NOISE SUPPRESSION BASED ON WAVELET PACKET DECOMPOSITION AND QUANTILE NOISE ESTIMATION FOR ROBUST AUTOMATIC SPEECH RECOGNITION

Erhard Rank, Tuan Van Pham, Gernot Kubin

Signal Processing and Speech Communication Laboratory
Graz University of Technology, Graz, Austria
erank@tugraz.at, v.t.pham@tugraz.at, g.kubin@ieee.org
http://spsc.tugraz.at

ABSTRACT

In this paper we address the application of a denoising algorithm based on wavelet package decomposition and quantile noise estimation to noise suppression for automatic speech recognition. The denoising algorithm is adapted to suit the different requirements in machine recognition, as compared to human perception, and is tested in combination with state-of-the-art speech recognition systems. The results show, that, if the proposed algorithm is integrated with the recognition system—including the training process—a performance comparable to recent high-quality noise suppression methods is achieved.

1. INTRODUCTION

Noise suppression is a highly important task to increase the recognition performance and user acceptance of current automatic speech recognition (ASR) systems. In particular when ASR systems are used in environments where it is not possible to use close talking microphones and with strong sources of background noise—like for hands-free applications in cars—robust noise suppression is required. A specific challenge are non-white and non-stationary noise sources, as well as tonal and approximately cyclo-stationary noise sources, emitted, for example, from machines or fans. The application of ASR in such a harsh environment is considered in the European project SNOW (Services for NOmadic Workers), where the task is to provide ASR for workers in a factory floor environment, namely in airplane maintenance.

In this paper we develop the wavelet packet decomposition (WPD)-based denoising algorithm which was proposed in [1] for the application in ASR systems. The advantage of the wavelet transform is the multiresolution analysis which provides a multiscale decomposition of a signal. For denoising of speech signals targeted at increased intelligibility and perceptual comfort for the human listener this algorithm was found to achieve a robust attenuation of background noise and still to preserve intelligibility and naturalness. However, the increase in perceptual quality of a speech signal may not be directly related to an increase in recognition rate of an ASR system. Thus, the proposed algorithm is modified, as described in the following Sect. 2, in order to better meet the requirements in ASR. In Sect. 3 we assess the performance of the enhanced algorithm regarding the influence on the recognition rate of ASR systems in the rough acoustic environment described above. The final section presents the conclusions and an outlook on steps for the further development.

We kindly acknowledge the support by the European Union for the FP6 IST STREP SNOW (FP6-511587). Furthermore, we sincerely thank our SNOW project partner Loquendo, in particular Luciano Fissore, for the speedily dispatch of running a recognition test on our data within less than 12 hours after our request!

2. NOISE SUPPRESSION ALGORITHM

The algorithm for noise suppression for ASR presented here is based on the denoising algorithm presented in [1]. The noisy speech signal is segmented into a sequence of buffers which have 960ms length and 480ms overlap. Each buffer consists of 47 speech frames which have 40ms length and 20ms overlap. By performing the WPD at the 7^{th} scale on each speech frame, the wavelet coefficients of 128 WPD channels are extracted. Next, the thresholds related to the noise levels are estimated based on the universal thresholds [2] and a quantile filtering algorithm using recursive buffering. To handle non-stationary and colored noise, the estimated thresholds are non-linearly adaptively weighted in the time-frequency domain. Then an optimized wavelet shrinking method is used to shrink wavelet coefficients which are below the weighted thresholds towards zero. Finally, the denoised speech frames are reconstructed by the wavelet packet reconstruction (WPR) of the thresholded wavelet coefficients. A schematic of the algorithm is shown in Fig. 1.

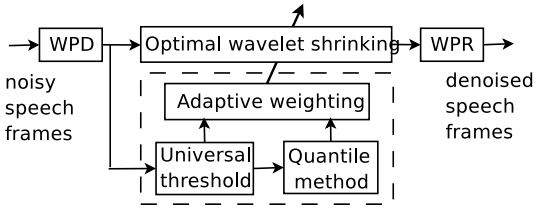

Fig. 1. Scheme of WPD-based noise suppression.

2.1. Wavelet shrinking and universal threshold

We consider a basically additive model of a signal s corrupted by noise e which creates a noisy signal x:

$$x = s + e\,. \tag{1}$$

Because of the linearity of the Wavelet transform, the WPD coefficients of noisy speech $X_{k,i}$ at the i^{th} frame can be expressed as the sum of WPD coefficients of clean speech $S_{k,i}$ and noise $E_{k,i}$ as:

$$X_{k,i} = S_{k,i} + E_{k,i}\,, \tag{2}$$

where $k = 2^j$ is number of WPD channels derived by performing full WPD at scale j, with $j \in \mathbb{Z}$ [$k \in 2^{\mathbb{Z}}$. A simple denoising process is thresholding the noisy wavelet coefficients which are lower than certain thresholds to zero (hard thresholding [3]). An enhanced wavelet shrinking method is proposed by [4] by smoothing the hard thresholding:

$$\tilde{X}_{k,i} = \begin{cases} X_{k,i}\ , & \text{if } |X_{k,i}| > T_{k,i}\ , \\ \dfrac{T_{k,i}\,\mathrm{sign}(X_{k,i})}{\mu_{k,i}} A_{k,i}\ , & \text{if } |X_{k,i}| \le T_{k,i}\ , \end{cases} \tag{3}$$

with $A_{k,i}$ and the adaptive parameter $\mu_{k,i}$ defined as:

$$A_{k,i} = (1+\mu_{k,i})^{\dfrac{|X_{k,i}|}{T_{k,i}}} - 1\ , \tag{4}$$

$$\mu_{k,i} = \theta\, \frac{\max\limits_{m}\{|X_{k,i}[m]|\}}{T_{k,i}}\ , \tag{5}$$

where θ is a constant factor and m is the index of the coefficients in the WPD channels. In contrast to hard thresholding, this shrinking preserves non-zero values for coefficients which are below the thresholds. Examples of the function in (3) are given in Fig. 2.

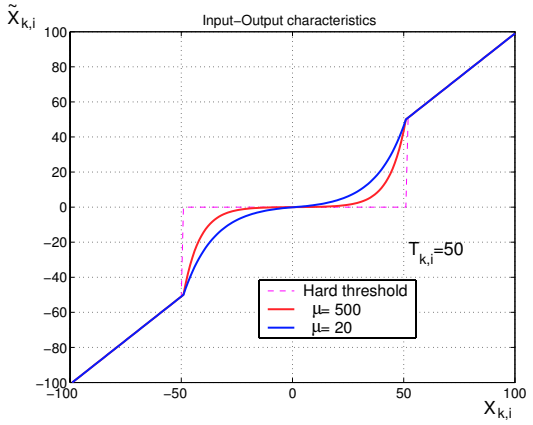

Fig. 2. Smoothed hard thresholding.

The universal threshold procedure in [2] which uses a robust estimate of standard deviation is applied to estimate the thresholds of the WPD coefficients as:

$$T_{k,i} = \frac{1}{\gamma_{\text{MAD}}}\text{Median}(|X_{k,i}|)\sqrt{2\log N_{\text{WP}}}\ , \tag{6}$$

where $\gamma_{\text{MAD}} = 0.6745$ and $D_{k,i}$ is the WPD coefficient sequence having length N_{WP} in each channel for each frame.

2.2. Quantile threshold estimation of noise level

A quantile-based algorithm to estimate the threshold related to noise level has been presented in [1]. The threshold is estimated by taking the q^{th} quantile over the duration of the buffer in every WPD channel. The algorithm is implemented using a recursive buffer which is constructed from the overlapping buffers as follows:

- First, the current threshold buffer is built from universal thresholds $T_{k,i}$ which is calculated for all WPD channels of all frames in the current buffer.
- Second, the recursive buffer is formed by merging the thresholds $T_{k,i}$ of the frames $i = 24, \ldots, 47$ in the current threshold buffer with the thresholds $T_{k,i}$ which are selected from quantile range $q = 0.1, \ldots, 0.6$ of the previous sorted threshold buffer.
- Then, for each WPD channel, the thresholds in the recursive buffer are sorted in ascending order which results in $T_{k,i'}$, where $i' = 1, \ldots, N_f$ is the frame index after sorting with $N_f = 47$. This sorted recursive buffer is used for the next loop.
- Finally, the threshold related to the noise levels, Γ_k, for all frames in the sorted recursive buffer at the k^{th} channel is determined as the q^{th} quantile :

$$\Gamma_k = T_{k,i'}\ \big|_{i'=\lfloor qN_f \rfloor} \tag{7}$$

From our experiments, the quantile $q = 0.2$ is selected out of a candidate range $q = 0.0, 0.1, \ldots, 0.6$, as the value yielding the best performance.

2.3. Nonlinearly adaptive weighting

To handle non-stationary and colored noise, the noise threshold Γ_k of each frame i in each channel k is weighted as follows:

$$\tilde{\Gamma}_{k,i} = \lambda_{k,i}\eta_k\Gamma_k, \tag{8}$$

where $\lambda_{k,i}$ and η_k are nonlinear functions in the time-frequency and in the frequency domain as described below. $\tilde{\Gamma}_{k,i}$ is the weighted estimate of the noise threshold.

2.3.1. *Frequency weighting*

Obviously, the energy distribution of colored noise is not equal over all frequency channels. Thus, the scale-dependent estimation of noise level is necessary in this case. As a solution proposed in [1], the colored noise is handled by weighting the threshold Γ_k with the nonlinear parameter η_k as:

$$\eta_{k0} = (10\Gamma_k)^{-(10\Gamma_k)^{a_0}} + d_0\ , \tag{9}$$

where $a_0 = 0.55, d_0 = 0.6$. The function η_{k0} in Fig. 3 amplifies strongly the threshold levels in the quantile range $[0, ..., 0.3]$ and less those in the remaining range. This means that more small noisy coefficients in high-frequency channels will be shrunk by the amplified thresholds while the large coefficients at low-frequency channels are only slightly impacted. This leads to low background noise while maintaining high quality of the denoised speech signal.

In case of low-frequency colored noise, the smaller impact on low-frequency channels results in a high level of background noise after denoising. As reported in subjective tests [1], this can be tolerated in speech enhancement applications because the naturalness of speech is maintained. However, the high background noise is sensitive in speech recognition applications and may reduce the word error rate. For this reason, we develop another weighting function η_{k1} shown as a dashed line in Fig. 3 that puts stronger weighting on the large Γ_k stemming from low-frequency WPD channels:

$$\eta_{k1} = (a_1\Gamma_k)^{b_1} + d_1\ , \tag{10}$$

where $a_1 = 70, b_1 = 0.5, d_1 = 0.4$ are selected manually from our experiments to achieve high performance of ASR. Figs. 4 and 5 shows the denoising results of a recording corrupted by low-frequency colored noise at 2dB SNR for the two different functions η_{k0} and η_{k1}.

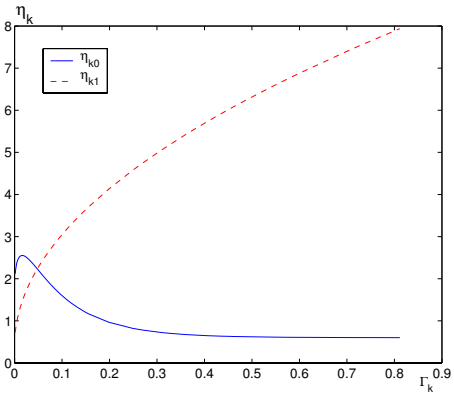

Fig. 3. Weighting on quantile thresholds in the frequency domain.

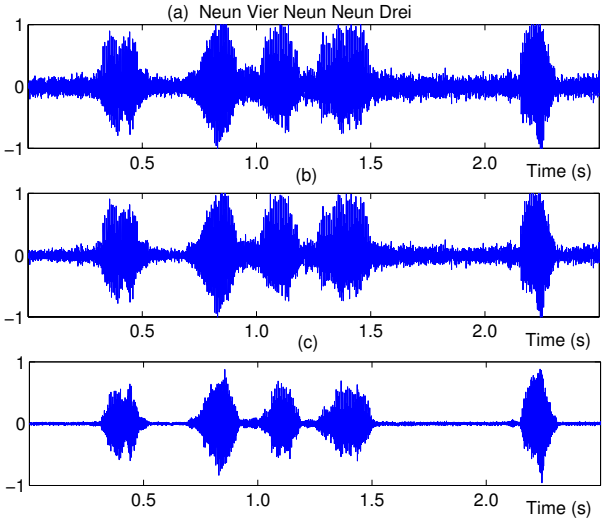

Fig. 4. Speech waveform of (a) noisy recording, and denoised recordings by (b) η_{k0}, and (c) η_{k1}.

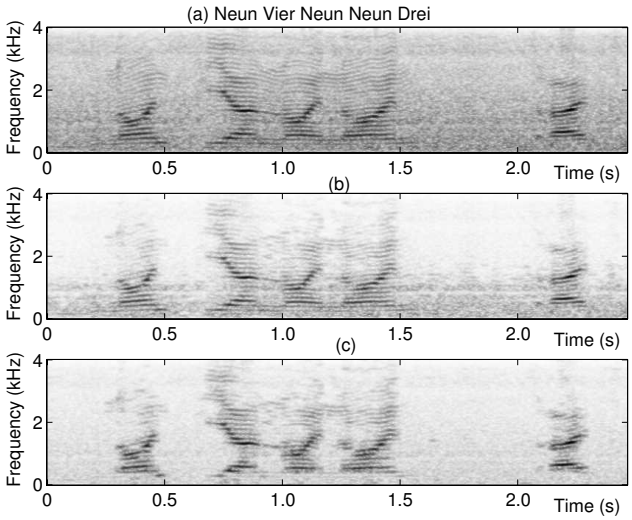

Fig. 5. Spectrogram of (a) noisy recording, and denoised recordings by (b) η_{k0}, and (c) η_{k1}.

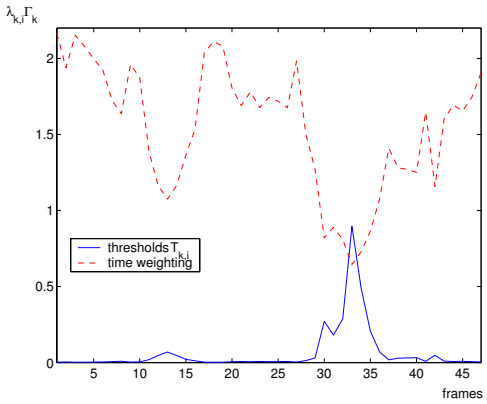

Fig. 6. Time adaptive weighting based on TDC.

2.3.2. Time adaptive weighting

The time-variant threshold dependent curve (TDC) $\lambda_{k,i}$ is built to track where the speech or noise appears along the time axis as follows:

$$\lambda_{k,i} = {T_{k,i}}^{-a_2} + d_2 \ , \tag{11}$$

where $a_2 = 0.14, d_2 = 0.2$ are constants, $T_{k,i}$ is the universal threshold from (6) in the current threshold buffer. Frames with smaller thresholds $T_{k,i}$, which are assumed to hold noise, will yield an increase of the threshold values and thus result in aggressive noise suppression. Frames with large $T_{k,i}$, which are considered to contain speech information, will result in a smaller threshold, to preserve speech quality. An example of a TDC in relation to the universal thresholds for one wavelet channel is shown in Fig. 6.

2.4. Optimal wavelet shrinking

To avoid speech distortion, especially for unvoiced consonants, the optimal wavelet shrinking is proposed by making the factor θ in (5) adaptive for each wavelet channel:

$$\theta_{k,i} = \exp\left(\alpha \frac{\tilde{\Gamma}_{k,i}}{\max\limits_{i}\{\tilde{\Gamma}_{k,i}\}}\right) \tag{12}$$

where $\alpha = 5.8$ is a slope constant. Due to the direct influence of $\theta_{k,i}$ on $\mu_{k,i}$, the part of function (3) (cf. also Fig. 2) for $|X_{k,i}|$ smaller than the threshold $T_{k,i}$ is automatically closer to the identity function and thus preserves more coefficients for the speech frames, and closer to the hard thresholding function to compress more noisy coefficients for the non-speech frames.

3. RECOGNITION EXPERIMENTS

To assess the suitability of the presented noise suppression algorithms for the use in ASR systems, a number of tests were carried out. Firstly, the algorithm was tested as a pre-processing stage to the front-ends specified by [5] (standard MFCC) and [6] (advanced front-end, AFE) in combination with the HTK recognizer [7] for

the German Aurora 3 SpeechDat-Car corpus. This corpus includes samples of series of digits recorded in a car environment under various driving conditions (car stopped with motor running, town traffic, driving at low speech on rough road, driving with high speed on good road; all recorded with close talking microphone and with a hands-free microphone). The noise environment for the SpeechDat-Car corpus is not as adverse as the factory floor environment in the SNOW project, however, the corpus is widely used for assessing ASR systems and thus allows for direct comparison of noise suppression algorithms, and particularly the training/test set with "high-mismatch" (i. e., with rather clean samples used as training data and very noisy samples as test data) should allow for a rough assessment of how the proposed noise suppression algorithm would behave in a more adverse environment.

Using the proposed noise suppression algorithm as a pre-processing stage for a given ASR system, however, does not increase the recognition rate: For the "high-mismatch" training/test set of the German SpeechDat-Car corpus and the HTK recognizer trained with the according front-end, the word recognition rate (WRR) is reduced from $66.7\,\%$ to $65.6\,\%$ (accuracy from $63.2\,\%$ to $60.4\,\%$) for the standard MFCC front-end, and from $89.8\,\%$ to $85.7\,\%$ (accuracy from $89.5\,\%$ to $77.3\,\%$) for the AFE. We attribute this mainly to the different training and test conditions, and probably to a negative interference between the proposed noise suppression algorithm and the denoising algorithm in the AFE.

The second test thus comprises the training of the HTK recognizer using the proposed noise suppression algorithm as a pre-processing stage for the standard MFCC front-end, and as a replacement for the Wiener filter denoising in the AFE. Here, the WRR of $66.7\,\%$ using the standard MFCC front-end for training and testing is increased to $75.3\,\%$ with the proposed algorithm (accuracy from $63.2\,\%$ to $73.2\,\%$), however, the WRR for the AFE of $89.8\,\%$ is reduced to $85.2\,\%$ (accuracy from $89.5\,\%$ to $83.9\,\%$).

A third test was performed on a corpus set up in the scope of the SNOW project, comprising 435 utterances (a total of 1135 words, utterances are commands for controlling a graphical browser display) recorded by 4 female and 4 male speakers under work conditions in an airplane maintenance facility. For this recognition test the Loquendo ASR system [8] using elaborate spectral subtraction denoising [9] was utilized, using a grammar where all the vocabulary words can be looped without any constraints. The proposed wavelet noise suppression algorithm was again used as a pre-processing stage, in addition to the denoising in the ASR front-end. Like in the experiment with HTK, the WRR is reduced, too, from originally $83.7\,\%$ to $78.4\,\%$. A summary of the test results is given in Table 1.

Table 1. Summary of recognition results: WRR and accuracy (Acc.) of the proposed denoising algorithm (this) in comparison to the original processing (orig) of [5] (MFCC), [6] (AFE), and [8] (LOQ ASR).

	German SpeechDat-Car/HTK				SNOW/LOQ ASR
	w.o retraining		with retraining		w.o retraining
	MFCC	AFE	MFCC	AFE	–
WRR orig	66.7%	89.8%	66.7%	89.8%	83.7%
WRR this	65.6%	85.7%	75.3%	85.2%	78.4%
Acc. orig	63.2%	89.5%	63.2%	89.5%	75.2%
Acc. this	60.4%	77.3%	73.2%	83.9%	67.1%

4. CONCLUSION AND OUTLOOK

In [1] we have shown that an elaborate WPD-based speech enhancement algorithm allows for consistent attenuation of background noise while preserving speech naturalness and intelligibility. In this paper the algorithm has been adapted to suit the requirements of noise suppression for the use with ASR systems. The modifications to the algorithm allow for a more aggressive suppression of background noise compared to the previous setting [1], as exemplified in Figs. 4 and 5.

The experiments with ASR systems show, on the one hand, that, for the application in adverse noise environment, no improvement in recognition rate can be achieved when the proposed algorithm is used as a pre-processing module in addition to ASR internal noise reduction methods without re-training. On the other hand, if the ASR system is trained with the noise suppression algorithm, a significant improvement is achieved using the ETSI 201 108 standard front-end, and the proposed algorithm almost achieves the performance of the noise reduction in the ETSI 202 050 advanced front-end.

Thus, this promising WPD-based algorithm can be applied for both speech enhancement and ASR[1], and should be further investigated and optimized for ASR in future research. In particular, the combination with voice activity detection, as used in the AFE, should be beneficial to increase accuracy (reduce the number of 'insertions'). Within the SNOW project, the aim is to fully integrate this noise suppression algorithm with the feature extraction front-end used in the Loquendo speech recognition system.

5. REFERENCES

[1] T. V. Pham and G. Kubin, "WPD-based noise suppression using nonlinearly weighted threshold quantile estimation and optimal wavelet shrinking," in *Proc. Interspeech*, Lisboa, Portugal, Sept. 2005, pp. 2090–2093.

[2] D. L. Donoho, "De-noising by soft-thresholding," *IEEE Trans. Inform. Theory*, vol. 41, no. 3, pp. 613–627, 1995.

[3] M. Jansen, *Noise reduction by wavelet thresholding*, Springer, USA, 2001.

[4] B. Kotnik, *Robust speech parameterization based on joint wavelet packet decomposition and autoregressive modeling*, Ph.D. thesis, Univ. of Maribor, Maribor, Slovenia, Sept. 2004.

[5] *ETSI ES 201 108 V1.1.1 Speech Processing, Transmission and Quality Aspects (STQ); Distributed speech recognition; Front-end feature extraction algorithm; Compression algorithms*.

[6] *ETSI ES 202 050 V1.1.3 Speech Processing, Transmission and Quality Aspects (STQ); Distributed speech recognition; Advanced front-end feature extraction algorithm; Compression algorithms*.

[7] "The Hidden Markov Model Toolkit (HTK)," http://htk.eng.cam.ac.uk, visited: Jan. 2006.

[8] "Loquendo ASR brochure," http://www.loquendo.com/en/brochure/ASR.pdf, visited: Jan. 2006.

[9] Roberto Gemello, Franco Mana, and Renato De Mori, "A modified Ephraim-Malah noise suppression rule for automatic speech recognition," in *Proc. Int. Conf. Acoust. Speech Sign. Process.*, Montreal, Canada, 2004, vol. 1, pp. 957–960.

[1] In spite of the more aggressive noise suppression approach targeted at the application with ASR, the proposed algorithm still provides better performance regarding intelligibility of the speech signal in a comparison with other denoising algorithms in informal listening tests.