16th European Signal Processing Conference (EUSIPCO 2008), Lausanne, Switzerland, August 25-29, 2008, copyright by EURASIP

SPEECH ENHANCEMENT USING A VARIABLE SUPPRESSION RULE IN HILBERT DOMAIN

M. Omidi, M.H. Savoji

Department of Electrical and Computer Engineering, Shahid Beheshti University
Evin, 1983963113, Tehran, Iran
phone: + (98)21-29902258, email: omidi_sbu@yahoo.com,m-savoji@sbu.ac.ir

ABSTRACT

We describe a speech enhancement system which combines a variable, input adaptive noise suppression rule with a recently developed spectral analysis framework in Hilbert domain. The variable suppression rule is an extension to a formula which encompasses well known noise reduction algorithms such as power subtraction and Wiener filtering. Time-varying parameters which are based on the input signal to noise ratio and its spectral shape are embedded in this formula. The framework consists of using Hilbert transform in sub-bands in conjunction with wavelet packet decomposition. This spectral analysis accounts for perceptual features and it was shown to be more effective than the common Fourier transform. The experiments show speech quality improvement in terms of perceptual measures.

1. INTRODUCTION

The noise reduction from human speech signal is an ongoing research subject with increasing applications. Generally, noise reduction algorithms consist of a spectral analysis section followed by a noise spectrum estimation procedure; and finally, an enhancement filter is applied to the noisy signal. Traditionally, spectral analysis is preformed by segmentation of the input signal and calculation of the short time Fourier transform. Noise is removed by modifying the frequency bins of every segment to achieve short time spectral noise attenuation. We focus on subtractive-type family of algorithms which attempt to enhance the short time spectrum of speech by subtracting a noise estimate from the input noisy signal. Due to low complexity and simplicity of implementation, this family of algorithms is widely used in speech enhancement systems. The major drawback of these types of algorithms is the residual noise, referred to as musical noise, introduced due to error in noise estimation. Besides, high degree of noise suppression, on the basis of this estimation, will bring about distortion of speech components. Thus, there is a trade-off between the amount of distortion and residual noise. Current research is focused on higher degree of suppression in low SNR (Signal to Noise Ratio) regions and lower suppression when a high SNR value is observed. According to this idea many variations have been proposed to subtractive algorithms, some embedding an adaptive parameter in the formula of these algorithms. We utilize a general formula devised in [1] which can take the form of a wide range of subtractive algorithms by changing a couple of parameters. This gives much flexibility and the opportunity of fast adaptation of the filter suppression rule based on the noisy signal.

Furthermore, we intend to use this adaptive suppression rule in Hilbert domain. Instead of using DFT filter banks, we use Perceptual Wavelet Packet Transform (PWPT) as introduced in [2] and [3] which was shown to be more in correlation with critical bands of human ear. Next, we utilize squared analytic signal envelope as a representation of sub-band power. This is based on the experiments reported in [4] and [5] which show that the analytic signal is more reliable than Fourier transform to reveal local variations in non-stationary signals. In fact, the psychoacoustic model presented in [6], states that a process similar to the analytic decomposition of a sound wave is performed in the basilar membrane.

The outline of this paper is as follows. In section 2 the spectral analysis including perceptual wavelet packet transform and analytic decomposition is explained in detail. Subsequently, we describe the proposed noise reduction technique in section 3. Experimental results are presented in section 4 and finally, we conclude the paper in section 5.

2. SPECTRAL ANALYSIS

2.1 Perceptual Wavelet Packet Transform

Instead of Fourier transform we employ discrete wavelet transform as a tool for spectral analysis in our implementation. Wavelet transform is suitable for the study of non-stationary processes since it lacks the limitation of fixed size transform window existing in short time Fourier transform. The regular dyadic wavelet decomposition results in logarithmic bandwidths of wavelet sub-bands. On the other hand, it has been shown that human ear acts similar to a filter bank with a particular structure. The frequency response of human ear is identical inside certain bands and different in others in terms of bandwidth. This structure is referred to as critical bands.

To obtain the critical band structure we must use the generalized form of wavelet transform called wavelet packet transform. 25 critical bands are determined for human ear in the range of 20Hz to 20 KHz. As we assume the input signal to be wideband speech (0-8KHz) no more than the first 19 critical bands are used in our work. We utilize a particular decomposition tree as depicted in Fig. 1. The frequency ranges for resulting sub-bands are illustrated in table 1. In

contrast to common methods, we may express a sub-band by joining a combination of outputs in different levels. In such cases, to obtain a single signal as a sub-band we reconstruct these different decomposition parts until they unite with the same sampling rates. For example, the low frequency component of sub-band 19 is up-sampled and filtered prior to being added to the higher frequency component.

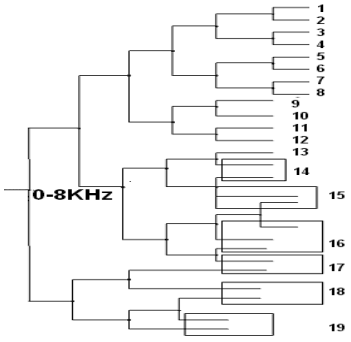

Figure 1 - Wavelet Packet Analysis Tree Corresponding to Used Critical Subbands

Table 1- Critical subbands

Band number	critical band range (Hz)
1	0-125
2	125-250
3	250-375
4	375-500
5	500-625
6	625-750
7	750-875
8	875-1000
9	1000-1250
10	1250-1500
11	1500-1750
12	1750-2000
13	2000-2250
14	2250-2750
15	2750-3125
16	3125-3750
17	3750-5000
18	5000-6500
19	6500-8000

Since the high and low band wavelet decomposition filters have some overlap, the approximation and detail signals bear some aliasing. The particular wavelet reconstruction pair of filters can be designed to cancel this effect when we synthesize the sub-bands using them. In our speech enhancement system the decomposed signals undergo some changes through the noise reduction formula. Thus the reconstructed signal will have some aliasing effect. To overcome this problem we must use filters with high selectivity and order to reduce their overlap. Besides, increasing the order of the filter is at the cost of higher computational load. Thus, there is a trade off between the filter order and the computational load. We have chosen *sav* wavelet filters primarily as introduced in [7] due to their sharpness of cut off. The explicit formula for the wavelet kernel is defined in (1). This formula is derived from the raised cosine function in frequency domain. Where f = 0 … k is the normalized frequency and k = N/2.

$$T(f)=\begin{cases}1 & |f|\le (r-\beta)\times k\\ \cos\left[\dfrac{\pi}{4\beta}\left(\dfrac{f}{k}-r+\beta\right)\right] & |f|\le (r+\beta)\times k\\ 0 & |f|>(r+\beta)\times k\end{cases} \tag{1}$$

N defines the frequency resolution and equals the filter order. To realize a half-band filter, r must equal 0.5. Using IFFT the filter coefficients are obtained and normalized. The wavelet kernel corresponding to this filter is called '*savK*' where K equals N/2.

We use sav16 as the suitable wavelet kernel with β parameter chosen to be 0.25 for sub-band decomposition. We also use these filters in analytic decomposition [7] as described in the next section. The sav filter has the advantage of being linear phase. In addition, it is easy to customize sav for use as a Hilbert transformer by changing the β parameter.

2.2 Analytic Decomposition

The analytic counterpart of a signal *s(t)* is defined to be:

$$s_a(t) = s(t) + j\hat{s}(t) \tag{2}$$

Where $\hat{s}(t)$ is the Hilbert transform of the signal. The analytic signal can be decomposed into instantaneous envelope and phase described by the following equations:

$$s_a(t) = a(t)e^{j\phi(t)} \tag{3}$$

Where

$$a(t) = |s_a(t)| \quad \phi(t) = \angle s_a(t) \tag{4}$$

The envelope of the analytic signal is of interest since it gives local information about the signal behavior, suitable for analysis of non-stationary processes. Among all methods of analytic decomposition, using a half-band filter is most suitable for our purpose. As described in [8], by filtering the input signal with a half-band low pass filter which is shifted to the right by $\pi/2$ the analytic counterpart of the real signal is obtained. This is illustrated in Fig 2.

We choose the low-pass half-band filter for analytic decomposition to be a *sav* filter. Note that the low-pass filter must equal 0.5 in the normalized frequency of 0.5, thus we must use $\pi/3$ instead of $\pi/4$ in (1). Furthermore, more accuracy and sharpness is needed and the β parameter must decrease. The value we use here is β = 0.01 with filter order being the same as wavelet kernel, 32. The coefficients of the filter are modulated by $\exp(-jn\pi f_s/4)$ to yield a $\pi/2$ shift in Fourier domain. The analytic signal is computed directly through convolving input signal with this complex filter.

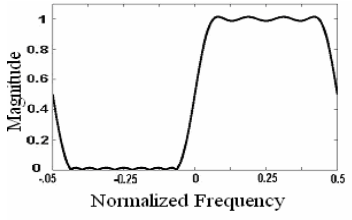

Figure 2 - A Sample Halfband Lowpass Filter Modulated by $\exp(-jn\pi f_s/4)$

3. ADAPTIVE SPEECH ENHANCEMENT SCHEME

3.1. General Noise Suppression Rule

Subtractive type speech enhancement techniques assume that the noise is additive and uncorrelated with speech. Thus, the noisy speech can be represented by

$$x(n) = s(n) + d(n) \tag{5}$$

The enhanced signal is computed in a frame-by-frame basis. As it is known that the human ear is not much sensitive to phase changes, the phase of the enhanced signal is left unchanged. For the particular case of power subtraction the suppression rule is formulated as (6).

$$|S(\omega)|^2 = \begin{cases} |X(\omega)|^2 - |\hat{D}(\omega)|^2, & if\ |X(\omega)|^2 > |\hat{D}(\omega)|^2 \\ 0 \quad otherwise \end{cases} \tag{6}$$

Where $|\hat{D}(\omega)|^2$ represents the noise power spectrum estimate. The method actually involves linear filtering in frequency domain. The frequency response of the filter is defined in every frame as follows:

$$H(\omega) = \sqrt{1 - \frac{|\hat{D}(\omega)|^2}{|X(\omega)|^2}} \tag{7}$$

A consideration in filter design is to choose an optimal filter characteristic. The concept of Wiener filter is the result of such consideration. The optimality of this filter is in finding the minimum of the mean square error defined as below:

$$E\left(\left(s(n) - \sum_{k=-\infty}^{\infty} h_k x(n-k)\right)^2\right) \tag{8}$$

Assuming normal distribution for speech and noise signals will end up in the following frequency response:

$$H(\omega) = \frac{|S(\omega)|^2}{|S(\omega)|^2 + |\hat{D}(\omega)|^2} \tag{9}$$

Another technique for noise suppression was developed by McAuley and Malpass [9] called maximum likelihood envelope estimation which models speech as a deterministic waveform of unknown amplitude. The corresponding filter is given as:

$$H(\omega) = \frac{1}{2} + \frac{1}{2}\left(\frac{|S(\omega)|^2 - |\hat{D}(\omega)|^2}{|S(\omega)|^2}\right)^{1/2} \tag{10}$$

If the suppression formulae for the above three methods are considered it becomes apparent that the Wiener filter is most suppressive and maximum likelihood yields the least suppression. The above algorithms are easily shown to be special cases of the generalized formula of the following equation (11). By changing $\alpha$and βparameters from 0 to 1 a wide range of filters are obtained.

$$S(\omega) = \left[(1-\alpha) + \alpha\left(\frac{|X(\omega)|^2 - |D(\omega)|^2}{|X(\omega)|^2}\right)^{\beta}\right].X(\omega) \tag{11}$$

α = ½, β = ½ … maximum likelihood
α = 1, β = ½ … power subtraction
α = 1, β = 1…. Wiener filter

3.2. Selection of Parameters

The selection of α and β parameters can be adaptively related to the noisy input signal condition. Once the SNR of the signal is low, it is desirable that the filter functions similar to a suppressive filter such as Wiener filter and in case of high SNR the filter magnitude should increase to close to unity to avoid distortion of speech components.

By inspecting (11), it is apparent that α parameter controls the filtered portion of the signal. We choose noise to signal ratio (NSR) as a proper value for α. It is calculated on a frame-by-frame basis and its maximum value is set to 1.

$$\alpha = \min\left(\frac{|\hat{D}(\omega)|^2}{|X(\omega)|^2}, 1\right) \tag{12}$$

The parameter β affects the amount of suppression applied at a given frequency bin. We intend to make it speech content dependent and choose the filter suppression to be higher when the signal spectrum is flat and represents a noise-like spectrum, and a lower suppression is required when the detected spectrum is tone-like. Among various mathematical descriptors of audio and speech signal features investigated in [10], spectral flatness measure (SFM) and spectral crest measure (SCM) are most closely related to our requirements. Equation (13) defines the spectral flatness as the ratio of the geometric mean to the arithmetic mean of the energy spectrum:

$$SFM(num\_band) = \frac{\left(\prod_{k \in num\_band} a(k)\right)^{1/k}}{\frac{1}{k}\sum_{k \in num\_band} a(k)} \tag{13}$$

Where *a(k)* is the amplitude in frequency band number *k*. for tonal signals SFM is close to 0 whilst for noisy signals it is close to 1. The other spectral shape descriptor is spectral crest factor which is computed by the ratio of the maximum value within the band to the arithmetic mean of the energy spectrum value.

$$SCM(num\_band) = \frac{\max(a(k \in num\_band))}{\frac{1}{k}\sum_{k \in num\_band} a(k)} \tag{14}$$

Thus inverse SCM (SCM^{-1}) seems to be a possible value for β. However, unlike the SFM, the SCM theoretically has no upper bound and using SCM^{-1} in place of β means the filter magnitude can be irregularly small. Thus, similar to what we did for α, we limit the minimum value of SCM^{-1} to be 0.1.

3.3. Noise Reduction Framework

In this subsection the implementation details of the algorithm is described. The block diagram of the system is shown in Fig.3. Through some experiments, it was observed that a window length of 64ms time length with 50% overlap is suitable. The following step is to decompose the frame to its critical band components using PWPT. The 19 obtained sub-bands are passed to the analytic decomposition section. Similar to most methods, we do not modify the phase of the signal. Prior to applying the suppression rule, we utilize minimum tracking algorithm for noise estimation as described in [11] which is suitable for slowly varying noise

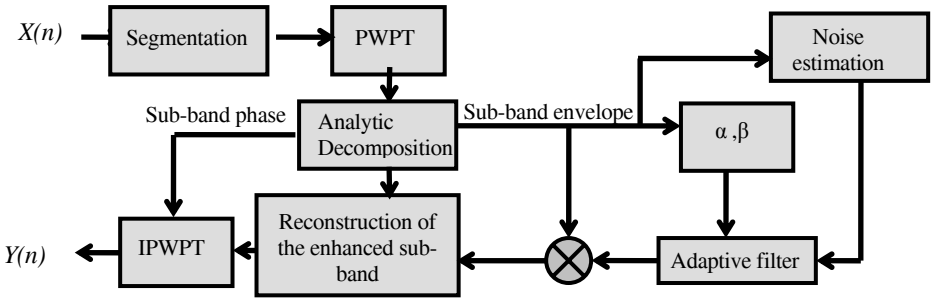

Figure 1 - Framework of the Proposed Noise Reduction System

level. Here we use squared envelope to represent the short time power of the sub-band.

The minimum tracking algorithm tracks minima of a smoothed periodogram. The smoothing process is preformed with a recursive iir filter:

$$P_x(\lambda,k) = \delta P(\lambda-1,k) + (1-\delta)\,|\,X(\lambda,k)\,|^2 \qquad (15)$$

Where $|X(\lambda,k)|^2$ represents the short time power of the sub-band k in frame λ. The smoothing coefficient δ is constant and proportional to the sampling frequency of the sub-band. The next step is to compute the parameter α of equation (11). We use estimated noise power along with the smoothed periodogram derived from (15) in (12). To compute the proper value for β based on abovementioned formulation of spectral shape descriptors we calculate the spectral measure by employing samples of $|X(\lambda,k)|^2$ in frame number λ *as* $a(k)$ in expression (13) or (14) . Next we apply the filter based on the general rule of (11). We also use a floor constant as defined by Berouti [12] for the lower bound of the filter. Subsequently, we reconstruct the sub-bands with the enhanced envelope and the phase of noisy input. The final step is to join the sub-bands through the synthesis step phase of the inverse wavelet packet transform.

4. EXPERIMENTAL EVALUATION

To perform objective evaluations of our method we used different measures. A segmental version of the objective measure proposed in an ETSI's standardization project [13] was used. This measure computes the segmental SNR improvement (SNR_{Imp}) separately in three energy classes of speech: high, medium and low energy. The final value is computed by averaging over the three classes.

Segmental SNR_{Imp} improvement in each class is defined as:

$$SNR_{imp,C} = \frac{10}{M_C}\sum_{m=0}^{M_C-1}\log_{10}\frac{\sum_{n=mN}^{(m+1)N-1} n_C^2(n)}{\sum_{n=mN}^{(m+1)N-1}\left(s_c(n)-\hat{s}_c(n)\right)^2} \qquad (16)$$

in which n_c, s_C and $\hat{s}_C$ are noise, clean and reconstructed signals in specified class, C. M_C is the number of segments of class C, and N is the length of each segment and is set to 320 samples at the sampling rate of 16 KHz.

Besides, we employ the ITU standard, PESQ (Perceptual Evaluation of Speech Quality) [14] which predicts subjective MOS for a variety of speech distortions in communication systems. The investigations in [15] show that this measure is most correlated with subjective judgments of quality degradation of speech signal.

Wideband speech signals sampled at 16 KHz were employed as test signals to evaluate the performance of the system. We used 12 sentences from 6 male and 6 female speakers. We added white Gaussian, babble and destroyer engine noise to these sentences at different SNR levels. The results are shown (as average on all files both in terms of subjective measure of MOS and objective measure of SNR improvement) in table 2. These results show a significant improvement especially in terms of predicted PESQ MOS scores when using our parametric filter varying both α and β (employing SCM^{-1}) in comparison with Wiener and Power Subtraction method.

Table 3 presents the predicted MOS measures when SFM and SCM^{-1} are utilized in the general rule and compares the result with the case of fixed $\beta = 1$ (α is kept constant and equal to 1 in these experiments). In most cases an improvement is observed. Nevertheless, it can be seen that SCM^{-1} gives slightly better results. In fact using SCM^{-1} makes the algorithm biased toward less distortion especially in speech segments while SFM guarantees more suppression in case a relatively flat spectrum is detected. This is a justification for its use in our parametric filter. It is noted that it was first established that using a parametric filter changing α only (with $\beta = 1$) was beneficial.

Table 2 - Results for three types of noise: White Gaussian, Babble and Destroyer engine noise

	SNR	20dB	10dB	5dB	3dB
WG	MOS parametric	3.42	2.80	2.39	2.28
	MOS power sub	3.11	2.53	2.29	2.25
	MOS wiener filt	3.38	2.76	2.37	2.29
	SNR_{Imp} parametric	1.08	1.27	1.93	2.88
	SNR_{Imp} power sub	0.72	0.89	2.11	2.35
	SNR_{Imp} wiener filt	1.11	1.85	2.07	2.79
BAB	MOS parametric	3.11	2.37	1.92	1.79
	MOS power sub	2.84	2.25	1.79	1.74
	MOS wiener filt	3.09	2.31	1.81	1.72
	SNR_{Imp} parametric	0.67	0.96	1.24	2.25
	SNR_{Imp} power sub	0.35	0.46	1.57	1.98
	SNR_{Imp} wiener filt	0.72	1.13	1.34	2.08
DEST	MOS parametric	3.28	2.53	2.15	1.98
	MOS power sub	2.93	2.27	2.09	1.69
	MOS wiener filt	3.25	2.41	2.12	1.93
	SNR_{Imp} parametric	0.95	1.15	1.74	2.75
	SNR_{Imp} power sub	0.58	0.72	1.96	2.23
	SNR_{Imp} wiener filt	0.98	1.57	1.77	2.58

Table 3 - Results for three types of noise comparing β measure

	SNR	PESQ SFM	PESQ SCM^{-1}	PESQ β =1	SNR SFM	SNR SCM^{-1}	SNR β =1
WG	20dB	3.12	3.21	3.27	0.75	1.09	0.38
	10dB	2.75	2.74	2.68	0.93	1.63	0.67
	5dB	2.38	2.42	2.28	1.12	0.98	1.11
	3dB	2.25	2.35	2.14	2.04	2.79	1.59
BAB	20dB	2.84	2.57	2.75	0.34	0.23	0.17
	10dB	2.43	2.23	2.38	0.56	0.54	0.42
	5dB	2.18	2.05	2.12	0.71	0.67	0.74
	3dB	2.09	1.89	1.91	1.58	2.12	1.09
DEST	20dB	3.09	2.84	3.01	0.58	0.82	0.27
	10dB	2.69	2.38	2.45	0.76	1.23	0.58
	5dB	2.32	2.13	2.21	0.98	0.79	0.93
	3dB	2.17	2.05	2.01	1.78	2.37	1.34

As far as complexity is concerned our previous experiments [5] comparing Martin's spectral subtraction algorithm [11] implemented in Fourier domain with our implementation in Hilbert domain and critical sub bands, showed that our algorithm not only gives better results but is even faster in terms of cpu time usage; although it seemed that the particular implementation used [16] may not have been optimized for speed. It is worth mentioning that the parametric implementation presented here does not add considerably to the complexity of our previous implementation.

5. CONCLUSION

In this research we established a speech enhancement system based on a general rule which is adaptable to the input noisy signal. Time-varying parameters which are related to SNR and spectral shape of noisy speech control the shape of the filter. It was observed that this selection results in less musical noise and distortion. Furthermore we combined this enhancement system with a spectral analysis scheme to reduce the effects of the residual noise. We employed perceptual features of wavelet and Hilbert transform along with their capabilities to reveal local, non-stationary characteristics of speech signal. A comparative study through different measures revealed improved quality of speech by the application of the proposed method.

We intend to evaluate and compare the performance of newer algorithms such as MMSE Log Spectral Amplitude Estimator with the algorithm presented here in the future.

6. ACKNOWLEDGMENT

The authors wish to thank the PhD student Ms. Sh. Gholami for her help in getting the final results in this study.

REFERENCES

[1] M.H. Savoji, “Effective noise reduction of speech signals using lattice filtering, segmentation and soft decision”, IEE colloquium on adaptive signal processing, London, UK, February 1993.

[2] I.Cohen, "Enhancement of Speech Using Bark-Scaled Wavelet Packet Decomposition", Eurospeech, 2001.

[3] T. Guelzow et al, "Spectral-subtraction speech enhancement in multirate systems with and without non-uniform and adaptive bandwidths", Signal Processing, Vol. 83, issue 8, Aug. 2003.

[4] N. Derakhshan, and M.H. Savoji, “Perceptual speech enhancement using a hilbert transform based time-frequency representation of speech,” 11th Int. Con. of Speech and Computer , St. Petersburg, 25-29 June 2006.

[5] M. Omidi, N. Derakhshan and M.H. Savoji, "The advantage of implementing Martin's noise reduction algorithm in critical bands using wavelet packet decomposition and Hilbert transform", CSICC2008, Kish Island, Iran, 9-11 March 2008.

[6] Flanagan, J.L., “Phase vocoder,” The Bell System Technical Journal, Vol. 45, pp. 1493-1509, 1966.

[7] M. Omidi, M.H. Savoji "A new hilbert transformer based on parametric wavelet kernel application to analytic signal decomposition of speech subband", 5th Int. Symposium on Image and Signal Processing and Analysis, ISPA 2007, Istanbul, Turkey, Sep. 2007.

[8] A. Reilly, G. Frazer and B.Boashash, “Analytic signal generation - tips and traps,” IEEE Trans. Signal Processing, Vol. 42, No.11, pp. 3241-3245, Nov. 1994.

[9] R. J. McAulay, M.L. Malpass, “Speech enhancement using a soft-decision noise suppression filter,” in Magnetism, vol. III, G. T. Rado and H. Suhl, Eds. New York: Academic, 1963, pp. 271–350.

[10] G. Peeters, “A large set of audio features for sound description (similarity and classification)” CUIDADO project CUIDADO I.S.T. Project Report 2004.

[11] R. Martin. “Spectral subtraction based on minimum statistics”. In Proc EUSIPCO, pages 1182-1185, Edinburgh, Sept 1994.

[12] M. Berouti, R. Schwartz, J. Makhoul, 1979. “Enhancement of speech corrupted by acoustic noise” Proc. ICASSP-79, pp. 208–211.

[13] 3rd Generation Partnership Project (3GPP™), “Technical specification group services and systems aspects, Minimum Performance Requirements for Noise Suppresser Application to the AMR Speech Encoder,” Technical report 3GPP™ GSM Technical Specification TS 26.077, 3rd Generation Partnership Proj., 1999.

[14] Perceptual Evaluation of Speech Quality (PESQ), an objective method for end-to-end speech quality assessment of narrowband telephone networks and speech codecs. ITU-T Recommendation P.862, February 2001.

[15] Marzinzik, 2000. Noise reduction schemes for digital hearing aids and their use for the hearing impaired. Ph.D. dissertation, University of Oldenburg.

[16] 'Voicebox' speech procesing toolbox for MATLAB: http://www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html