SINGLE MICROPHONE WIND NOISE REDUCTION USING TECHNIQUES OF ARTIFICIAL BANDWIDTH EXTENSION

Christoph Matthias Nelke, Niklas Nawroth, Marco Jeub, Christophe Beaugeant[⋆]*, and Peter Vary*

Institute of Communication Systems and Data Processing (ind)
RWTH Aachen University, Germany
[⋆]Intel Mobile Communications, Sophia-Antipolis, France
{nelke,jeub,vary}@ind.rwth-aachen.de christophe.beaugeant@intel.com

ABSTRACT

In this contribution, we propose a method to enhance single channel speech signals which are degraded by wind noise. In contrast to common speech enhancement systems, a special processing is required due to the highly non-stationary characteristics of wind signals. The basic idea is to exploit the fact that wind noise is mainly located at low frequencies and thus, a large frequency range of the speech is almost noise free. Techniques which artificially extend the bandwidth of telephone speech towards lower frequencies are applied to replace the highly disturbed low frequency parts. Here, the discrete model of speech production is used to reconstruct the required parts of the speech signal. Important parameters for this model are pitch frequency, the spectral envelope and a spectral gain. In this context, an evaluation is carried out which determines the robustness of several pitch estimators against wind noise. The frequency range of the reconstructed speech is finally adapted to the actual level of wind noise. Based on realistic scenarios it is shown that the influence of the wind noise can greatly be reduced by the proposed concept. This includes a comparison with a state-of-the-art speech enhancement system and an algorithm specially designed to reduce wind noise.

1. INTRODUCTION

Nowadays, mobile communication devices can be used in almost any acoustic environment. This leads to the drawback that the quality and intelligibility of the captured speech signals can be greatly degraded by interfering background noise. Most of the occurring noise types such as inside car noise, babble noise or traffic noise can be assumed to be stationary over a certain period of time. In contrast to that, wind noise is characterized by a high degree of instationarity. Well-established speech enhancement systems for single channel signals apply spectral weighting based on an estimate of the short-term power spectral density (PSD) of the noise signal (e.g. [1], [2]). State-of-the-art algorithms for the estimation of the noise PSD can be found e.g. in [3], [4]. Despite, these methods are able to track time-varying noise signals, for wind noise they deliver insufficient estimates.

Hence, the reduction of wind noise requires a special class of speech processing. In [5] a dual channel system was presented which exploits the low correlation of the wind signal between the channels. However, for many applications the use of more than one microphone is not feasible. Single channel methods, which were proposed in the past based on spectral weighting without directly estimating the noise PSD can be found in [6], [7].

In contrast to these speech enhancement methods, our new approach completely removes distorted parts from the speech signal and replaces them by artificially generated speech samples. Techniques which are used for artificial bandwidth extension are applied here, e.g. [8]. The aim of these techniques is to blindly estimate a wideband signal (50-7000 Hz) from a given narrowband telephone signal (300-3400 Hz). This is mainly related to the reconstruction of the higher frequency range (3400-7000 Hz). Figure 1 shows the long-term power of speech and wind signals averaged over a duration of 30 seconds. Based on the property that the wind noise is mainly located at low frequencies (0-500 Hz), our method uses parameters extracted both from the higher frequency range and the noisy spectrum to extend the undistorted speech towards lower frequencies. The artificially generated speech is provided by a discrete time model of speech production, e.g. [2].

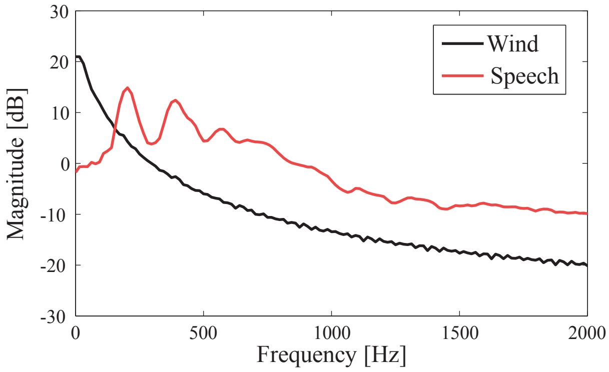

Fig. 1. Long-term power spectra of speech and wind signal

The remainder is organized as follows. First, the discrete model of speech production is introduced, followed by an overview of the proposed system for wind noise reduction. Section 4 shows the estimation of the parameters for the synthesis of the low frequency speech parts. For the adaptation of the system to the actual wind noise power, a wind noise detection method is presented in Section 5. An evaluation based on real measurements is given in Section 6. This contribution is finally summarized in Section 7.

© EURASIP, 2012 - ISSN 2076-1465

2. DISCRETE MODEL OF SPEECH PRODUCTION

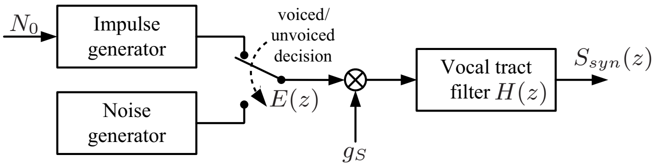

Fig. 2. Discrete model of speech production (c.f. [2])

A commonly used model for the process of speech production is given by the system shown in Fig. 2 (cf. [2]) which provides a synthetic speech signal in the z-Domain

$$S_{syn}(z) = g_S \cdot E(z)H(z). \tag{1}$$

Based on the type of the speech segment the time domain excitation signal $e(k)$ is generated, where k is the discrete time index. For voiced speech segments a periodic impulse sequence

$$e_{voiced}(k) = \sum_{i=-\infty}^{+\infty} \delta(k - iN_0) \tag{2}$$

and for unvoiced segments a white noise signal is used as excitation signal. N_0 is referred to the pitch period and defines the time lag of the impulses δ in the voiced speech segments. The gain factor g_S determines the amplitude of the excitation signal. The human vocal tract is modeled by filtering the excitation signal with the time-varying autoregressive synthesis filter $H(z)$.

3. SYSTEM OVERVIEW

The structure of the proposed system is depicted in Fig. 3. The noisy input signal $x(k)$ is assumed to be an additive superposition of the clean speech signal $s(k)$ and the wind noise $n(k)$. The entire system applies a frame-based signal processing by first segmenting the input signal into overlapping frames with index λ, reducing the wind noise and then constructing the output signal $\hat{s}(k)$ via overlap-add.

The noise-free part of the input signal $x_{HP}(\lambda)$ pass the adaptive high-pass filter with a variable cut-off frequency f_c in the upper branch of the system. The lower branch represents the artificial bandwidth extension. First the parameters N_0, a, g_S of the speech production model are estimated from the prefiltered noisy signal $\tilde{x}(\lambda)$ and then a synthetic signal $s_{syn}(\lambda)$ is generated using the model introduced in Section 2. In [8] this model was chosen for an artificial extension of narrowband telephone speech towards higher frequencies. Essential parameters for this model are the pitch period N_0, the linear predictive coding (LPC) coefficients $a_1, ..., a_M$ for the vocal tract filter $H(z)$ and a gain factor g_S. The prefilter in Fig. 3 is realized by a fixed high-pass filter which reduces the influence of the wind noise on the parameter estimation.

The spectral power distributions of wind noise and unvoiced speech segments shows only a minor overlap and can thus be suppressed by the adaptive high pass filter. Hence, in our system the high-pass filtered signal $x_{HP}(\lambda)$ provides a good noise reduction for unvoiced speech segments. Besides, the wind signal as well as the unvoiced speech are both noise like signals. Consequently, a certain amount of residual noise leads to no severe degradation of the unvoiced speech. Thus, the speech production model of Fig. 2 is only used for the generation of voiced speech segments. The voiced/unvoiced decision can be derived from the pitch estimators presented in 4.1. For a voiced excitation signal equally spaced pulses are generated in the time domain based on the actual pitch period using Eq. 2.

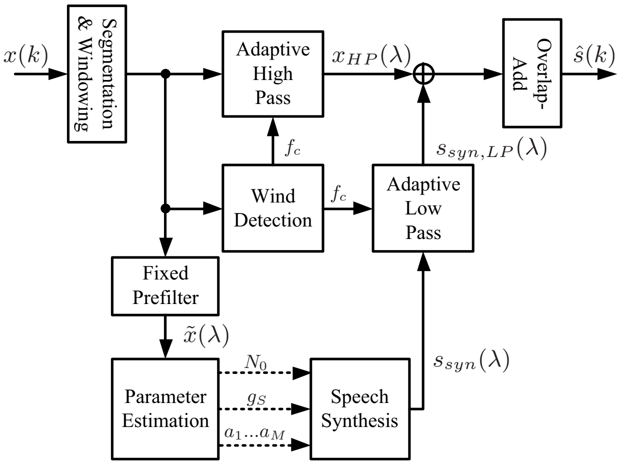

Fig. 3. Proposed wind noise reduction system

Finally, the high-pass filtered parts of the original signal $x_{HP}(\lambda)$ and the low-pass filtered synthetic signal $s_{syn,LP}(\lambda)$ are merged to provide an enhanced speech signal as depicted in Fig. 4. The red curve presents the artificially generated speech signal while the blue curve shows the unprocessed input signal. Both the high-pass and the low-pass filter are realized as FIR filters with complementary passbands. The wind noise detector in Fig. 3 determines the variable cut-off frequency f_c between these two parts based on the actual power of disturbance. Thus it can be guaranteed that the speech signal will not be modified by the system in noise-free conditions.

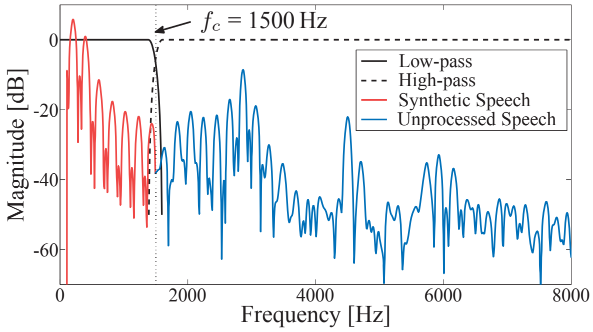

Fig. 4. Merging of synthetic and original speech signal

4. PARAMETER ESTIMATION

4.1. Pitch Period

A broad variety of algorithms to determine the pitch period in speech signals exists, cf. [9]. Due to the frame based processing of the signal only short-term pitch estimators are considered here. These methods calculate the pitch period of short signal segments. Investigations on the robustness of pitch estimators in terms of additive noise are mostly carried out with white Gaussian noise, cf. [10], [11], [9]. Results from this investigations are not valid here because of the non-whiteness of wind signals. A performance study of several pitch estimators is carried out in this section. The goal is to identify the method which shows the most reliable results during the occurrence of wind noise. Subsequently, a voiced/unvoiced classification and a smoothing of the pitch estimate is proposed.

It turns out that during the occurrence of wind noise algorithms working in the time domain are much more error-prone compared to frequency domain methods. Therfore three frequency-domain estimators and one time-domain estimator used as a reference are investigated in more detail in the following.

- **CEP:** In [12], a pitch estimator working in the cepstrum is proposed. This transform leads to a separation of the excitation signal and spectral envelope described above. The lower cepstral coefficients represent the spectral envelope while the pitch frequency and its harmonics are mapped to a single higher cepstral coefficient. This results in a local maximum in the cepstrum. Furthermore, investigations have shown that the wind noise mainly effects the lower cepstral coefficients which correspond to the spectral envelope of the speech signal.
- **HPS:** The method proposed in [13] called *Harmonic Product Spectrum* takes advantage of the harmonic structure of the voiced segments of a speech signal. In this case the signal spectrum consists of the pitch frequency and equally spaced harmonics. Although the pitch frequency is often covered by the low frequency wind noise the higher harmonics are mostly undistorted. The HPS is defined by a weighted product of the harmonics for a potential pitch frequency and the actual short-term spectrum. The final estimate is then given by the frequency which maximizes the HPS.
- **PEFAC:** Similar to the HPS, the *Pitch Estimation Filter with Amplitude Compression* exploits the harmonic structure of the speech signal [11]. Here, a convolution of the spectrum with a filter function is applied. The filter function is constructed in a way that all harmonic peaks are mapped to the pitch frequency. In addition, the amplitude of parts of the spectrum is compressed in order to suppress narrowband noise.
- **AUTOC:** Besides, an autocorrelation based time domain algorithm [9] is evaluated. The pitch estimate is provided by the local maximum in a fixed search range of the autocorrelation. Other time domain methods as [14] or [10] show similar results.

The described pitch estimators are evaluated with 5.6 minutes of speech signals taken from [15]. This database provides real pitch measured by a laryngograph as reference. The speech samples were superposed by wind noise from real recordings. More details regarding the recordings will be given in Sec. 6. In literature a common measure to evaluate pitch estimators is the gross-error rate (GER) which determines the number of pitch estimates which deviate more than 20 % from the real pitch frequency [9]. Fig. 5 depicts the GER for the four different estimators for different SNR values. For all estimators the range of possible pitch estimates was limited to 50-400 Hz which is the normal frequency range for both female and male speakers [9]. It can be seen that the methods operating in the frequency domain clearly outperform the time domain algorithm (AUTOC) in terms of estimation accuracy for SNR values below 10 dB. In total the cepstrum based method provides the most reliable pitch estimate.

In the system shown in Fig. 3 the pitch estimate provided by the cepstrum based method is used to produce the required excitation signal for voiced speech segments. To reduce the effect of single outliers the actual estimated pitch frequency $\tilde{f}_0$ is smoothed over time using

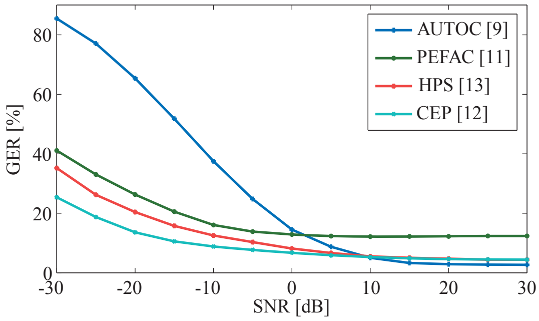

Fig. 5. Gross-error rate of pitch estimators

$$f_0(k) = \alpha_p \cdot f_0(k-1) + (1-\alpha_p) \cdot \tilde{f}_0(k). \qquad (3)$$

The CEP pitch estimator is based on a local maximum search. By means of the "peakedness" of this maximum a voiced/unvoiced classification can be made for the speech synthesis. This is derived by comparing the local maximum with an adjacent range in the cepstral domain. The ratio V of the local maximum and the average value of the search range $c_{min}, ..., c_{max}$ of the complex cepstrum $x_{cc}(c)$ is applied is a detection method for voiced speech method

$$V = \frac{\max\{|x_{cc}(c_{min}, ..., c_{max})|\}}{\text{mean}\{|x_{cc}(c_{min}, ..., c_{max})|\}}. \qquad (4)$$

If V exceeds a certain threshold V_{th} the actual frame is classified as voiced.

4.2. Vocal Tract Filter

As depicted in Fig. 2, the filter $H(z)$ represents the human vocal tract which is often approximated as an all-pole filter. A common way to estimate this filter is to compute the LPC coefficients using the Levinson-Durbin algorithm [2]. Fig. 6 shows the LPC spectra of a voiced speech segment. The dashed lines illustrate the influence of the wind noise on the spectral envelope. The clean input signal is depicted by the black solid line. The LPC spectra of the unprocessed noisy speech and the prefiltered noisy input are represented by the dashed black and red lines, respectively. The momentary SNR in this segment is -5 dB.

It can be seen that the wind noise effects mainly the low frequency parts of the envelope and parts at frequencies higher than 4000 Hz. The low frequency wind noise disturbance can easily be suppressed by a high-pass filter which is realized by a fixed prefilter as shown in Fig. 3. The high-frequency mismatch can be neglected because these parts of the speech will pass the system without any modification through the upper branch in Fig. 3. For the proposed method we used the LPC coefficients of the prefiltered noisy input signal. It turns out that using a high-pass filter with a constant cut-off frequency of 200 Hz results in a sufficient estimate of the LPC coefficients from the noisy speech although, there are some deviations in for extremely low frequencies (below 100 Hz).

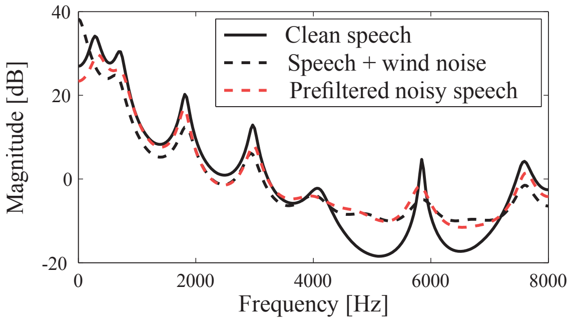

Fig. 6. LPC spectra of a voiced speech segment, SNR = -5 dB

4.3. Gain Factor

The gain factor introduced in Fig. 2 controls the power of the excitation signal. Ideally, the power of the reconstructed excitation signal should be equal to the power of the excitation signal of the clean speech signal. The residual signal of the LPC analysis described in Sec. 4.2 is used for as an estimate of the clean excitation signal power. Again, the prefiltered speech $\tilde{x}(\lambda)$ is used to reduce the effect of the wind signal. The constant factor γ is introduced to compensate the high-pass effect on the power of the excitation signal.

$$g_S = \gamma \cdot \sqrt{\frac{\sum_{k=0}^{M-1} \tilde{e}_{HP}^2(k)}{\sum_{k=0}^{M-1} e_{syn}^2(k)}}. \tag{5}$$

5. WIND NOISE DETECTION

The detection of the occurrence of wind noise is necessary to adjust the range of reconstructed speech. The intention is to determine the disturbed frequency range in the actual frame. As shown in Fig. 1, the wind noise signal has the main power distribution at low frequencies which is rapidly descending towards higher frequencies. Thus, using the power ratio between a low frequency range up to f_H and the whole frequency range is used as a wind indicator. This leads to the heuristically chosen equation for the cut-off frequency f_c between the original and the reconstructed speech, where μ_H is the discrete frequency bin corresponding to f_H:

$$f_c(\lambda) = f_{max} \cdot \frac{\sum_{\mu=0}^{\mu_H} |X(\mu,\lambda)|^2}{\sum_{\mu=0}^{M} |X(\mu,\lambda)|^2}. \tag{6}$$

The upper bound for the cut-off frequency is given by f_{max} which determines the maximum frequency range of the artificially generated speech. To prevent artefacts caused by sudden changes of the cut-off frequency, f_c is smoothed over time with a smoothing factor α_c.

6. EVALUATION

6.1. Experimental Setup

The proposed method is tested with wind noise recorded with a mock-up mobile phone mounted in the hand-held position on an artificial head (HEAD acoustics HMS II.3 & HHP III). The recordings were taken on a windy day with wind speeds up to 50 km/h on a roof terrace. The wind noise was captured by a microphone (Beyerdynamic MM1) assembled to the mock-up phone without any kind of wind screen. In order to have a reference for the evaluation the noisy speech was generated by a superposition of clean speech from [15] with the wind noise recordings. This may not represent non-linear effects of wind noise such as clipping of the audio signal.

The frame length of the overlap-add system was set to 20 ms with 50 % overlap. For a precise pitch estimation longer frames of 50 ms were used for the cepstrum pitch estimator. The threshold V_{th} for the voiced/unvoiced classification was set to 5. The order of the LPC vocal tract filter was 20. Investigations showed that the correction factor for the gain factor in Eq. 5 should be chosen between 0.5 and 2 and was set to 0.5 for the evaluation. The smoothing factors α_p and α_c were chosen to 0.3 and 0.8, respectively. The upper bound for the reconstructed speech was set to f_{max} = 1500 Hz. The frequency f_H for the wind noise detection in Eq. 6 was set to 100 Hz. The sampling frequency was 16 kHz.

6.2. Results

In Fig. 7 the spectrograms of noisy and enhanced signals are shown at an input SNR of -5 dB. For a better view on the wind noise effects only frequencies up to 4 kHz are depicted. It can clearly be seen that signal parts influenced by the wind noise are removed in speech pauses and replaced by the synthetic signal during speech activity, respectively.

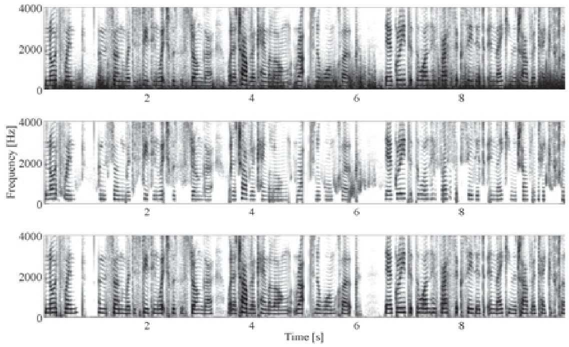

Fig. 7. Spectrograms of unprocessed (top), enhanced (middle) and clean (bottom) speech at a SNR of -5 dB

Because of the non-linear filtering of the proposed method common quality measures like speech and noise attenuation can not be applied. The evaluation is carried out with the Speech Intelligibility Index (SII) [16] and the improvement of Perceptual Evaluation of Speech Quality (ΔPESQ $=$ PESQ$_{out}$ $-$ PESQ$_{in}$) [17]. The SII provides a value between 0 and 1 where a SII higher than 0.75 indicates a good communication system and values below 0.45 correspond to a poor system. The averaged results for different input SNRs are shown in Fig. 8.

The proposed method is compared with a standard speech enhancement system with a MMSE estimator for both the noise PSD [3] and the clean speech DFT coefficients [1] and a method which was explicitly designed for wind noise reduction [7]. The latter uses an adaptive postfilter concept based on low order LPC spectra of speech and wind. The chosen range of the input SNR for the evaluation reflects wind noise levels which occur during real outdoor measurements.

The evaluation shows an enhancement in terms of intelligibility for all algorithms over the whole SNR range. Among the inves-

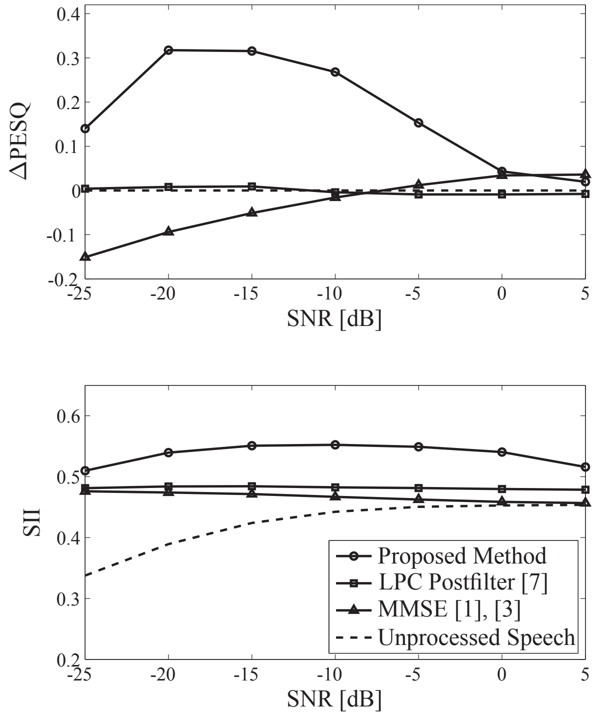

Fig. 8. ΔPESQ and SII of unprocessed and enhanced speech

tigated methods the proposed system achieves the highest SII improvements. The ΔPESQ score shows only an improvement for the proposed algorithm. This especially applies for the SNR range of -20 to 0 dB, which describes realistic scenarios. Here, the MMSE noise reduction even shows a degradation of the quality expressed by the negative values. Informal listening tests confirmed the results of this experiments. Nevertheless, there are some audible artefacts resulting from the synthetic speech which may occur in very low SNR conditions when large parts of speech have to be reconstructed.

7. CONCLUSIONS

In this contribution, we have introduced a single microphone system for the reduction of wind noise. It was shown that state-of-the-art methods, which require an accurate noise estimate have only a limited performance of enhancing the signal because of the non-stationary characteristics of wind noise. Our method uses techniques from artificial bandwidth extension to replace the disturbed parts of the speech signal. In this context, the performance of several pitch estimators was investigated. Experimental results with real wind noise measurements showed that the proposed system outperforms both a standard system given by [3] and [1] and a method explicitly designed for wind noise reduction [7].

8. REFERENCES

[1] J.S. Erkelens, R.C. Hendriks, R.. Heusdens, and J.. Jensen, "Minimum mean-square error estimation of discrete fourier coefficients with generalized gamma priors," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 15, pp. 1741 –1752, 2007.

[2] P. Vary and R. Martin, *Digital Speech Transmission. Enhancement, Coding and Error Concealment*, Wiley-VCH Verlag, 2006.

[3] R.C. Hendriks, R. Heusdens, and J. Jensen, "MMSE based noise PSD tracking with low complexity," in *Proc. of IEEE Intern. Conf. on Acoustics, Speech, and Signal Process. (ICASSP)*, Dallas, Texas USA, 2010.

[4] R. Martin, "Noise power spectral density estimation based on optimal smoothing and minimum statistics," *IEEE Trans. Speech Audio Process.*, vol. 9, no. 5, pp. 504–512, 2001.

[5] J. M. Kates, *Digital Hearing Aids*, Plural Publishing, Inc, 1 edition, 2008.

[6] B. King and L. Atlas, "Coherent modulation comb filtering for enhancing speech in wind noise," in *Proc. of Intern. Workshop on Acoustic Echo and Noise Control (IWAENC)*, Seattle, Washington USA, 2008.

[7] E. Nemer and W. Leblanc, "Single-microphone wind noise reduction by adaptive postfiltering," in *Proc. of IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, New Paltz, New York USA, 2009.

[8] P. Jax and P. Vary, "Bandwidth extension of speech signals: A catalyst for the introduction of wideband speech coding?," *IEEE Communications Magazine*, vol. 44, no. 5, pp. 106–111, 2006.

[9] W. Hess, *Pitch Determinaton of Speech Signals*, Springer Verlag, 1983.

[10] H. Kobayashi and T. Shimamura, "A weighted autocorrelation method for pitch extraction of noisy speech," in *Proc. of IEEE Intern. Conf. on Acoustics, Speech, and Signal Process. (ICASSP)*, Istanbul, Turkey, 2000.

[11] S. Gonzalez and M. Brookes, "A pitch estimation filter robust to high levels of noise (PEFAC)," in *Proc. of European Signal Processing Conf. (EUSIPCO)*, Barcelona, Spain, 2011.

[12] A. Noll, "Cepstrum pitch determination," *The Journal of the Acoustical Society of America (JASA)*, vol. 41, no. 2, pp. 293–309, 1967.

[13] A. Noll, "Pitch determination of human speech by the harmonic product spectrum, the harmonic sum spectrum and a maximum likelihood estimate," in *Proc. of the Symposium on Computer Processing in Communications*, 1970, vol. 14, pp. 779–797.

[14] M. Ross, H. Shaffer, A. Cohen, R. Freudberg, and H. Manley, "Average magnitude difference function pitch extractor," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 22, no. 5, pp. 353 – 362, oct 1974.

[15] F. Plante, G. Meyer, and W.A. Ainsworth, "A pitch extraction reference database," in *European Conference on Speech Communication and Technology*. ESCA, September 1995, vol. 4.

[16] ANSI S3.5-1997, "Methods for the calculation of the speech intelligibility index," 1997.

[17] A.W. Rix, J.G. Beerends, M.P. Hollier, and A.P. Hekstra, "Perceptual evaluation of speech quality (pesq)-a new method for speech quality assessment of telephone networks and codecs," in *Proc. of IEEE Intern. Conf. on Acoustics, Speech, and Signal Process. (ICASSP)*, 2001.