PERCEPTUAL WAVELET FILTERING FOR ROBUST SPEECH RECOGNITION

Tuan Van Pham, Michael Stark, Gernot Kubin

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

ABSTRACT

In this paper an enhanced noise reduction for robust speech recognition is implemented by means of a perceptual wavelet filtering algorithm. The psychoacoustic model is applied to map the universal thresholds to the perceptually universal thresholds for each critical wavelet subband. By improving our quantile filtering method, the change of noise level is tracked more adaptively. The denoising algorithm is compared with well-known noise reduction methods embedded in different state-of-the-art speech recognizers. We achieve almost similar recognition performance with the HTK recognizer on AURORA3 SPEECHDAT-Car corpus and an improvement with the Loquendo recognizer on SNOW-Factory corpus.

Index Terms— wavelet shrinkage, quantile filtering, critical subband, speech recognition, speech enhancement.

1. INTRODUCTION

Ambient noise is a very important factor that can enormously reduce the recognition rate of automatic speech recognition (ASR) systems. In order to keep robustness of the speech recognizers, which are operated in environments with a wide range of noise sources such as car, factory noise, etc., noise reduction is integrated in the front-end units of ASR systems to compensate environmental mismatch between training and testing phases. Speech recognition in car noise environments is studied in the AURORA3 project [1] while the impact of factory noise to ASR systems used in airplane maintenance factories has been examined in the European SNOW project [2]. Dealing with complex noise in such harsh environments raises the challenge for most noise suppression algorithms.

In this paper a so-called perceptual wavelet filtering (PWF) algorithm is developed from the previous wavelet denoising (pWD) method proposed in [3]. We address here novel contributions for estimating perceptually universal thresholds and enhancing our quantile filtering technique which indirectly increase the recognition rates. As shown in Fig. 1, the universal thresholds are firstly calculated from the wavelet coefficients which are obtained by implementing the binary wavelet packet decomposition (WPD) on the noisy speech frames. After that, based on the psychoacoustic model, the perceptually universal thresholds are derived for every critical wavelet subband via a threshold-mapping module. The noise thresholds are then adaptively estimated for each critical subband by an enhanced quantile filtering method. Thanks to its ability of successively estimating the noise thresholds for each input speech frame, the fast change of a non-stationary noise is trackable. The quantile noise thresholds are afterwards weighted and fed into an optimized shrinking function to enhance the noisy wavelet coefficients. Finally, the denoised speech frames are reconstructed by the wavelet packet reconstruction (WPR). This PWF algorithm is then tested as a preprocessing stage to the front-end units of the speech recognizers.

Through Section 2, the estimate of noise thresholds for every critical subband using the enhanced quantile filtering is explained. Section 3 points out some optimal characteristics of the shrinking gain function. In Section 4, by carrying out several recognition experiments, performance of the proposed algorithm is evaluated and discussed. The final section gives a conclusion and future research.

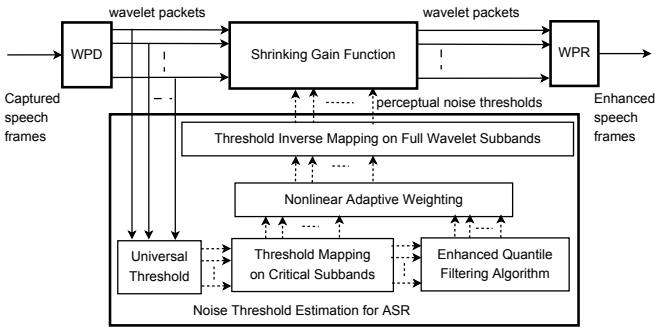

Fig. 1: Block scheme of the proposed PWF algorithm.

2. NOISE THRESHOLD ESTIMATION

The wavelet denoising approach, which is considered as a non-parametric statistical estimation, has been recently developed by using wavelet shrinkage [4]. The wavelet coefficients of noisy speech $Y^k_{m,i}(n)$ can be expressed as the sum of the wavelet coefficients of clean speech $X^k_{m,i}(n)$ and noise $D^k_{m,i}(n)$ as:

$$Y^k_{m,i}(n) = X^k_{m,i}(n) + D^k_{m,i}(n)\,, \tag{1}$$

where $\chi^k_{m,i}(n)$[1] describes a sequence of wavelet packet coefficients (i.e. $d^{2k}_{m,n}$ and $d^{2k+1}_{m,n}$) of each packet node $\{m,k\}$ derived at m^{th}

This research was carried out in the context of COAST-ROBUST, a joint project of Graz University of Technology, Philips Speech Recognition Systems, and Sail Labs Technology. We gratefully acknowledge funding by the Austrian KNet Program, ZID Zentrum fuer Innovation und Technology, Vienna, the Steirische Wirtschaftsfoerderungsgesellschaft mbh, and the Land Steiermark. We are thankful Erhard Rank, Philips Speech Recognition Systems, Vienna for his excellent support. Furthermore, we sincerely thank Loquendo, in particular Luciano Fissore, for running intensive recognition tests. Finally, thanks to Kshitij Gupta, IIMA, India for his nice collaboration.

[1] For simplicity, χ represents the signals Y, X, and D.

scale of the i^{th} speech frame (with $k = 1, \ldots, 2^m$ the packet channel index) by using filter bank as follows:

$$d_{m,n}^{2k} = \sum_p d_{m-1,p}^k \, h(p-2n) = d_{m-1}^k * \overline{h}(2n), \quad (2)$$

$$d_{m,n}^{2k+1} = \sum_p d_{m-1,p}^k \, g(p-2n) = d_{m-1}^k * \overline{g}(2n), \quad (3)$$

where $h(n)$ and $g(n)$ form a pair of conjugate mirror filters used at the analysis stage with $g(n) = (-1)^{1-n} h(1-n)$, and $h(-2n) = \overline{h}(2n)$ and $g(-2n) = \overline{g}(2n)$ are synthesis filters. In this study, the WPD is implemented at a selected decomposition scale $m = 7$ so m is discarded in the notation, and superscript k becomes subscript k to simplify the notation of wavelet packet coefficients as $\chi_{k,i}(n)$ with n the coefficient index. An algorithm to estimate noise level from these wavelet coefficients is proposed in the following parts.

2.1. Perceptually universal threshold

The threshold estimate is based on the minimization of the risk function consisting of bias and variance terms as follows:

$$\begin{aligned} E\{R(T)\} = E\left\{\|E\{\widehat{X}_k(n)\} - X_k(n)\|^2\right\} \\ + E\left\{\|\widehat{X}_k(n) - E\{\widehat{X}_k(n)\}\|^2\right\}. \end{aligned} \quad (4)$$

where $E\{\cdot\}$ is the expectation estimation and $\widehat{X}_k(n)$ is the enhanced coefficients derived from a shrinking function that will be explained in Sec. 3. Under assumption of the i.i.d. noise with variance σ^2, the universal threshold proposed in [4] is proportional to σ and the length N_i of the coefficient sequence at the i^{th} frame:

$$T_i = \sigma\sqrt{2\log N_i} \quad (5)$$

To handle non-white noise where the noise power varies over different wavelet packets, we estimate the universal threshold for every wavelet packet separately. Beside that, a robust estimate of the standard deviation is applied by estimating the median absolute deviation (MAD) of the sequence of coefficients at every packet. The universal threshold T_i therefore is rewritten as:

$$T_{k,i} = \frac{1}{\gamma_{MAD}} \text{Median}(|Y_{k,i}(n)|)\sqrt{2\log N_{k,i}}\,, \quad (6)$$

where $\gamma_{MAD} = 0.6745$ is the conversion factor between the standard deviation and MAD in case of white Gaussian noise.

In order to improve the accuracy of the noise estimation, we propose a perceptual noise threshold estimation method that is based on psychoacoustic model. In literature, a so-called perceptual WPD that its wavelet subbands are designed to match the auditory critical subbands has been used for speech enhancement with an increased performance as reported in [5]. Within our new proposal, we still implement the full WPD. However, the estimation of noise thresholds is carried out only on the critical wavelet subbands (CWS). According to the specifications of the center frequencies, and the corresponding critical bandwidths (CBW) defined for Bark psychoacoustical scale in [6], there are approximately 17 CWSs obtained for the bandwidth of 4 kHz (which is the bandwidth of the recorded speech signal in the SNOW and AURORA3 databases). The perceptual thresholds $P_{j,i}$ of each critical subband j, at the i^{th} frame are estimated by calculating the mean of the universal thresholds $T_{k,i}$ from the corresponding wavelet packets k as defined by a following equation:

$$P_{j,i} = \frac{1}{Cu_j - Cl_j + 1} \sum_{k=Cl_j}^{Cu_j} T_{k,i}, \quad \text{with } j = 1, \ldots, 17. \quad (7)$$

where $[Cl_j \ldots Cu_j]$ are orders of wavelet packets derived by the full WPD. The mapping of these channels into the critical wavelet subbands are described in Tab. 1. By this process, the complexity of the system is reduced due to processing on limited number of critical subbands only, while noise is removed efficiently from all wavelet packets. Moreover, the estimate of noise on the critical wavelet subbands helps to improve the quality of features for ASR which are extracted from mel-frequency channels as used in Sec. 4.

Table 1: Mapping between critical subbands and wavelet packets.

CWS_j	$[Cl_j..Cu_j]$	CBW[kHz]	CWS_j	$[Cl_j..Cu_j]$	CBW[kHz]
1	[1..4]	0 - 0.125	10	[41..48]	1.25 - 1.5
2	[5..8]	0.125 - 0.25	11	[49..56]	1.5 - 1.75
3	[9..12]	0.25 - 0.375	12	[57..64]	1.75 - 2
4	[13..16]	0.375- 0.5	13	[65..72]	2. - 2.25
5	[17..20]	0.5 - 0.625	14	[73..80]	2.25 - 2.5
6	[21..24]	0.625 - 0.75	15	[81..96]	2.5 - 3
7	[25..28]	0.75 - 0.875	16	[97..112]	3 - 3.5
8	[29..32]	0.875 - 1	17	[113..128]	3.5 - 4
9	[33..40]	1 - 1.25			

2.2. Enhanced quantile filtering

In this section, we present an improvement of the quantile filtering method, which was proposed in [3], to track non-stationary noise properly for every captured frame.

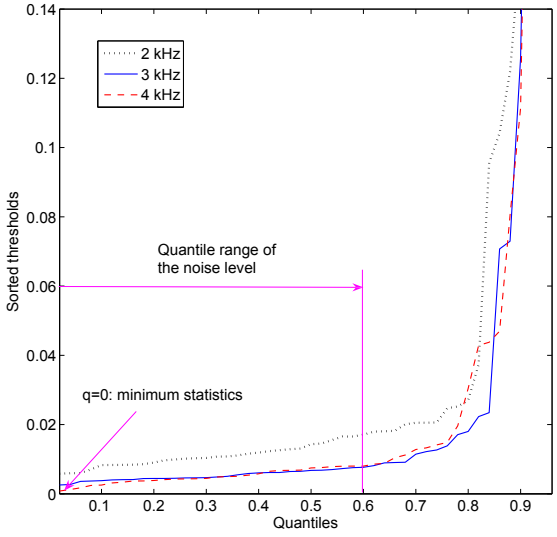

Fig. 2: Quantiles of sorted threshold values over a buffer at three selected wavelet packets.

From the analysis of the universal threshold approach, we conclude that this is a very local estimate. The universal thresholds are calculated from the wavelet packets at a certain frame. Thus, the temporal characteristics of speech and noise are not accounted for the estimation of the threshold, especially for non-stationary noise. This drawback is surmounted by applying the quantile estimate that

is actually a generalization of the minimum statistics approach in [7]. Both approaches are based on the fact that speech information does not always appear in all frequency channels simultaneously, even in speech intervals. While the minimum statistics method shows a bias estimate when tracking the minimum of the noisy signal power spectral density, the quantile filtering method avoids that problem by tracking the noise level determined by a q^{th} quantile selected out of a range $q = 0.0, 0.1, \ldots, 1$ over a window of the utterance.

In this paper, a sliding window of 960 ms length consisting of $N_f = 47$ overlapped speech frames is constructed. In order to track the fast changes of the non-stationary noise properly, instead of using recursive buffers proposed in [3], the window is slided over the whole utterance at a rate of one frame. The noise threshold is estimated for each sliding window successively. The estimated quantile noise threshold is then used to compress noise for every last speech frame in the windows. To carry out the estimation, firstly the values of perceptual noise thresholds $P_{j,i}$ derived from the N_f frames are stored in the b^{th} buffer. After sorting the threshold values for every CWS, we observe that the threshold values derived from non-speech frames occupy up to 60% of the buffer as depicted in Fig. 2. Thus, we carry out the estimate of quantile noise thresholds in two following steps:

- Sort $P_{j,i}$ in ascending order over the buffer b to get $P_{j,i'}$ with $i' = [1 \ldots N_f]$.
- Determine an adaptive threshold $\Gamma_j(b)$ by taking the q^{th} quantile as: $\Gamma_j(b) = P_{j,i'} \mid_{i'=\lfloor qN_f \rfloor}$

The quantile factor $q = 0.2$ was selected out of a candidate range $q = 0.0, 0.1, \ldots, 0.6$, as the value yields the best performance from our experiments. Under the assumption that the noise can not change faster than speech over time, the estimated quantile noise threshold is smoothed by applying a simple first-order recursive model:

$$\Gamma_{j,i}(b) = \alpha\Gamma_{j,i-1}(b-1) + (1-\alpha)\Gamma_{j,i}(b), \tag{8}$$

where the value of the forgetting factor α is set to 0.94.

2.3. Nonlinearly adaptive weighting

A refinement on the estimated quantile noise threshold is done by weighting it nonlinearly in both time and frequency domains as:

$$\tilde{\Gamma}_{j,i}(b) = \lambda_{j,i}(b)\eta_{j,i}(b)\Gamma_{j,i}(b), \tag{9}$$

$$\lambda_{j,i} = (a_1 P_{j,i})^{b_1} + d_1 \ , \tag{10}$$

$$\eta_{j,i} = (a_2 \Gamma_{j,i})^{b_2} + d_2 \ . \tag{11}$$

The temporal weighting function $\lambda_{j,i}(b)$ is built to track where speech and noise appear along the buffer. The frames with smaller estimated thresholds $P_{j,i}$ might correspond to noise and will undergo stronger thresholding. The frames with large $P_{j,i}$ values always contain more speech information and are treated in the reverse way to preserve speech quality. A frequency weighting $\eta_{j,i}(b)$ is introduced to take into account the correlations between packets at a certain frame that are not considered by the universal threshold. The function, which is designed to meet requirements of ASR, produces stronger weighting on the large quantile thresholds Γ_j stemming from wavelet packets containing large coefficients of speech and noise. Obviously, this results in a denoised speech which is less natural, but still intelligible at a very low remaining noise level. The constants $a_1 = 10, b_1 = -1, d_1 = 2.5$ and $a_2 = 64, b_2 = 0.5, d_2 = 0.1$ are selected manually from our experiments to achieve high performance of ASR.

Finally, an inverse mapping is implemented to provide the shrinking gain function for the estimated noise thresholds of all 128 wavelet packets as shown in Fig. 1.

3. SHRINKING GAIN FUNCTION

Wavelet shrinking is a promising tool to remove noise from an observed noisy signal. The principle is based on thresholding or shrinking the wavelet coefficients towards zero. Due to the decorrelation property of the WPD, the noise is spread out over all wavelet coefficients. This means the WPD leads to a sparse representation that allows replacing the noisy coefficients by zero. Hard and soft thresholding are proposed by [4] as the simple but sub-optimal denoising functions. An enhanced shrinking function used in [3] presents a smoothed hard thresholding based on the μ-law:

$$\widehat{X}_{k,i}(n) = \begin{cases} Y_{k,i}(n) \ , & \text{if } |Y_{k,i}(n)| > \tilde{\Gamma}_{k,i} \ , \\ \dfrac{\tilde{\Gamma}_{k,i}\,\mathrm{sgn}\{Y_{k,i}(n)\}}{\mu_{k,i}} A_{k,i}(n) \ , & \text{if } |Y_{k,i}(n)| \le \tilde{\Gamma}_{k,i} \ , \end{cases} \tag{12}$$

where $A_{k,i}(n)$ and the adaptive parameter $\mu_{k,i}$ are defined in [3] as:

$$A_{k,i}(n) = (1+\mu_{k,i})^{\frac{|Y_{k,i}(n)|}{\tilde{\Gamma}_{k,i}}} - 1 \ , \tag{13}$$

$$\mu_{k,i} = \exp\left(\beta \frac{\tilde{\Gamma}_{k,i}}{\max\limits_i\{\tilde{\Gamma}_{k,i}\}}\right) \frac{\max\limits_n\{|Y_{k,i}(n)|\}}{\tilde{\Gamma}_{k,i}} \ , \tag{14}$$

The shrinkage presents a compromise between soft thresholding (larger variance but smaller bias) and hard thresholding (higher bias but smaller variance). In other words, hard thresholding tends to keep closeness to the signal, while soft thresholding achieves smoothness of the signal. A big advantage of the μ-law shrinkage over others is that it does not strictly set to zeros all or parts of the wavelet coefficients, whose absolute values are below the threshold, as done by hard and soft thresholding [4]. Furthermore, it preserves the larger coefficients and has a smooth transition from noisy coefficients to signal coefficients. We improve this shrinking function by introducing the exponential term in Eq. 14 with a slope constant $\beta = 5.8$, which preserves more coefficients for the speech frames and removes more noise coefficients for the non-speech frames.

4. EVALUATION AND DISCUSSION

We carried out several tests to evaluate performance of the proposed PWF algorithm in terms of the recognition rates. The algorithm is compared with several well-known noise reduction methods which are embedded in different state-of-the-art speech recognizers. In this study, we use the HTK recognizer [8] for the tests on the German AURORA3 SpeechDat-Car corpus [1]. Both standard front-end (SFE) MFCC specified by [9] and advanced front-end (AFE) defined by [10] are examined with the HTK recognizer. The AURORA3 corpus consists of audio samples recorded with a close talking microphone and a hands-free microphone in a car environment with a variety of driving conditions. For testing on the SNOW-Factory corpus which was built during the SNOW project [2], the Loquendo speech recognizer [11] with general purpose acoustic models is applied. This second corpus was collected in the halls of the Airbus airplane maintenance factory. Background noise consists of strong stationary, non-stationary noise, and music noise.

In the first test, the proposed noise suppression algorithm is used as a pre-processing stage (i. e., without (wo.) retraining the acoustic models) for the HTK ASR system. For the "high-mismatch" (hm) training/test set of the first corpus, as presented in Table 2, the word recognition rate (WRR) is increased from 66.70% to 73.13% (word accuracy (WAC) from 63.23% to 70.77%) for the SFE. However, by

using the AFE, the WRR is reduced from 89.78% to 86.63% (WAC from 89.45% to 68.87%). We assume this decrease mainly due to an impact resulted from a double denoising by the proposed algorithm and the Wiener filtering in the AFE. By comparing with the results derived from the pWD method in [3], we see that recognition performance is mostly improved (7.53% WRR and 10.37% WAC for the SFE, and 1.53% WRR and 5.57% WAC for the AFE).

Table 2: Recognition performance as WRR/WAC using baseline (BSL) and the proposed PWF algorithms (Algs.).

Recognizer/Database			HTK/SpeechDat-Car	
Front-ends			SFE	AFE
Conditions	Algs.	Modes		
	pWD	wo.	65.60 / 60.40	85.10 / 77.30
	pWD	wi.	75.30 / 73.20	85.20 / 83.90
	BSL	-	66.70 / 63.23	89.78 / 89.45
hm	PWF	wo.	73.13 / 70.77	86.63 / 82.87
	PWF	**wi.**	**77.71** / **76.73**	**89.45** / **86.63**
	BSL	-	78.48 / 76.43	89.53 / 89.02
mm	PWF	wo.	69.03 / 56.99	86.75 / 57.83
	PWF	**wi.**	**81.92** / **78.99**	**88.65** / **85.29**
	BSL	-	90.48 / 87.92	95.55 / 94.65
wm	PWF	wo.	90.08 / 84.33	94.29 / 81.31
	PWF	**wi.**	**92.91** / **91.20**	**95.07** / **93.25**

In the second test, we consider a retraining of the acoustic model for the HTK recognizer. During the training phase, the PWF is used as a pre-processing stage for the SFE. The Wiener filtering is replaced by the PWF for the AFE. With (wi.) retraining mode, in the hm condition, the obtained WRRs are now almost similar to the ones of the BSL, e.g. 89.45% to 89.78% for the AFE while the WAC is slightly lower, e.g. 86.63% to 89.45%. For the SFE, the WRR and WAC are significantly increased to 11.01% and 13.5%. The recognition performance is much improved as compared to the one obtained from the pWD (12.11% WRR and 16.33% WAC for the SFE, and 4.25% WRR and 2.73% WAC for the AFE).

In the third test, we repeat the simulations done in the first and the second tests for medium-mismatch (mm) and well-match (wm) conditions. From Table 2, we observe that usage of noise reduction as a pre-processing stage contributes to the significant increase of the recognition performance if there is a high-mismatch condition (i. e., with quite clean samples used during training phase and very noisy samples for test phase). In the hm condition, the WRR is improved up to 11.11% (from 66.70% to 77.71%) while the WRR increase is smaller, 3.44% and 2.43%, for the mm and wm conditions, respectively. We also realize that the need of model retraining depends on the mismatch conditions no matter front-end units are used. Thus the need of retraining models depends on operating conditions.

In the final test, the PWF algorithm is tested with the Loquendo speech recognizer on the SNOW database. By replacing the elaborate spectral subtraction of Loquendo (SSL) [11] by the PWF, the WAC is slightly increased from 94.78% to 94.69%. In case of combining the end-point detector (EPD) that tries to identify the starting point and ending point of user utterance, the WAC is improved from 68.20% using the SSL to 88.98% when using the PWF. It is interesting that the PWF algorithm really helps the EPD to reduce the deletion of speech portions.

5. CONCLUSIONS

This paper clearly shows that, by applying a proposed perceptual wavelet filtering algorithm as a pre-processing stage, the robustness of ASR systems in harsh environments is increased.

By retraining the acoustic model of the recognizer with the proposed noise reduction algorithm, similar recognition performance is achieved using the ETSI 202 050 advanced front-end, and a significant improvement is obtained by using the ETSI 201 108 standard front-end. In addition, the proposed denoising algorithm really improves recognition performance in very adverse environments as aircraft maintenance factories. We attribute these attracted outcomes to the estimate of noise threshold on critical wavelet subbands and the excellent quantile filtering technique with its ability in accurately tracking the change of non-stationary noise levels. The results show that, the perceptual wavelet filtering algorithm is comparable to state-of-the-art noise reduction algorithms for robust ASR.

For the future research, a test on a wide range of background noise types allows for assessment of robustness of the proposed algorithm in ASR applications. Furthermore, because the distributions of speech information between different frequency channels are not the same, an intelligent quantile filtering with adaptive selection of quantile factors for different channels is motivated. Improvement of word accuracy will be considered.

6. REFERENCES

[1] "AURORA Project Database - Subset of SpeechDat-Car German database (AURORA/CD0003-03)," Evaluations and Language resources Distribution Agency, Tech. Rep., 2001.

[2] "Services for NOmadic Workers," European Commission. [Online]. Available: http://www.snow-project.org/

[3] E. Rank, T. V. Pham, and G. Kubin, "Noise suppression based on wavelet packet decomposition and quantile noise estimation for robust automatic speech recognition," in *Proc. ICASSP*, vol. 1, 2006, pp. 477–480.

[4] D. L. Donoho and I. M. Johnstone, "Adapting to unknown smoothness via wavelet shrinkage," *Journal of the American Statistical Association*, vol. 90, no. 432, pp. 1200–1224, 1995.

[5] Y. Shao and C.-H. Chang, "A generalized time-frequency subtraction method for robust speech enhancement based on wavelet filter banks modeling of human auditory system," *IEEE Trans. Systems, Man and Cybernetics, Part B*, vol. 37, no. 4, pp. 877–889, 2007.

[6] E. Zwicker and E. Terhardt, "Analytical expression for critical band rate and critical bandwidth as a function of frequency," *Journal of the Acoustical Society of America*, vol. 68, pp. 1523–1525, 1980.

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

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

[9] *ETSI ES 201 108 V1.1.1 Speech Processing, Transmission and Quality Aspects (STQ), Distributed speech recognition, Front-end feature extraction algorithm, Compression algorithms*, ETSI, 2000.

[10] *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*, ETSI, 2003.

[11] R. Gemello, F. Mana, and R. D. Mori, "A modified Ephraim-Malah noise suppression rule for automatic speech recognition," in *Proc. ICASSP*, vol. 1, Montreal, Canada, 2004, pp. 957–960.