International Journal of Information, Control and Computer Sciences
ISSN: 2517-9942
Vol:11, No:12, 2017

Speech Enhancement Using Wavelet Coefficients Masking with Local Binary Patterns

Christian Arcos, Marley Vellasco, Abraham Alcaim

***Abstract*—**In this paper, we present a wavelet coefficients masking based on Local Binary Patterns (WLBP) approach to enhance the temporal spectra of the wavelet coefficients for speech enhancement. This technique exploits the wavelet denoising scheme, which splits the degraded speech into pyramidal subband components and extracts frequency information without losing temporal information. Speech enhancement in each high-frequency subband is performed by binary labels through the local binary pattern masking that encodes the ratio between the original value of each coefficient and the values of the neighbour coefficients. This approach enhances the high-frequency spectra of the wavelet transform instead of eliminating them through a threshold. A comparative analysis is carried out with conventional speech enhancement algorithms, demonstrating that the proposed technique achieves significant improvements in terms of PESQ, an international recommendation of objective measure for estimating subjective speech quality. Informal listening tests also show that the proposed method in an acoustic context improves the quality of speech, avoiding the annoying musical noise present in other speech enhancement techniques. Experimental results obtained with a DNN based speech recognizer in noisy environments corroborate the superiority of the proposed scheme in the robust speech recognition scenario.

***Keywords*—**Binary labels, local binary patterns, mask, wavelet coefficients, speech enhancement, speech recognition.

I. Introduction

IN the presence of various kinds of background noise, the performance of many real-world speech processing applications, such as hearing aids design, hands-free mobile telephony, speech transmission and robust speech recognition, is far from being satisfactory. The noise degrades the systems to levels where their use may become definitely unacceptable. In recent years, speech enhancement has attracted much research effort to deal with this issue. The goal of enhancement is to improve the intelligibility and quality of a speech signal degraded in adverse conditions. Many methods have been proposed in the literature to handle with the noise problem [1], [2] under various assumptions. Some of them are based mainly on the estimation of the short-term noise power spectrum to suppress its components and reconstruct the clean signal. Spectral subtraction (SS) [3] is a classical method for noise suppression. It averages the noisy signal over non-speech sections through a voice activity detector (VAD), subtracting an estimate of the short-term noise spectrum and providing a measure of the present noise floor. An important problem of this method is that assumptions regarding background noise are required to make them work reasonably well. They depend on characteristics such as the stationary of noise, the SNR of the observed signal, etc. Another technique proposed in the literature is the Wiener filtering [2]. The goal of this filter is the reduction of noise from the second order statistics. It assumes that the voice signal and the additive noise are stationary stochastic processes with known spectral characteristics, being an optimal filter to recover clean speech in the Minimum Mean Square Error (MMSE) [4] sense. However, a crucial problem usually found in these traditional methods based on noise suppression is that the resulting speech is modified by an annoying artefact known as 'musical noise'.

Another well-known enhancement approach is the wavelet denoising (WD) proposed by Donoho [5]. Unlike the previous ones, this technique attempts to enhance speech signal without requiring explicit speech pause detection for noise level or SNR estimation. It leads to a good representation of stationary as well as nonstationary segments of the speech signal. The wavelet-based algorithm employs the discrete wavelet transform as subband decomposition. It can be used to extract the localized contributions of the signal of interest.

In contrast to noise suppression methods, based on noise estimations, and inspired by the human auditory processing, Wang et al. in [6] introduces a speech segregation approach, to separate speech from background noise. This approach has shown considerable promise to improve the speech enhancement results. It considers that the sounds that reach the ear are subject to a process called Auditory Scene Analysis (ASA for its acronym in English) [7]. Based on this process, it has been proposed the classical ideal binary mask (IBM) [8], [9], which has been suggested as a primary computational goal for Computational Auditory Scene Analysis (CASA) systems. The IBM may be seen as a binary classification of time-frequency (T-F) units constructed from premixed target and interference. Each unit in the T-F representation of the noisy signal is identified as speech domain when the T-F unit exceeds a threshold or noise domain otherwise.

Following these research lines, we propose a speech enhancement technique, referred to as Wavelet LBP (WLBP), using the relevant information of the wavelet transform according to scale or resolution, and a mask based on the Local Binary Patterns (LBP) [10], [11] approach. It is often used in 2-D image processing for texture description. Our proposal provides, for every one coefficient of each high-frequency subband, the masking which converts it into a value encoded with a higher level of information through the LBPs instead of being eliminated through a threshold. Hence, the method aims at indicating which coefficient of the wavelet transformation of noisy speech is dominated by

C. Arcos, M. Vellasco and A. Alcaim are with the Center for Telecommunications Studies CETUC and the Department of Electrical Engineering, PUC-Rio, Rio de Janeiro, RJ, 22451-900 Brazil (e-mail: christian@cetuc.puc-rio.br, marley@ele.puc-rio.br, aalcaim@gmail.com).

noise. The effectiveness of this scheme relies primarily on the fact that with the LBP codes the value of the original coefficient is encoded with the values of the neighbouring coefficients. Therefore, the coded information will take into account the highest level of information. We evaluate the proposed WLBP on the AURORA-4 tasks (clean and corrupted speech) [12]. The performance assessments are carried out in terms of the objective speech quality measure p.862, known as perceptual evaluation of speech quality (PESQ) standard. as well as the word error rate (WER) in a DNN based continuous speech recognition system. The rest of the paper is organized as follows. Section II provides a brief overview of prior works related to the wavelet-based speech enhancement algorithm and the local binary patterns technique. In Section III we introduce the proposed speech enhancement method using wavelet masking with local binary patterns (WLBP). Simulation results are presented in Section IV and finally, Section V contains some concluding remarks.

II. Prior Work

A. Wavelet-Based Speech Enhancement: Wavelet Denoising (WD)

Let $y(t) = x(t) + r(t)$ denote the noisy signal, with $x(t)$ and $r(t)$ representing the clean speech and noise, respectively. The general wavelet denoising algorithm proposed in [5], [13], attempts to recover a signal $x(t)$ from the noisy data $y(t)$. The wavelet-based speech enhancement algorithm known as wavelet denoising (WD) is summarized as follows:

Step 1: Apply a J-level wavelet decomposition to the noisy signal to produce the corrupted wavelet coefficients.
Step 2: Apply the appropriate thresholding nonlinearity to the detail (high-frequency) coefficients in order to shrink the wavelet coefficients of the noisy signal. The threshold rule can be either soft or hard. In this paper we use the soft-threshold function defined by

$$\eta_y(\beta_{jk}, t) = \begin{cases} sgn(\beta_{jk})(|\beta_{jk}| - \delta_j), & |\beta_{jk}| \geq \delta_j \\ 0, & otherwise \end{cases} \quad (1)$$

where β_{jk} is the *k*th detail (high-frequency) coefficient of the noisy signal at level *j*, at the *t-th* frame (time) in a particular resolution, and the function *sgn(.)* is $+1$ if the argument is positive and -1 otherwise.
Step 3: Inverse wavelet transform of the thresholded wavelet coefficients to obtain the denoised signal.

Note that the above algorithm was developed taking into account the method of denoising denominated Visushrink introduced by Donoho [5], where δ_j in (1) represents a soft-threshold, which is proposed as an universal threshold estimate given by

$$\delta_j = \sigma_j \sqrt{2ln(N_j)} \quad (2)$$

where N_j represents the size of the coefficients in level *j* and σ_j is a rough estimate of the noise level. This estimate is given by

$$\sigma_j = \frac{mad(\beta_{jk})}{0.6745}, \quad k = 0, ..., N_j - 1 \quad (3)$$

where *mad* is the median absolute deviation of the detail coefficients at the highest resolution level *($j = J$)*.

B. Local Binary Patterns Technique

The Local Binary Patterns technique (LBP) was originally developed for Digital Image Processing, introduced as a complementary measure for local image contrast [10]. It has become one of the best texture descriptors, in terms of its performance and highly discriminative abilities [14], [15]. The aim of this scheme is to summarise the local structure in an image by comparing each pixel with its *p* neighbours. The original LBP operator typically works in a $3X3$ pixel block of an image (see Fig. 1 for illustration), where every single pixel in the block is thresholded by its central pixel value. This procedure results in a binary number which is summed to each neighbour binary value to be transformed into a decimal number, obtaining a label for the centre pixel. As mentioned in [16], the neighbourhood consists of 8 pixels, a total of $2^8 = 256$ different labels can be obtained depending on the relative grey values of the centre and the pixels in the neighbourhood. The LBP code for the central pixel is given in a decimal form as

$$LBP_p = \sum_{i=0}^{p-1} sgn(f_p - f_c)2^p \quad (4)$$

where f_c represents the gray value of the center pixel, f_p the gray values of the *p*=8 surrounding pixels, and the function *sgn* is the same as in (1).

In [11], the authors adapted the 2-D LBP operator to 1-D LBP and presented a theoretically very simple, yet efficient, 1-D LBP approach for voice activity detection (VAD). The concept of the one-dimensional LBP method consists in a binary label describing the abrupt local changes of the 1-D signal which attempts to estimate periods of speech and non-speech. The LBP 1-D code is obtained from a sliding window with an odd number of samples through the signal, where each neighbouring sample is threshold against central samples of the processing window. An example of the 1-D LBP operator and their binary codes are given in Fig. 2, where *p*, the number of neighbourhoods, is set to 8 (1x8 mask pattern).

III. Wavelet Coefficients Masking Based on Local Binary Patterns

The aim of masking techniques is to separate speech from noise sources. As mentioned in Section I, the IBM mask is considered to be a goal of CASA in order to achieve this purpose. It consists of a T-F binary matrix constructed from pre-mixed speech and noise, where each T-F unit is set to 1 if the local SNR is greater than a threshold and 0 otherwise. This mask has been widely used in the literature and it has been shown that under certain constraints it is the optimal binary

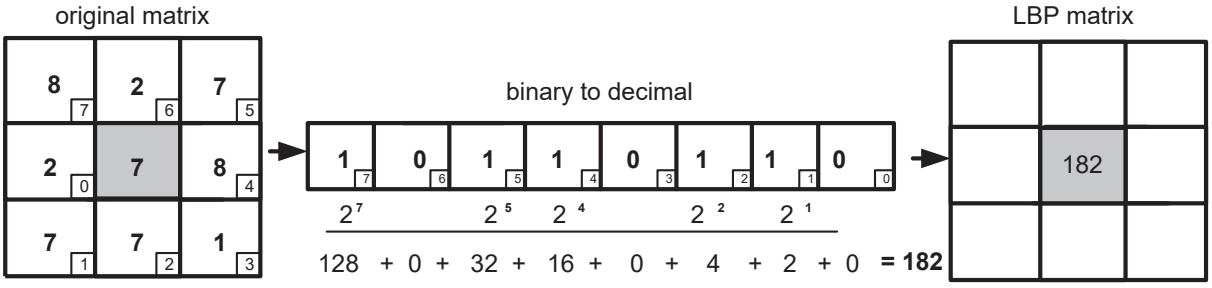

Fig. 1 Computing the binary code of eight neighbouring pixels

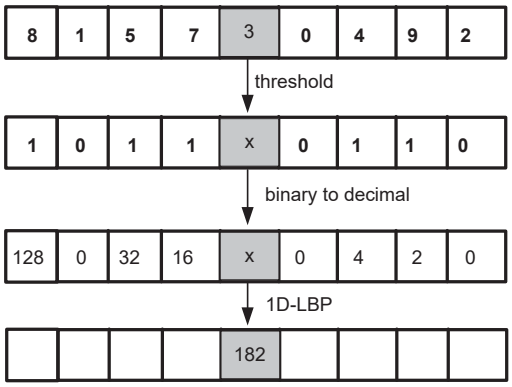

Fig. 2 Computation of the 1-D LBP operator of eight neighbouring samples

mask in terms of Signal-to-Noise Ratio (SNR). However, this algorithm presents a significant limitation affecting speech quality. When spectral frequency components are reduced to zero, they produce a musical noise. Furthermore, this kind of mask is based on true speech spectrum requiring access to the true local (instantaneous) SNR. On the other hand, one critical decision of this kind of methods is to choose a suitable T-F domain to represent the time varying contents of the signal. Traditionally, they use the short time Fourier transform (STFT) to produce a time-frequency representation for the sound mixture.

In this paper, we propose a new masking technique based on a compromise between precise temporal information and frequency localization. This is offered by the wavelet transform, which presents a solution to overcome the shortcomings of the Fourier transform. This is due to its ability to incorporate additional temporal information that covers multiple frames in the characteristic vector. In the proposed scheme we apply the LBP mask to the high-frequency subbands of the wavelet denoising technique. The diagram of our mask estimation based on wavelet coefficients is shown in Fig. 3. Details are given in the following paragraphs.

As mentioned in Section II, in the *step 1* we compute the wavelet packet of the input signal using the wavelets represented by

$$\beta_{jk,t} = \frac{1}{\sqrt{N_j}} \sum_{t=0}^{N-1} y(t)\psi_{j,k}(t) \tag{5}$$

where $\beta_{jk,t}$ are the coefficients of the wavelet transform for each *j-th* level of decomposition (scale j) and position k and $\psi_{j,k}(t)$ is the family of wavelet functions with scale j, position k. In (5), t is the index of the frame in a particular high-frequency, and the wavelet function is defined by

$$\psi_{j,k}(t) = \sqrt{2^j}(\psi 2^j t - k) \tag{6}$$

The decomposition of the signal into different frequency bands is obtained by successive low-pass and high-pass filtering in the time domain. The input corrupted speech signal $y(t)$ is first filtered by a low-pass filter and a high-pass filter. The result will be a low-pass signal a_1 and a high-pass signal d_1, each containing half of the samples of the input signal $y(t)$. The high-pass filter produces the wavelet coefficients where the LBP mask will be applied for level *J*. The low-pass filter produces the scale function for the next level of the hierarchical decomposition. When the low-frequency bands are input to another filter bank system, identical to the first one, a tree structure is created, which divides the spectrum of the original signal into octaves. The decomposition produces *J* levels of wavelet coefficients (see Fig. 4) corresponding to individual signals where the high-frequency ones will be used for the proposed masking. In Fig. 4 the input signal is decomposed into 5 levels, where the signal $y(t)$ (Fig. 4(b)) corresponds to the signal to be analyzed. The signal a_5 is the low-frequency component of the input signal since it is the output of the last low-pass filter of the decomposition tree. The signals d_j (j = 1 ... 5) are the high-frequency components, being d_1 the highest one because it is obtained from the first filter of the tree. These signals are referred to as detail signals.

The output of any high-frequency filter is subdivided into overlapping consecutive time frames of 32 ms and 10 ms time shift. This process generates a matrix of two dimensions $\Gamma_{M,N}$, where *M* represents the number of frames and *N* is the number of wavelet coefficients in each frame. This produces a T-F units matrix for each level. They will be enhanced according to the adapted LBP operator to work on each row of the matrix $\Gamma_{M,N}$ for each *j-th* level of decomposition. A mathematical description of our adapted LBP for each $j-th$

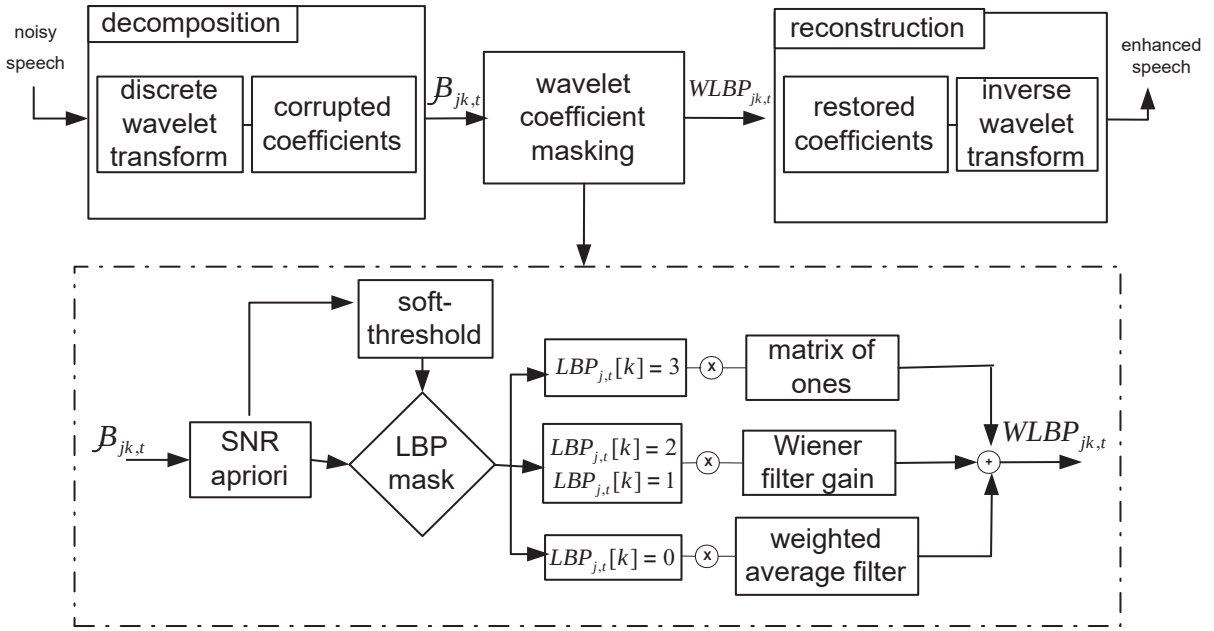

Fig. 3 Block diagram of proposed method, based on local binary patterns and wavelet transform

level, $k-th$ coefficient and a particular time (frame) t is given as follows

$$LBP_{j,t}[k] = \sum_{i=1}^{p/2} \Big\{ sgn[\gamma[k-i] - \gamma[k]] 2^{p/2-i} + sgn[\gamma[k+i] - \gamma[k]] 2^{p/2+i-1} \Big\} \quad (7)$$

where p is the number of neighbouring coefficients surrounding each T-F unit in analysis. The function $sgn[.]$ is set to 1 if the difference between the neighbouring coefficients and the T-F unit of analysis is greater than the threshold δ_j, given in (2) and is important because it avoids the influence of very small noises, $\gamma[k]$ represents the *a priori* SNR for each T-F unit in dB estimated directly from the noisy coefficients. To estimate $\gamma[k]$ we have used the algorithm reported in [17]. For the case $p = 2$, we obtain LBP codes ranging from 0 to 3. For each LBP code, we set the corresponding mask value to 1 if the LBP code is 3. This represents the situation where the speech energy is significantly higher than the noise. When the LBP codes are 1 or 2, these values are smoothed using a Wiener filter gain function. Finally, a temporal smoothing through a weighted average filter is carried out when the LBP code is 0 to reduce fluctuations between the local energy of the noisy speech and the one when the speech energy is greater than the noise. This procedure enhances speech presence in neighbouring coefficients, smoothing all T-F units with dominant noise energy, instead of removing them, as in IBM. Quantitatively, for the case where $p = 2$ the WLBP mask for each $j-th$ level, $k-th$ coefficient and a particular time t is defined as

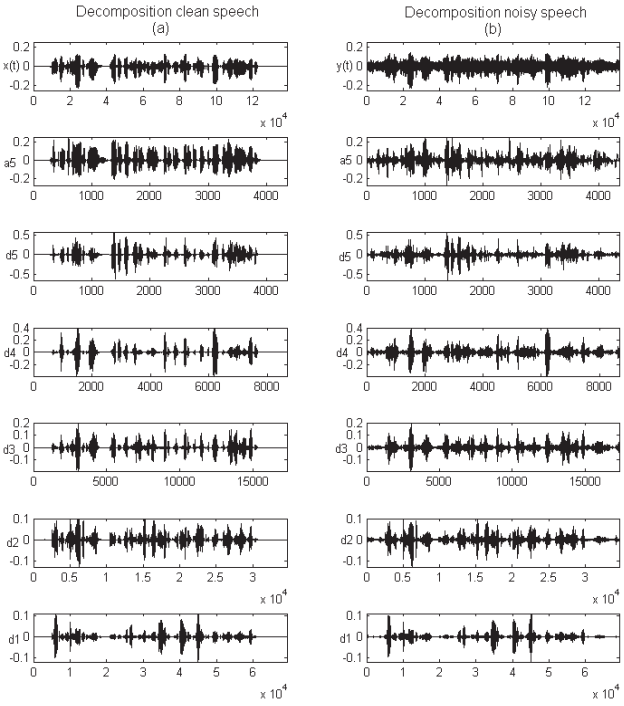

Fig. 4 Wavelet decomposition of the input signal into 5 levels: (a) clean speech and (b) corrupted speech with babble noise at 0dB SNR

$$WLBP_{jk,t} = \begin{cases} \dfrac{\gamma[k-1] + 2\gamma[k] + \gamma[k+1]}{4}, & LBP_{j,t}[k] = 0 \\ \sqrt{\dfrac{\gamma[k]}{1+\gamma[k]}}, & LBP_{j,t}[k] = 1 \; or \; 2 \\ 1, & LBP_{j,t}[k] = 3 \end{cases} \quad (8)$$

Finally, the inverse transform is applied in order to obtain the synthesis of the signal. In this step, the improved speech signal is passed through the high-pass and low-pass synthesis filters. For the high-pass filters, the reconstructed signal is derived from the retained coefficients. The inverse wavelet transform of the detail signals is given by

$$\beta_{jk,t}^{-1} = \frac{1}{\sqrt{N_j}} \sum_j \sum_k WLBP_{jk,t} \psi_{j,k}(t) \tag{9}$$

In Algorithm 1 we summarize the proposed technique for the case where the number of neighbours is $p = 2$.

Algorithm 1 Computing the WLBP mask for p=2

Input: Corrupted speech signal y(t).
Output: $WLBP_{jk,t}$.

Compute the wavelet transform to the noisy signal usnig Daubechies 10 mother wavelet with J=5 levels of decomposition.
Segment each j-th detail level, j=1,...,5, into 32-ms frames (256 samples at an 8kHz sampling frequency) with 10-ms intervals
Compute the SNR *a priori* $\gamma[k]$ for each high-frequency wavelet subband according to [17].
while $\gamma[k]$ True (for $k = [0 : 255]$) **do**
  Perform the $LBP_{jk,t}[k]$ code on the slide analysis window of length $p = 2$.
  if $\gamma[k \pm 1] \geq \gamma[k]$ **then**
    return $W_p = 1$; increment p
  else
    return $W_p = 0$; increment p
  end if
  if $p = 2$ **then**
    return $LBP_{jk,t}[k] = 2^{W_p} + 2^{W_{p+1}}$
  end if
end while
Separate all segments which diferent values of LBP.
Compute $WLBP_{jk,t}$ according to (8)

IV. Simulation Results

In this section, we present and discuss the simulation results of the proposed algorithm (WLBP), the classical spectral subtraction (SS), the wavelet denoising method (WD), and the estimated binary mask (EBM). In the case of the WLBP and EBM algorithms, we estimate the *a priori* SNR $\gamma[k]$, by using the Improved Minima Controlled Recursive Averaging (IMCRA) algorithm proposed by Cohen [17]. All experiments were conducted on the noisy subset of Aurora-4 task [12]. The chosen subset consists of 330 clean speech utterances mixed with 6 environmental noises (babble, airport, restaurant, street, car, train) ranging from 0dB to 15dB. The original signal was sampled at a frequency of 8kHz. Performance evaluations were carried out with two measures for assessing the quality of the enhanced speech:

1. The ITU-T P.862 Perceptual Evaluation of Speech Quality (PESQ) recommendation, which is an objective measurement for estimating subjective quality obtained in listening-only test [18].

2. The proposed technique was also evaluated by a speech recognition system that was implemented using a baseline hybrid deep neural network-HMM (DNN-HMM). This system was trained using Kaldi recipe 's5' [19]. The training set used for the experiments was the subset train-si84 (7138 utterances). We used the dataset Nov'92 (330 utterances) as the test set. The audio data was preprocessed into 40-dimensional log Mel filter-banks, with deltas and accelerations. The trigram language model used for the task was provided with the WSJ CD. The forced alignments were generated from Kaldi recipe tri4b, corresponding to LDA preprocessing of data, with MMLT and SAT for adaptation. There were a total 3385 triphone states in the alignments.

Performance results in terms of PESQ scores are given in Table I. From this table, we observe that *noisy* results obtained without any enhancement technique, severely affect the speech signal. The efficiency of the algorithms is better when Local Binary Patterns is used in the wavelet masking process. As can be seen, the performance of the proposed masking scheme improves the PESQ measure in all scenarios of SNR averaged over 6 environmental noises.

TABLE I
$PESQ$ Averaged over the Different Kinds of Noise

SNR	noisy	SS	WD	WLBP	EBM
0	1.06	1.19	1.20	1.30	1.11
5	1.19	1.40	1.42	1.56	1.22
10	1.45	1.75	1.78	1.94	1.47
15	1.87	2.20	2.25	2.40	1.89

Although a formal subjective evaluation was not carried out, it was observed from informal listening tests that the proposed method does not present the uncomfortable musical noise present in the other enhancement techniques.

Finally, the performance of a DNN based continuous speech recognition system was assessed by the average word error rate (WER) performance measure (experimental conditions were previously reported in this section). In clean conditions, it produces a WER of 4.71%. Table II shows the performance of our mask compared with the other techniques for six environmental noises. Each method is averaged over 0, 5 10 and 15dB. We can see that the wavelet coefficient masking based on Local Binary Patterns overperforms the spectral subtraction, wavelet denoising and EBM mask in all noise scenarios averaged over the different conditions of SNR. Taking the average of the $\overline{WER}$ over all kinds of noise in Table II we can see that for the WLBP scheme this average is 32.6% while for the spectral subtraction, wavelet denoising and EBM methods this value is 34.11% 37.11% and 36.52% respectively.

TABLE II
Word Error Rate $\overline{WER}$ on the Noisy Subset of the AURORA-4 Corpus, Averaged over the Different Conditions of SNR

system	babble	airport	restaurant	street	car	train
Noisy	59.24	56.80	58.88	46.18	56.27	34.90
SS	43.95	38.48	46.57	31.94	39.15	22.62
WD	36.29	35.82	42.63	34.84	39.94	21.00
WLBP	**36.09**	**34.29**	**40.24**	**31.14**	**34.29**	**18.77**
EBM	56.24	54.28	57.34	44.67	54.02	32.86

V. Conclusions

In this paper, a method to improve speech enhancement has been proposed. This method employs the

wavelet denoising scheme and applies a new mask to threshold the high-frequency wavelet coefficients. The proposed scheme (WLBP) exploits the spectro-temporal characteristics of speech to perform enhancement of the signal, by employing the local binary patterns mask in each high-frequency subband signal of the wavelet decomposition. We have compared our method with well-known enhancement techniques (SS, WD, EBM) in six real noisy environments where the SNR estimation does not depend on the true or ideal condition of knowing all signals a priori. We have shown that the results provided by the proposed scheme are better in objective quality scores, showing to be a good technique for speech enhancement. An experiment was also carried out with a DNN based continuous speech recognizer. We have shown that the WLBP algorithm yields superior speech recognition results, as compared to the SS, WD and EBM schemes. This reveals that not only with respect to objective speech quality but also in terms of the word error rate of a DNN based speech recognizer, the WLBP is more effective in noise reduction.

REFERENCES

[1] J. Benesty, S. Makino, J. Chen, *Speech Enhancement*, Springer, 2005.
[2] P. C. Loizou, *Speech enhancement: theory and practice*, CRC press, 2013.
[3] S. Boll, *Suppression of acoustic noise in speech using spectral subtraction*, IEEE Transactions on acoustics, speech, and signal processing, 27, pp. 113-120 1979.
[4] Y. Ephraim, D. Malah, *Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator* IEEE Transactions on Acoustics, Speech, and Signal Processing, 32, pp. 1109-1121, 1984.
[5] D. L. Donoho, *De-noising by soft-thresholding*, IEEE transactions on information theory, 41, pp. 613-627, 1995.
[6] Y. Wang, K. Han, and D. Wang, *Exploring monaural features for classification-based speech segregation*, IEEE Transactions on Audio, Speech, and Language Processing, 21(2), pp. 270-279, 2013.
[7] D. Wang, G. J. Brown*Computational auditory scene analysis: Principles, algorithms, and applications*, Hoboken, NJ, USA Wiley-IEEE press, 2006.
[8] D. Wang,*On ideal binary mask as the computational goal of auditory scene analysis*, Speech separation by humans and machines, p. 181-197, 2005.
[9] Y. Jiang, H. Zhou, and Z. Feng *Performance analysis of ideal binary masks in speech enhancement* In 4th International Congress Image and Signal Processing (ICISP), Vol. 5, pp. 2422-2425, october. 2011.
[10] T. Ojala, M. Pietikinen, D. Harwood *A comparative study of texture measures with classification based on featured distributions*, Pattern recognition, pp. 51-9, 1996.
[11] N. Chatlani, JJ. Soraghan, *Local binary patterns for 1-D signal processing*, EUSIPCO, p. 95-99, 2010.
[12] D. Pearce, and H. G. Hirsch, *The Aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions.* In Sixth International Conference on Spoken Language Processing, 2000.
[13] D. L. Donoho, I. M. Johnstone,*Threshold selection for wavelet shrinkage of noisy data*, In Engineering in Medicine and Biology Society, Vol. 1, pp. A24-A25, nov. 1994.
[14] S. Liao, M. W. Law, A. C. Chung,*Dominant local binary patterns for texture classification*, IEEE transactions on image processing, 18(5), pp. 1107-1118, 2009.
[15] J. Chen, S. Shan, C. He, G. Zhao, M. Pietikainen, X. Chen, and W. Gao, *WLD: A robust local image descriptor*, IEEE transactions on pattern analysis and machine intelligence, 32(9), pp. 1705-1720. 2010.
[16] D. Gupta, and A. Jindal. *Content based image retrieval using enhanced local tetra patterns* International journal of innovative research in science and engineering, January 2017.
[17] I. Cohen, *Noise spectrum estimation in adverse environments: Improved minima controlled recursive averaging*, IEEE Transactions on Speech and Audio Processing, 11(5), pp. 466-475. 2003.
[18] J. G. Beerends, A. P. Hekstra, A. M. Rix,and M. P. Hollier, *Perceptual evaluation of speech quality (pesq) the new itu standard for end-to-end speech quality assessment part ii: psychoacoustic model*. Journal of the Audio Engineering Society, 50(10), pp. 765-778, 2002.
[19] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, and J. Silovsky. *The Kaldi speech recognition toolkit*, In IEEE workshop on automatic speech recognition and understanding, IEEE Signal Processing Society, Dec 2011.