting the speech signal, show that all speech enhancement techniques improve the SNR

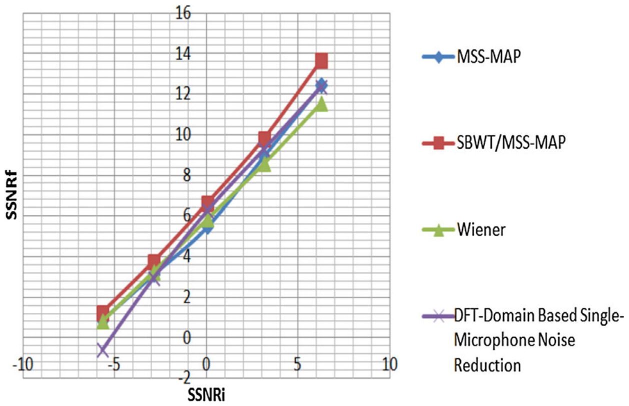


Fig. 6 Speech signal corrupted by volvo noise

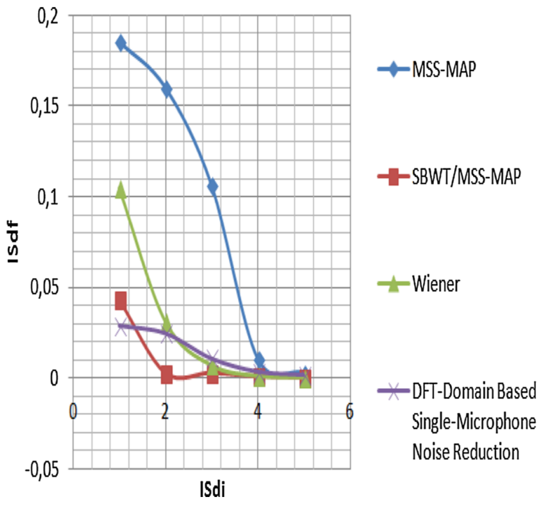


Fig. 7 Speech signal corrupted by volvo noise

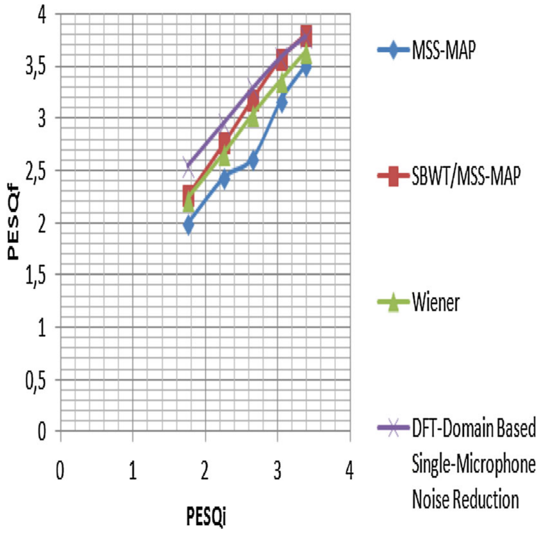


Fig. 8 Speech signal corrupted by volvo noise

(SNRf > SNRi). Moreover, the proposed technique outperforms all the techniques used in our evaluation precisely the technique based on MSS-MAP (Yang and Loizou 2011) and the technique DFT domain based single-microphone noise reduction (Hendriks et al. 2013) which in turn outperforms the two others techniques: MSS-MAP and Wiener filtering.

The results obtained from SSNR computation and in case of volvo noise corrupting the speech signal, show that all speech enhancement techniques improve the SSNR (SSNRf > SSNRi). Moreover, the proposed technique outperforms all the techniques used in our evaluation precisely the technique based on MSS-MAP (Yang and Loizou 2011) and technique DFT domain based single-microphone noise reduction (Hendriks et al. 2013) which in turn outperforms the two others techniques: MSS-MAP and Wiener.

According to the results obtained from ISd computation and in case of Volvo noise corrupting the speech signal, the proposed speech enhancement technique (SBWT/MSS-MAP) gives the lowest values of ISd compared to others techniques. Therefore in term of ISd, the proposed technique (SBWT/MSS-MAP) outperforms the three others techniques: MSS-MAP, Wiener and DFT-domain based single noise reduction (Hendriks et al. 2013).

According to the results obtained from PESQ computation and in case of Volvo noise corrupting the speech signal, the proposed technique (SBWT/MSS-MAP) and the technique DFT-domain based single noise reduction (Hendriks et al. 2013), outperform the two others technique: Wiener and MSS-MAP. For the higher values of SNRi, the values of the PESQ after enhancement (PESQf), obtained from the application of the proposed technique (SBWT/MSS-MAP), are almost the same the values obtained from the application of the technique DFT-domain based single noise reduction (Hendriks et al. 2013). Whereas For the lower values of SNRi, the technique DFT-domain based single noise reduction (Hendriks et al. 2013) outperforms the proposed technique (SBWT/MSS-MAP).

The Fig. 9 illustrates an example of speech enhancement using the proposed technique.

This figure shows clearly that the proposed technique efficiently reduces the noise while preserving the quality of the original speech signal.

The evaluation of the different techniques [SBWT/MSS-MAP, MSS-MAP (Yang and Loizou 2011) and DFT-domain based single-microphone noise reduction (Hendriks et al. 2013)], is also performed on a speech sentence taken from TIMIT Database and corrupted by the noise. This speech sentence is the English sentence.

“She had your dark suit in greasy wash water all year” and is pronounced by a female voice. This sentence is corrupted by car noise with different values of SNR.

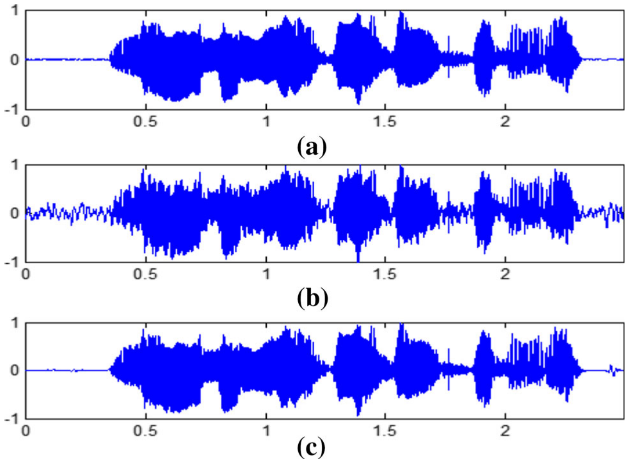


Fig. 9 An example of denoising speech signal corrupted by car noise: **a** clean speech, **b** noisy speech (SNR = 10 dB), **c** denoised speech signal using the proposed technique (SBWT/MSS-MAP)

In Tables 4, 5, 6 and 7, are listed the results obtained from the computation of the SNR, the SSNR, the ISd and the PESQ and this for the case of volvo noise.

The results obtained from SNR, SSNR and ISd computation (Tables 1, 2 and 3) show that the proposed technique (SBWT/MSS-MAP) outperforms the two techniques: MSS-MAP (Yang and Loizou 2011) and DFT-

Table 4 SNR computation (case of volvo noise)

SNRi (dB)	SNRf (dB)		
	Method		
	MSS-MAP (Yang and Loizou 2011)	(SBWT/ MSS-MAP)	DFT-Domain based single-microphone noise reduction (Hendriks et al. 2013)
−5	9.1904	12.16981	6.7524
0	13.7894	16.4734	11.5116
5	18.3689	18.8809	15.8715
10	22.6764	23.1859	21.4899
15	26.2160	26.8156	26.3789

Table 5 SSNR computation (case of volvo noise)

SSNRi (dB)	SSNRf (dB)		
	Method		
	MSS-MAP (Yang and Loizou 2011)	(SBWT/ MSS-MAP)	DFT-Domain based single-microphone noise reduction (Hendriks et al. 2013)
−6.3572	7.1139	7.2774	4.3178
−3.2400	11.0879	11.2451	8.3123
0.4822	15.1413	15.4914	12.0685
4.8450	18.7229	19.1495	17.7213
9.1621	21.8488	22.3828	21.9582

Table 6 ISd computation(case of volvo noise)

ISdi	ISdf		
	Method		
	MSS-MAP (Yang and Loizou 2011)	(SBWT/ MSS-MAP)	DFT-Domain based single-microphone noise reduction (Hendriks et al. 2013)
0.1009	0.0171	0.0026	0.0397
0.0855	0.0031	2.5812e-04	0.0103
0.0572	4.1817e-04	3.6254e-04	9.0442e-04
0.0231	1.3195e-04	9.3145e-05	1.1662e-04
0.0050	3.3776e-05	1.6663e-05	8.4648e-06

Table 7 PESQ computation (case of Volvo noise)

PESQi	PESQf		
	Method		
	MSS-MAP (Yang and Loizou 2011)	(SBWT/ MSS-MAP)	DFT-Domain based single-microphone noise reduction (Hendriks et al. 2013)
2.7811	3.1591	3.2993	3.4530
3.1403	3.4478	3.5466	3.7164
3.5639	3.7728	3.8505	3.9573
3.8282	3.9647	3.9998	4.1910
4.2065	4.0719	4.0517	4.2520

domain based single-microphone noise reduction (Hendriks et al. 2013).

The results obtained from PESQ computation (Table 4) show that the DFT-domain based single-microphone noise reduction (Hendriks et al. 2013) outperforms the two techniques: the proposed technique (SBWT/MSS-MAP) and the MSS-MAP technique (Yang and Loizou 2011).

We have also used others speech signals and an other denoising technique in our evaluation. This technique is supervised and online nonnegative matrix factorization (NMF) based noise reduction and was proposed in (Mohammadiha et al. 2013). Figures 11, 12, 13 and 14 show the different curves obtained from the SNR, the SSNR, the ISd and PESQ computations for the different values of SNR before speech enhancement. These results are obtained from the application of the proposed technique (SBWT/MSS-MAP) and the others three techniques [the DFT domain based single-microphone noise reduction technique (Hendriks et al. 2013), the technique MSS-MAP (Yang and Loizou 2011) and supervised and online NMF based noise reduction technique (Mohammadiha et al. 2013; Girish et al. 2015)] to a speech signal (Fig. 10) corrupted by different types of noise. This speech signal is sampled at 16,000 Hz and pronounced in English language by a male voice.

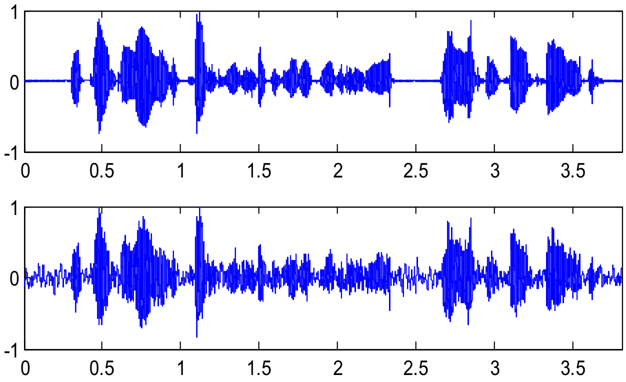


Fig. 10 An example of speech signal corrupted by volvo noise and used for evaluating the four techniques including the proposed one (SBWT/MSS-MAP)

According to the curves in Fig. 11 and in term of SNR computing, when the SNR before denoising, SNRi is higher, the proposed technique outperforms the others denoising techniques. Although, when the SNRi is lower, the best technique is supervised and online NMF based noise reduction technique (Mohammadiha et al. 2013).

According to the curves in Fig. 12 and in term of segmental SNR computing, the proposed technique outperforms the others denoising techniques.

According to the curves in Fig. 13 and in term of ISd computing, the proposed technique and MSS-MAP based one (Yang and Loizou 2011) outperform the others denoising techniques.

According to the curves in Fig. 14 and in term of PESQ computing, when the perceptual evaluation of speech quality before denoising (PESQi) is higher, the DFT Domain based single-microphone noise reduction technique (Hendriks et al. 2013) outperforms the others

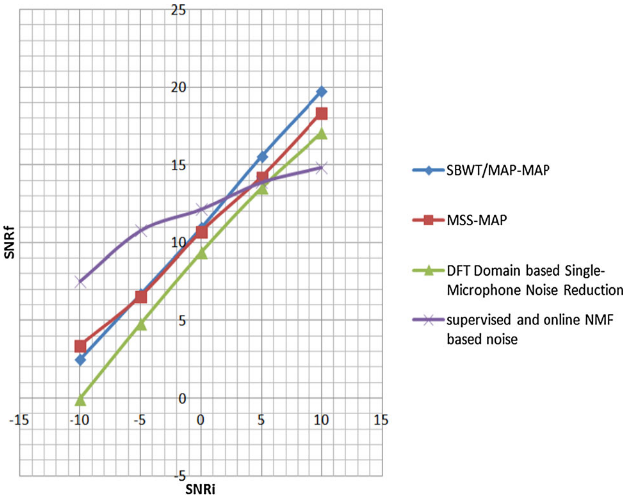


Fig. 11 Signal to noise ratio after denoising (SNRf) versus signal to noise ratio before denoising (SNRi): case of a speech signal (Fig. 10) corrupted by volvo noise

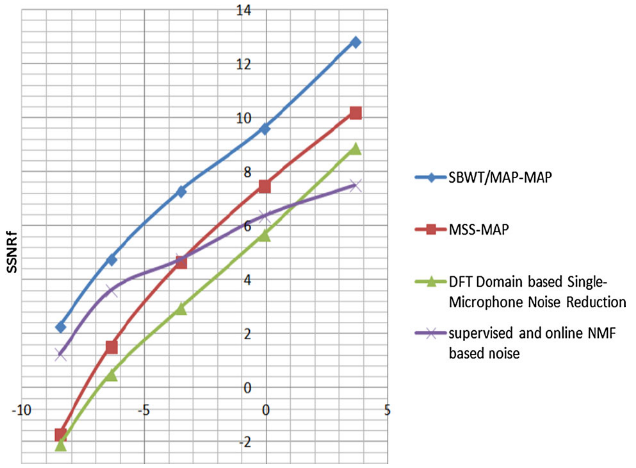


Fig. 12 Segmental signal to noise ratio after denoising (SSNRf) versus segmental signal to noise ratio before denoising (SSNRi): case of a speech signal (Fig. 9) corrupted by volvo noise

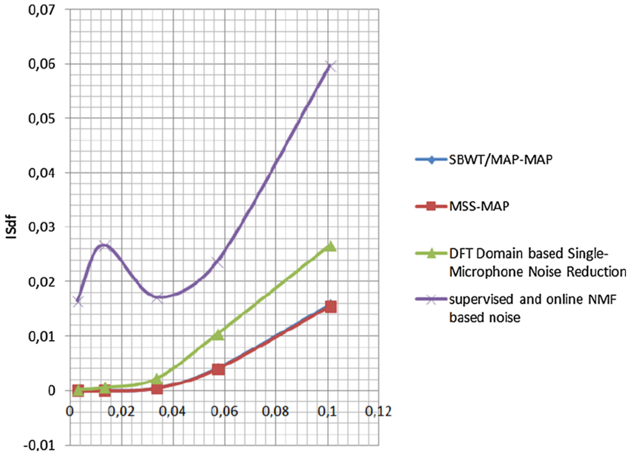


Fig. 13 Itakura–Saito distance (ISdf) versus Itakura–Saito distance (ISdi): case of a speech signal (Fig. 9) corrupted by volvo noise

denoising techniques. Although, when the PESQi is lower, the supervised and online NMF based noise reduction technique (Mohammadiha et al. 2013) outperforms the others techniques. In higher values of PESQi, the proposed technique is better than the two techniques MSS-MAP (Yang and Loizou 2011) and supervised and online NMF based noise reduction (Mohammadiha et al. 2013).

Figures 15, 16, 17 and 18 show others examples of speech enhancement using the proposed technique.

Where SNRi and SNRf are respectively signal to noise ratios before and after enhancement. SSNRi and SSNRf are respectively segmental signal to noise ratios before and after

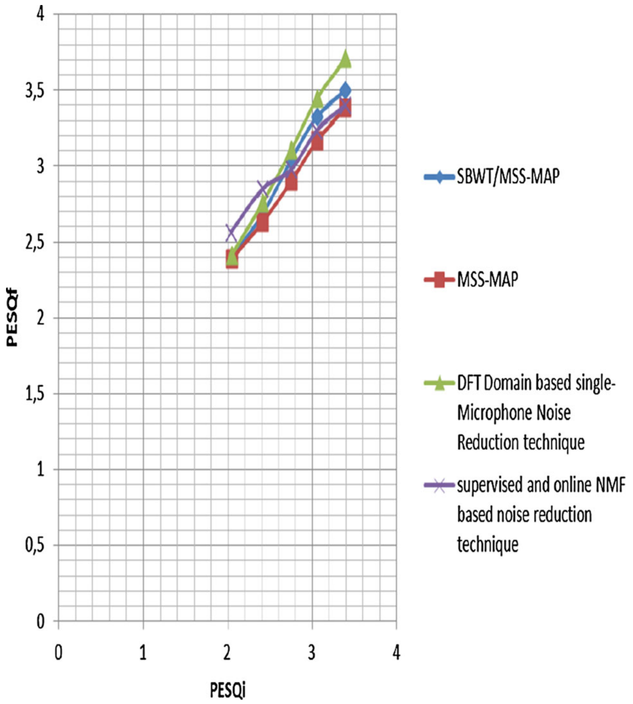


Fig. 14 Perceptual evaluation of speech quality after denoising (PESQf) versus perceptual evaluation of speech quality before denoising (PESQi): case of a speech signal (Fig. 9) corrupted by volvo noise

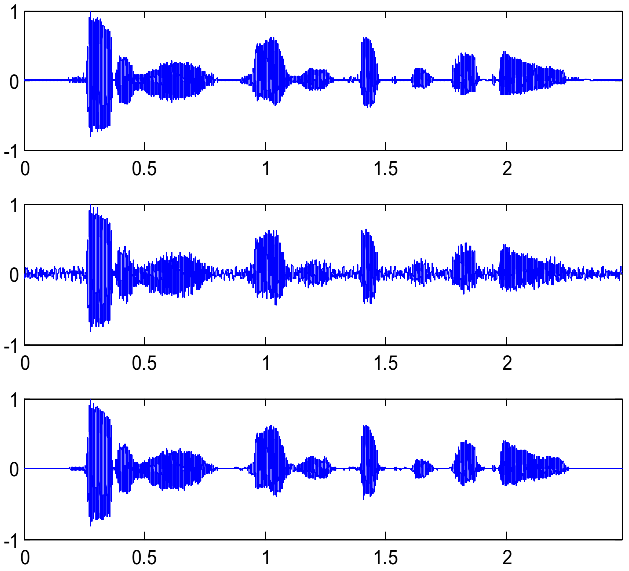


Fig. 15 A speech signal taken from Timit Database and corrupted by tank noise, enhanced by the proposed technique (SNRi = 10 dB, SNRf = 16.7383 dB, SSNRi = 1.7965 dB, SSNRf = 7.6179 dB, ISdi = 0.0182, ISdf = 3.7397e-04, PESQi = 2.6675, PESQf = 3.1143)

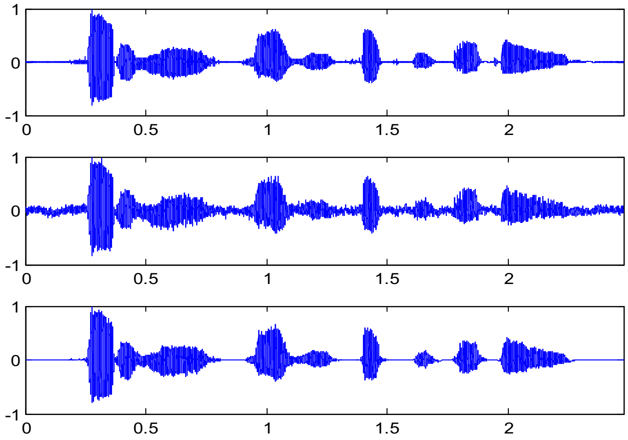


Fig. 16 A speech signal taken from Timit Database and corrupted by pink noise, enhanced by the proposed technique (SNRi = 10 dB, SNRf = 15.0956 dB, SSNRi = 1.5896 dB, SSNRf = 6.2249 dB, ISdi = 0.0768, ISdf = 0.0495, PESQi = 2.2660, PESQf = 2.7800)

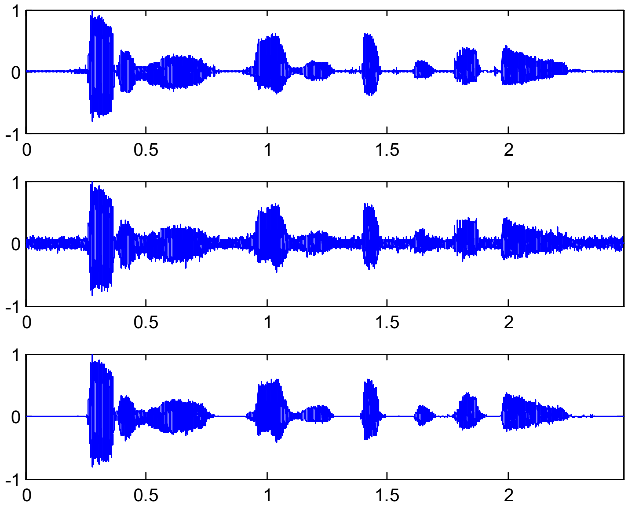


Fig. 17 A speech signal taken from Timit Database and corrupted by white noise and enhanced by the proposed technique (SNRi = 10 dB, SNRf = 14.5035 dB, SSNRi = 1.4850 dB, SSNRf = 6.0776 dB, ISdi = 0.5621, ISdf = 0.0495, PESQi = 2.0519, PESQf = 2.7304)

enhancement. ISdi and ISdf are respectively Itakura–Saito distances before and after enhancement. PESQi and PESQf are respectively PESQ before and after enhancement.

Figures 19 and 20 illustrate another example of speech denoising using the proposed technique (SBWT/MSS-MAP). In Fig. 20 are illustrated the spectrograms of the clean speech signal, the noisy speech signal and the enhanced speech signal.

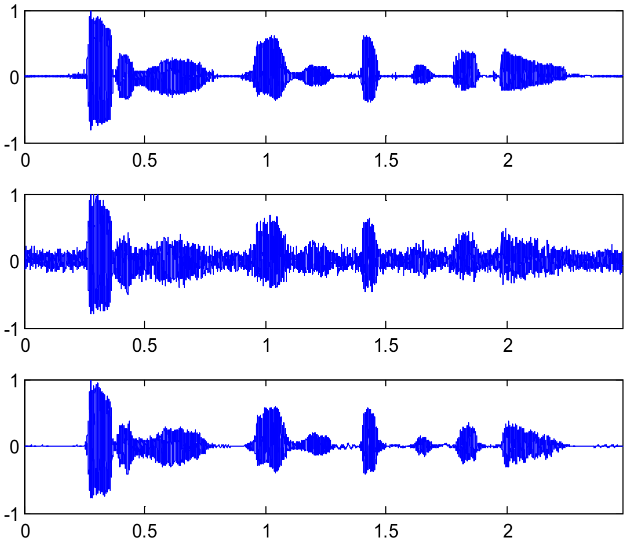


Fig. 18 A speech signal taken from Timit Database and corrupted by F16 noise, enhanced by the proposed technique (SNRi = 5 dB, SNRf = 11.4539 dB, SSNRi = 1.7233 dB, SSNRf = 3.2526 dB, ISdi = 0.4625, ISdf = 0.4826, PESQi = 1.8480, PESQf = 2.4521)

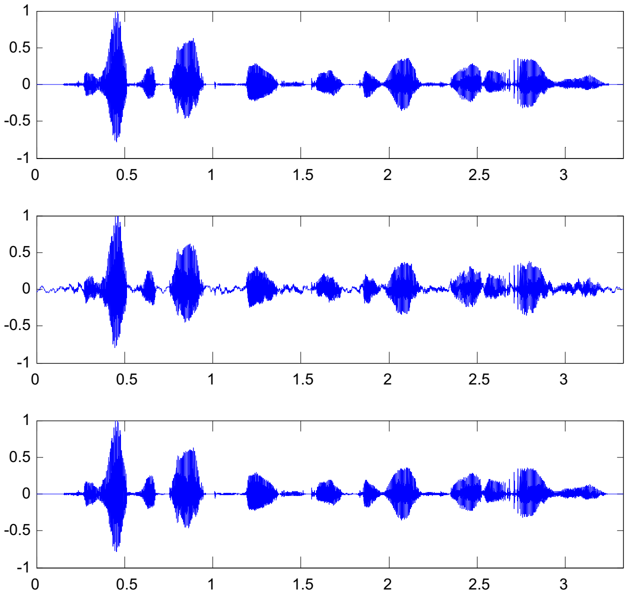
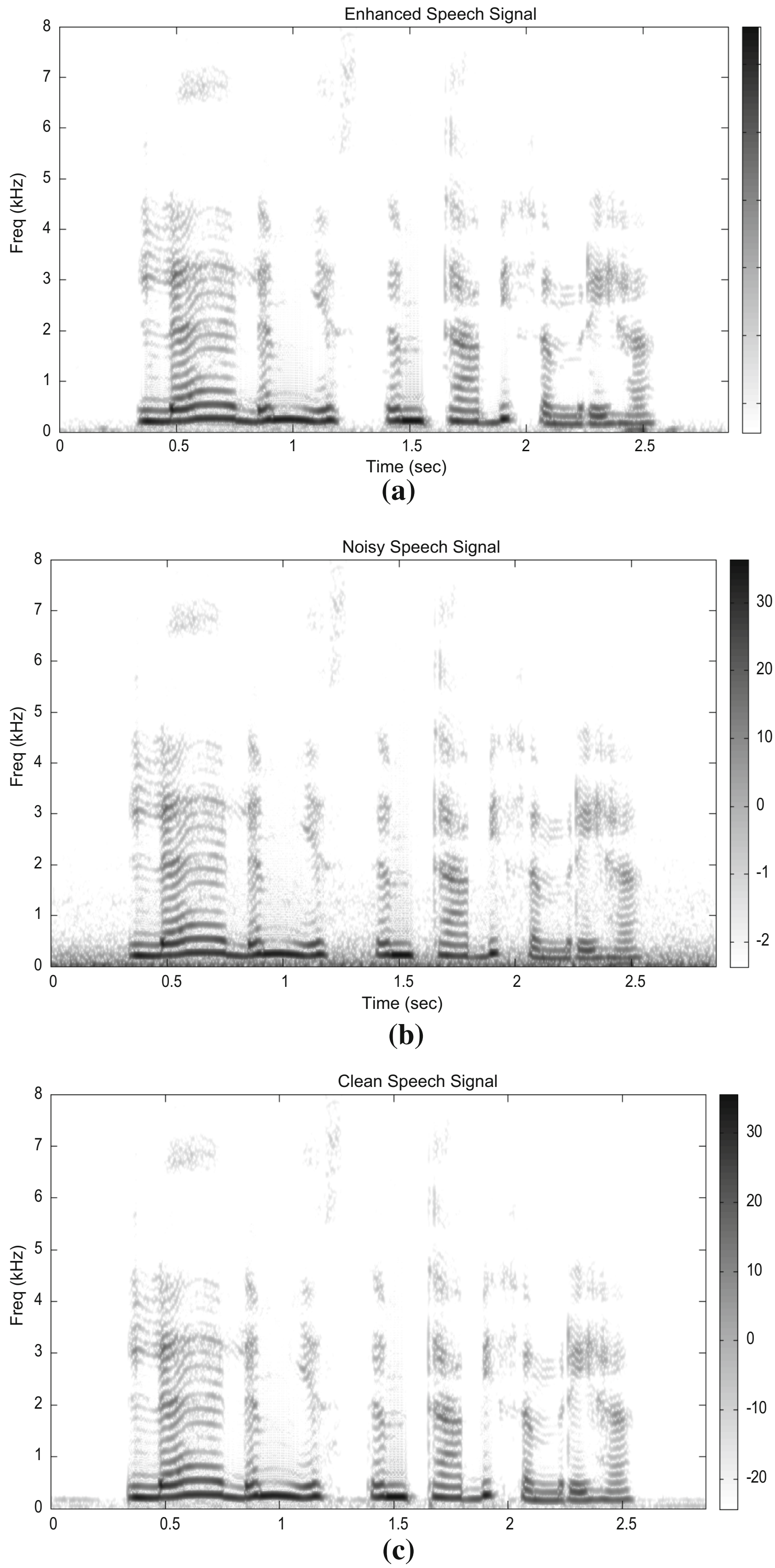


Fig. 19 An example of speech enhancement using the proposed technique (SBWT/MSS-MAP): Denoising of speech signal (taken from Timit Database) corrupted by volvo noise with SNR = 10 dB

The spectrogram (b) shows that the type of noise corrupting the speech signal is low pass because it is localized in low frequencies regions. The spectrogram (c) shows that the noise which is a car noise, is suppressed efficiently by using the proposed technique (SBWT/MSS-MAP).

Fig. 20 **a** The spectrogram of the clean speech signal. **b** The spectrogram of the noisy speech signal (speech signal corrupted by car noise with SNR = 10 dB). **c** The spectrogram of the enhanced speech signal



(a)

(b)

(c)

6 Conclusion

In this paper, we proposed a new speech enhancement technique, which integrates a new proposed wavelet transform (which we call SBWT) and the MSS-MAP. The SBWT is introduced in order to solve the problem of the perfect reconstruction associated with the BWT. The MSS-MAP estimation was used for estimation of speech in the SBWT domain. The performance of the proposed technique (SBWT/MSS-MAP) was compared to that of the techniques based on MSS-MAP estimation, the Wiener Filtering, the speech enhancement technique based on DFT and the supervised and online NMF based noise reduction technique. The evaluation was based on four objective metrics: SNR, SSNR, ISd and PESQ. We have also used in our evaluation a number of speech signals (ten sentences pronounced in Arabic language by a male voice and ten others pronounced by a female voice) and others speech sentences taken from TIMIT Database. We have also used different types of noises which are Car, White, F16, Tank and pink noises. The results obtained from the SNR, SSNR, ISd and PESQ computations, show that the proposed technique (SBWT/MSS-MAP) outperforms the technique based on MSS-MAP estimation and the Wiener Filtering. When compared with the technique supervised and online NMF based noise reduction, the proposed technique is better when the SNR is higher and we have the opposite when the SNR is lower.

References

Bahoura, M., & Rouat, J. (2001). Wavelet speech enhancement using the teager energy operator. *IEEE Signal Processing Letters, 8*, 10–12.

Bahoura, M., & Rouat, J. (2006). Wavelet speech enhancement based on time-scale adaptation. *Speech Communication, 48*(12), 1620–1637.

Biswas, A., Sahu, P. K., Bhowmick, A., & Chandra, M. (2014). Feature extraction technique using ERB like wavelet sub-band periodic and aperiodic decomposition for TIMIT phoneme recognition. *International Journal of Speech Technology, 17*(4), 389–399.

Boll, S. (1979). Suppression of acoustic noise in speech using spectral subtraction. *IEEE Transactions on Signal Processing, 27*(2), 113–120.

Chen, S. H., & Wang, J. F. (2004). Speech enhancement using perceptual wavelet packet decomposition and teager energy operator. *Journal of VLSI Signal Processing System, 36*(2–3), 125–139.

Cohen, I. (2001). Enhancement of speech using bark-scaled wavelet packet decomposition. In *Eurospeech 2001* (pp. 1933–1936). Aalborg, Denmark.

Daqrouq, K., Abu-Isbeih, I. N., Daoud, O., & Khalaf, E. (2010). An investigation of speech enhancement using wavelet filtering method. *International Journal of Speech Technology, 13*(2), 101–115.

Deller, J. R., Hansen, J. H. L., & Proakis, J. G. (2000). *Discrete-time processing of speech signals* (2nd ed.). New York: IEEE Press.

Ephraim, Y., & Malah, D. (1984). Speech enhancement using a minimum mean square error short time spectral amplitude estimator. *IEEE Transactions on Acoustics, Speech Signal Processing, 32*, 1109–1121.

Giguere, C. (1993). *Speech processing using a wave digital filter model of the auditoryperiphery*. Cambridge: University of Cambridge.

Giguere, C., & Woodland, P. C. (1994). A computational model of the auditory periphery for speech and hearing research. *Journal of Acoustical Society of America, 95*(1), 331–342.

Girish, K. V., Ramakrishnan, A. G., & Ananthapadmanabha, T. V. (2015). Adaptive dictionary based approach for background noise and speaker classification and subsequent source separation. *Journal of Latex Class Files, 14*(8).

Haykin, S. (1996). *Adaptive Filter Theory* (3rd ed.). Upper Saddle River, NJ: Prentice Hall.

Hendriks, R. C., Gerkmann, T., & Jensen, J. (2013). DFT-domain based single-microphone noise reduction for speech enhancement: A survey of the state of the art. *Synthesis Lectures on Speech and Audio Processing, 9*(1), 1–80.

Hirsch, H., & Pearce, D. (2000). The aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions. In ISCA Tutorial and Research Workshop ASR2000, Paris, France.

Hu, Y., & Loizou, P. C. (2004). Speech enhancement based on wavelet thresholding the multitaper spectrum. *IEEE Transactions on Speech and Audio Processing, 12*(1), 59–67.

Hu, Y., & Loizou, P. C. (2008). Evaluation of objective measures for speech enhancement. *IEEE Transactions on Speech, Audio Processing, 16*(1), 229–238.

Johnson, M. T., Yuan, X., & Ren, Y. (2007). Speech signal enhancement through adaptive wavelet thresholding. *Speech Communication, 49*(2), 123–133.

Loizou, P. C. (2007). *Speech enhancement theory and practice*. Abingdon: Taylor & Francis.

Lu, C. T., & Wang, H. C. (2003). Enhancement of single channel speech based on masking property and wavelet transform. *Speech Communication, 41*(2–3), 409–427.

Mahmoudi, D. (1997). A microphone array for speech enhancement using multiresolution wavelet transform. In Proceedings of Eurospeech'97 (339–342). Rhodes, Greece.

Mohammadiha, N., Smaragdis, P., & Leijon, A. (2013). "Supervised and unsupervised speech enhancement using nonnegative matrix factorization. *IEEE Transactions on Audio, Speech, and Language Processing, 21*(10), 2140–2151.

Mortazavi, S. H. & Shahrtash, S. M. (2008). Comparing denoising performance of DWT, WPT, SWT and DT-CWT for partial discharge signals. In Proceedings of the 43rd International Universities Power Engineering Conference (UPEC'08) (pp. 1–6). Padova, Italy.

Rix, A. W., Beerends, J. G., Hollier, M. P., & Hekstra, A. P. (2001). Perceptual evaluation of speech quality (pesq)—a new method for speech quality assessment of telephone networks and codecs. In Proceedings if ICASSP, IEEE International Conference on acoustics, speech and signal processing (Vol. 2, pp. 749–752).

Seok, J. W., & Bae, K. S. (1997). Speech enhancement with reduction of noise components in the wavelet domain. In ICASSP 97 (pp. 1223–1326). Munich, Germany.

Shao, Y., & Chang, C. H. (2007). A generalized time-frequency subtraction method for robust speech enhancement based on wavelet filter banks modeling of human auditory system. *IEEE Transactions on Systems, Man, and Cybernetics. Part B, Cybernetics, 37*(4), 877–889.

Shensa, M. J. (1992). The discrete wavelet transform wedding à trouse and Mallat algorithms. *IEEE Transactions on Signal Processing, 40*(10), 2464–2482.

Singh, S., Mutawa, A. M. (2016). A wavelet-based transform method for quality improvement in noisy speech patterns of Arabic language. *International Journal of Speech Technology*, 1–9.

Talbi, M., & Aicha, A. B. (2014). Enhancement of speech signal based on application of the maximum a posterior estimator of magnitude-squared spectrum in stationary bionic wavelet domain. In 2nd International Conference on mathematical, computational and statistical sciences (MCSS '14). Wseas, Gdansk, Poland.

Talbi, M., Salhi, L., Abid, S., & Cherif, A. (2010). Recurrent neural network and bionic wavelet transform for speech enhancement. *International Journal of Signal and Imaging Systems Engineering, 3*(2), 136–144.

Taşmaz, H., & Erçelebi, E. (2008). Speech enhancement based on undecimated wavelet packet-perceptual filterbanks and MMSE–STSA estimation in various noise environments. *Digital Signal Processing, 18*(5), 797–812.

Vaz, C., Ramanarayanan, V., & Narayanan, S. (2013). A two step technique for MRI audio enhancement using dictionary learning and wavelet packet analysis, In Proceedings InterSpeech (pp. 1312–1315).

Yang, L., & Loizou, P. C. (2011). Estimators of the magnitude-squared spectrum and methods for incorporating SNR uncertainty. *IEEE Transactions on Audio, Speech, and Language Processing, 19*(5), 1123–1137.

Yao, J., & Zhang, Y. T. (2001). Bionic wavelet transform: a new time-frequency method based on an auditory model. *IEEE Transactions on Biomedical Engineering, 48*(8), 856–863.

Yao, J., & Zhang, Y. T. (2002). The application of bionic wavelet transform to speech signal processing in cochlear implants using neural network simulations. *IEEE Transactions on Biomedical Engineering, 49*(11), 1299–1309.

Zavarehei, E., Vaseghi, S., & Yan, Q. (2006). Inter-frame modeling of DFT trajectories of speech and noise for speech enhancement using Kalman filters. *Speech Communication, 48*(11), 1545–1555.

Zheng, L., Zhang, Y. T., Yang, F. S., & Ye, D. T. (1999). Synthesis and decomposition of transient-evoked otoacoustic emission based on an active auditory model. *IEEE Transactions on Biomedical Engineering, 46*, 1098–1106.