SINGLE MICROPHONE WIND NOISE PSD ESTIMATION USING SIGNAL CENTROIDS

Christoph Matthias Nelke[1], *Navin Chatlani*[2], *Christophe Beaugeant*[3] *and Peter Vary*[1]

[1] Institute of Communication Systems and Data Processing (ind)
RWTH Aachen University, Germany
[2] Intel, Allentown, PA, USA [3] Intel, Sophia-Antipolis, France
{nelke,vary}@ind.rwth-aachen.de {navin.chatlani,christophe.beaugeant}@intel.com

ABSTRACT

This contribution presents an efficient technique for the enhancement of speech signals disturbed by wind noise. In almost all noise reduction systems an estimate of the current noise power spectral density (PSD) is required. As common methods for background noise estimation fail due to the non-stationary characteristics of wind noise signals, special algorithms are required. The proposed estimation technique consists of three steps: a feature extraction followed by a wind noise detection and the calculation of the current wind noise PSD. For all steps we exploit the different spectral energy distributions of speech and wind noise. In this context, the so-called signal centroids are introduced. Investigations with measured audio data show that our method can cope with the non-stationary characteristics and enables a sufficient reduction of wind noise. In contrast to other wind noise reduction schemes the proposed algorithm has low complexity and low memory consumption.

Index Terms— Wind noise reduction, speech enhancement, wind detection, noise PSD estimation, single microphone

1. INTRODUCTION

Mobile communication devices are quite often used in extreme acoustical environments. An annoying factor is the occurrence of noise which is picked up by the microphone during a conversation. Wind noise represents a special class of interference because it is generated by turbulences in an air stream around the edges of the device leading to a fast changing, non-stationary noise signal. In the case of a speech signal superposed by wind noise the quality and intelligibility can be greatly degraded. Most mobile devices do not offer space for a wind screen, therefore it is necessary to develop systems which can reduce the effects of wind noise by means of signal processing. The crucial part of all of these systems is the accurate estimation of the noise PSD. In the past decades many single microphone methods where proposed to estimate the noise PSD from noisy speech signals (e.g. [1], [2], [3]). All these algorithms rely on the assumption that the noise signal is slower varying over time than the speech signal. This is however not true for wind noise signals. The conventional algorithms provide no or only little noise reduction due to inaccurate noise PSD estimates. Some prior works exhibit methods which were especially designed for wind noise reduction. Efficient algorithms were proposed exploiting the coherence of dual microphone recording ([4], [5]). However, many mobile applications are equipped with a single microphone. Approaches dealing with the reduction of wind noise in single microphone signals can be found in [6], [7], [8], [9], [10]. While [6], [7] and [10] directly modify the noisy input signal, the methods of [8] and [9] provide an estimate of the wind noise PSD.

The approach proposed in this contribution estimates the wind noise PSD by exploiting the spectral characteristics of speech and noise. The $1/f$-decay of the magnitude spectrum towards higher frequencies of the wind noise and the harmonic structure of voiced speech signals are used for differentiation. In Sec. 2 the proposed system is presented. Then the general signal statistics and the proposed noise PSD estimation technique are introduced in Sec. 3 and 4. An evaluation and conclusions are given in Sec. 5 and 6.

2. WIND NOISE REDUCTION SYSTEM

It is assumed that the noisy signal $x(k)$ is the superposition of the clean speech signal $s(k)$ and the wind noise signal $n(k)$. The wind noise reduction system is realized as a short-time frequency domain overlap-add structure as depicted in Fig. 1. The noisy input signal $x(k)$ is first segmented into frames of 20 ms with 50% overlap on which a Hann window is applied. The frames are transformed into the frequency domain via a fast Fourier transform (FFT) yielding $X(\lambda, \mu)$ where λ and μ are the discrete frame index and frequency bin. The enhanced signal $\hat{S}(\lambda, \mu)$ is obtained by multiplying $X(\lambda, \mu)$ with spectral gains $G(\lambda, \mu)$. The enhanced time domain signal $s(k)$ is obtained by using the IFFT and overlap-add. The novel concept presented in this paper is the estimation of the wind noise PSD $\hat{\Phi}_n(\lambda, \mu)$ which can be separated into the three highlighted blocks.

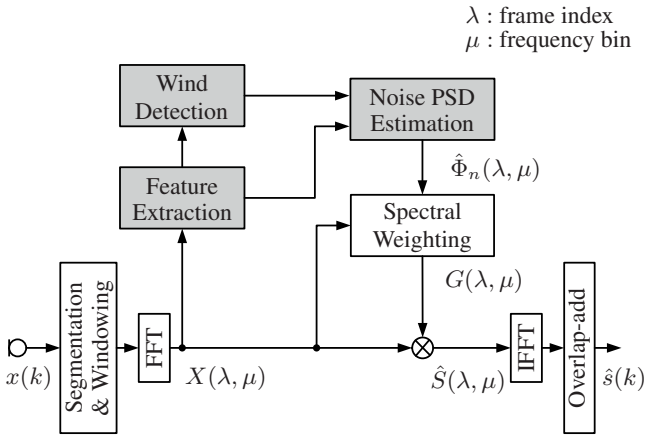

Fig. 1. Wind Noise Reduction System

978-1-4799-2893-4/14/$31.00 ©2014 IEEE

3. SIGNAL STATISTICS

As initially mentioned wind noise is mainly generated by a turbulent air flow around obstacles. The effects caused by the direct interaction of the wind with the microphone has only a small influence on the acoustic signal [11] but the turbulences in the wind flow induce transient acoustic signals. The duration of one wind gust varies from 100 ms up to several seconds. In order to detect wind noise segments a feature is required which is only dependent on the short-term statistics. A particular characteristic of wind noise is the spectral energy distribution. The spectrum has a constant level for low frequencies (< 10 Hz) and a $1/f$-behavior for higher frequencies [11]. The spectrum of speech signals differ greatly from this low pass characteristic. Segments of a speech signal can roughly be divided in two classes: voiced and unvoiced. While voiced segments have a harmonic structure unvoiced segments are noise-like. The spectral energy distributions of wind, voiced and unvoiced speech are shown in Fig. 2. The curves show the averaged spectra of speech segments from [12] and wind noise signals. More details on the wind noise recordings are given in Sec. 5.1. The wind noise is depicted by the red line and exhibits clearly visible the low pass characteristic. The main energy of voiced speech (black line) is located between 100 and 1000 Hz whereas unvoiced speech (gray line) is distributed in the frequency range above 3000 Hz. From the depicted energy distributions it can be seen that wind noise mainly overlaps the frequency range of voiced speech segments. Consequently, the proposed noise estimation and reduction is only realized in this frequency range below unvoiced speech (0-3000 Hz).

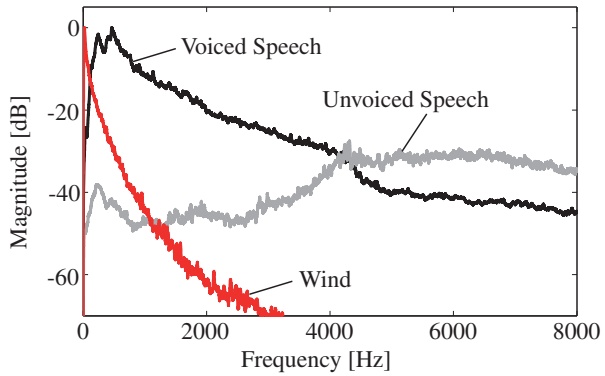

Fig. 2. Spectral Energy Distribution of Speech and Wind

4. WIND NOISE ESTIMATION

As described in Sec. 2, the wind noise PSD estimation is divided into three parts which will be explained more detailed in the following.

4.1. Feature Extraction

For the detection and estimation of wind noise so-called spectral subband centroids (SSCs) are introduced as a feature. The SSCs determine the spectral "center-of-gravity" of a signal in a certain subband and were formerly used to support automatic speech recognition systems [13], [14]. For the PSD $\Phi_x(\lambda,\mu)$ of a signal frame the SSC of the m-th subband is defined by

$$\mathrm{SSC}_m(\lambda) = \frac{f_s}{L}\frac{\sum_{\mu=\mu_{m-1}}^{\mu_m-1}\mu\cdot\Phi_x(\lambda,\mu)}{\sum_{\mu=\mu_{m-1}}^{\mu_m-1}\Phi_x(\lambda,\mu)}\,. \tag{1}$$

Here, the PSD is estimated by the squared magnitude of the noisy input $X(\lambda,\mu)$. The subbands are limited by the frequency bins μ_m, f_s is the sampling frequency (16 kHz) and $L =$ the FFT size (512). Because the SSCs refer to a position in the spectrum they will be treated as a frequency in the following. For our algorithm we only consider a single SSC representing the frequency range from 0 to 3000 Hz ($\mu_0 = 1 \ldots \mu_1 = 96$) which will be denoted as SSC_1. The observations made in Sec. 3 lead to the fact that SSC_1 is mainly affected by voiced speech segments and wind noise segments, whereas unvoiced speech segments have only marginal influence on the first centroid. For an ideal $1/f$-decay of a wind noise signal, the SSC_1 value is constant and independent of the absolute signal energy due to the normalization. In Fig. 3, the histograms of SSC_1 values from 20 ms segments of wind noise and voiced speech segments taken from [12] are shown. Here, about 25 minutes of speech samples and 6 minutes of wind noise were used for this measurement. The low pass characteristic of wind noise results in a narrowband distribution of the SSCs clearly below 100 Hz while the SSCs of voiced speech segments are distributed mostly between 300 Hz and 500 Hz.

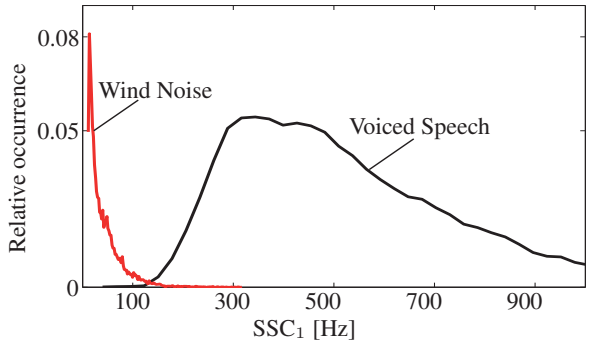

Fig. 3. Distribution of SSC_1

4.2. Wind Noise Detection

In the proposed algorithm it is necessary to reliably detect signal segments containing wind noise. This detection is based on the SSC_1 calculated in every signal frame. Based on the assumption that the speech and noise signals are uncorrelated, the PSD of the noisy signal is given by the sum of the speech PSD $\Phi_s(\lambda,\mu)$ and the noise PSD $\Phi_n(\lambda,\mu)$. With the definition of the *a posteriori* SNR

$$\gamma(\lambda,\mu) = \frac{\Phi_x(\lambda,\mu)}{\Phi_n(\lambda,\mu)} \tag{2}$$

Eq. 1 can be rewritten as

$$\mathrm{SSC}_1(\lambda) = \mathrm{SSC}_{1,\mathbf{S}}(\lambda)\cdot\left(1-\frac{1}{\overline{\gamma}(\lambda)}\right) + \mathrm{SSC}_{1,\mathbf{N}}(\lambda)\cdot\left(\frac{1}{\overline{\gamma}(\lambda)}\right) \tag{3}$$

with $\mathrm{SSC}_{1,\mathbf{S}}(\lambda)$ and $\mathrm{SSC}_{1,\mathbf{N}}(\lambda)$ representing the centroid frequencies of clean speech and pure wind and $\overline{\gamma}(\lambda)$ is the *a posteriori* SNR averaged over the frequency bins used for the SSC_1 computation in one frame. From Eq. 3 it can be seen that SSC_1 can be used as an indicator for clean voiced speech, pure wind noise, or a soft decision on a mixture of the two previous cases. While applying a threshold to SSC_1 can identify segments containing clean voiced speech or pure wind (e.g. 150 Hz in Fig. 3) a soft decision which determines the degree of disturbance can be based on SSC_1. Fig. 4 depicts the influence of the superposition of voiced speech and wind noise on SSC_1, where the same signal samples were used as for the experiment shown in Fig. 3. Here the SSC_1 of frames of voiced speech

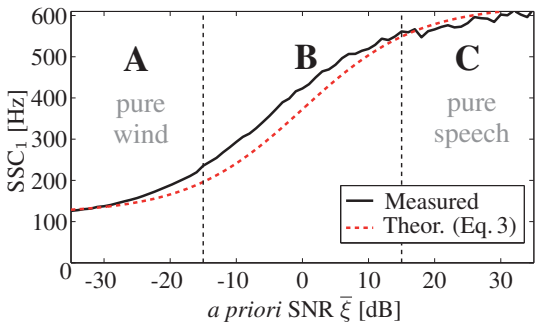

Fig. 4. SSC_1 of voiced speech disturbed by wind noise

segments disturbed by wind noise are computed and sorted according to the current signal-to-noise ratio (SNR) in this frame. The black solid curve in Fig. 4 shows the mean SSC_1 values as a function of the different SNR values. The theoretical relationship derived in Eq. 3 is given by the red dashed line. Here the dependency on the averaged *a priori* SNR $\overline{\xi} = \Phi_s/\Phi_n$ is given for reasons of clarity. It can clearly be seen that the measured centroid frequencies from the signal frames follow the theoretical curve. Based on this relation between the centroid frequency and the SNR three labeled ranges (**A**, **B** and **C**) are defined. In ranges **A** and **C**, pure wind and clean voiced speech are predominant, respectively. In range **B** both voiced speech and wind are active and the SSC_1 gives information on the degree of distortion. We denote the frequencies corresponding to the limits between **A**-**B** and **B**-**C** by f_1 and f_2. For SSC_1 values below f_1 pure wind and SSC_1 values above f_2 clean voiced speech is assumed. The energy of unvoiced speech segments is rather located at higher frequencies and show a spectrally flat characteristic in the frequency range of the SSC_1 (see Fig. 2). This leads just to marginal effects on the SSC_1 which can only be seen in periods without wind noise in terms of higher SSC_1 values (> 1000 Hz). Therefore the influence of unvoiced speech on the wind noise detection can be neglected.

4.3. Noise PSD Estimation

In general, the estimation of the PSD $\hat{\Phi}_n^2(\lambda,\mu)$ of a time varying signal is often realized via recursive smoothing of the noise component $N(\lambda,\mu)$ in consecutive signal frames as

$$\hat{\Phi}_n(\lambda,\mu) = \alpha(\lambda)\cdot\hat{\Phi}_n(\lambda-1,\mu) + (1-\alpha(\lambda))\cdot|N(\lambda,\mu)|^2, \quad (4)$$

where the smoothing factor $\alpha(\lambda)$ can take values between 0 and 1 and can be chosen fixed or adaptive. $|N(\lambda,\mu)|^2$ is called a noise periodogram and is not directly accessible since the input signal contains both speech and wind noise. Hence, for the proposed system the noise periodograms are estimated based on the classification defined in Sec. 4.2. Within the range **A**, where only wind noise occurs the input signal can directly be used for the calculation of the noise periodogram. Within range **C**, where we assume clean speech the noise periodogram is set to zero. For the estimation within the third range **B**, where both voiced speech and wind noise are active, a more sophisticated approach is used which exploits the spectral characteristics of wind noise and voiced speech. Due to the $1/f$-decay of the spectral power of wind signal, the noise periodograms are approximated with a simple exponential fit as

$$|\hat{N}_{exp}(\lambda,\mu)|^2 = \frac{\beta}{\mu^\nu}. \quad (5)$$

The parameters β and ν are introduced to adjust the power and the decay of the periodogram. For the computation of β and ν, two supporting points are required corresponding to the spectrum of the wind noise. Our technique exploits the harmonic structure of voiced speech where the spectrum exhibits local maxima at the pitch frequency and multiples of this frequency. The pitch frequency is dependent on the articulation and varies for different speakers. Between the multiples of the pitch frequency the spectrum reveals local minima where no or only very low speech energy is located and thus the spectrum of the wind noise is exposed. The approximation in Eq. 5 is now fit to these local minima which can be assigned to the wind noise spectrum. Typical values for the decay parameter ν are between 0.5 and 2. In Fig. 5 the dashed gray line depicts the noisy speech spectrum, the red line the wind noise spectrum and the green dashed line the approximation from Eq. 5 using the points marked by the black circles for the computation of β and ν.

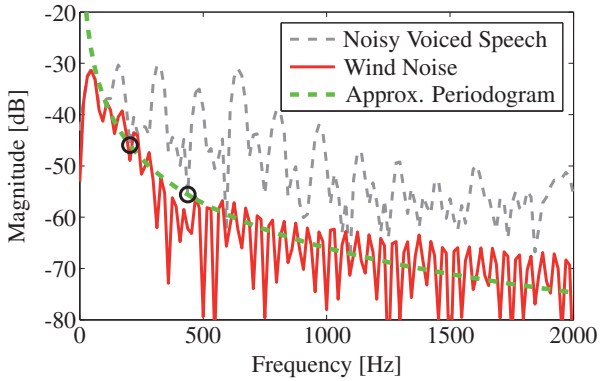

Fig. 5. Approximation of wind noise periodogram

Overestimation of the wind noise is avoided by limiting the estimate $|\hat{N}_{exp}(\lambda,\mu)|^2$ to the current noisy input frame. Finally, the calculation of the wind noise periodogram based on the current $SSC_w(\lambda)$ value can be summarized as:

$$|\hat{N}(\lambda,\mu)|^2 = \begin{cases} |\hat{X}(\lambda,\mu)|^2 & \text{, if } SSC_1(\lambda) < f_1 \\ |\hat{N}_{exp}(\lambda,\mu)|^2 & \text{, if } f_1 < SSC_1(\lambda) < f_2 \\ 0 & \text{, if } SSC_1(\lambda) > f_2 \end{cases} \quad (6)$$

The performance of the noise estimation is heavily dependent on the smoothing factor $\alpha(\lambda)$ in Eq. 4. On the one hand, a small smoothing factor allows fast tracking of the wind noise. This has the drawback that speech segments which are wrongly detected as wind noise have a great influence on the estimated noise PSD. On the other hand, a large smoothing factor reduces the effect of wrong detection during speech activity. However, this leads to slow adaptation of the noise estimate. Thus, an adaptive computation of $\alpha(\lambda)$ is favorable where low values are chosen during wind activity in speech pauses and high values during speech activity. Since the SSC_1 value is an indicator for the current SNR condition (see Fig. 4) the following linear mapping for the smoothing factor is used:

$$\alpha(\lambda) = \begin{cases} \alpha_{\min} & \text{, if } SSC_0(\lambda) < f_1 \\ \frac{1}{f_2-f_1}[\alpha_{\max}(SSC_0(\lambda)-f_1) & \\ \quad + \alpha_{\min}(f_2 - SSC_0(\lambda))] & \text{, if } f_1 < SSC_0(\lambda) < f_2 \\ \alpha_{\max} & \text{, if } SSC_0(\lambda) > f_2 \end{cases} \quad (7)$$

This relationship between the smoothing factor $\alpha(\lambda)$ and the $SSC_0(\lambda)$ value leads to a fast tracking and consequently accurate noise estimate in speech pauses ($SSC_0(\lambda) \approx f_1$) and reduces

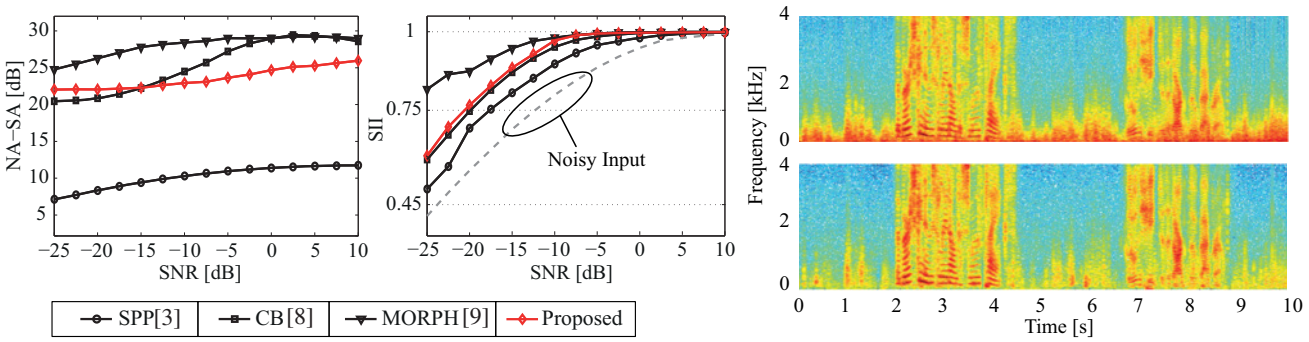

Fig. 6. Evaluation results: *left:* noise reduction performance (NA-SA: noise attenuation minus speech attenuation); *middle:* intelligibility enhancement (SII: speech intelligibility index); *right:* spectrograms of noisy speech signal at SNR = -5 dB (top) and enhanced output of the proposed system (bottom)

the risk of wrongly estimating speech as wind noise during speech activity ($\mathrm{SSC}_0(\lambda) \approx f_2$). The frequencies f_1 and f_2 are the values of SSC_1 for pure wind and pure speech, respectively.

5. EXPERIMENTS AND RESULTS

5.1. Experimental Setup

For this evaluation, wind noise was recorded with a mock-up mobile phone mounted in hand-held position on an artificial head (HEAD acoustics HMS II.3 & HHP III) on a windy day with wind speeds up to 15 m/s. In order to have a reference for the evaluation the noisy speech was generated by a superposition of clean speech from [12] with the wind noise recordings at different SNR values. The thresholds f_1 and f_2 for the classification were set to 200 and 600 Hz, respectively. The limits were chosen somewhat higher than shown in Fig. 4 and ensure less misclassification of speech as wind noise and thus leads to lower speech distortion. The range of the adaptive smoothing factor was set to $\alpha_{\min} = 0.1$ and $\alpha_{\max} = 0.9$. The required PSD of the input signal $\Phi_x(\lambda, \mu)$ was calculated by recursive smoothing as defined in Eq. 4 with a fixed smoothing factor of 0.5. For the computation of β and ν the first two local minima greater than 50 Hz are used. This makes the proposed system robust towards inaccuracy of the measurement hardware such as a highpass characteristic of the used microphone. Besides, the decay parameter ν is limited by the aforementioned range between 0.5 and 2.

5.2. Evaluation Results

The performance of the proposed approach is compared with a state-of-the-art single microphone algorithm based on the speech presence probability [3] (SPP) which was designed for general noise PSD estimation. In addition two methods explicitly designed for the estimation of the wind noise PSD were considered in this comparison: The authors of [8] propose to estimate the wind noise spectrum by using templates stored in a codebook (CB). In [9] morphological operations known from image processing are exploited and leads to estimate the wind noise PSD (MORPH). Therefore the spectrogram of a noisy signal is considered as an image in the time-frequency plane where connected regions are determined as signal parts degraded by wind noise. In all investigated algorithms the noise reduction was realized by applying spectral subtraction gain calculation [15]. The noise reduction performance is determined by means of the noise attenuation minus speech attenuation (NA-SA) measure (e.g. [16]), where higher values indicate an improvement and the results are shown in the left plot in Fig. 6. Besides, the speech intelligibility index (SII) [17] was calculated for the noisy as well as the enhanced signal. A SII higher than 0.75 indicates a good communication system and values below 0.45 correspond to a poor system. The results are shown in the middle plot of Fig. 6 where the dashed line corresponds to the SII of the noisy unprocessed input signal. Both measures indicate that the methods designed for wind noise reduction outperform the SPP noise estimator. This is caused by the non-stationary characteristics of wind noise which is not assumed for general background noise signals. The algorithms designed for wind noise reduction all provide similar results where the morphological approach shows the highest improvement especially for low SNR conditions (< 5 dB). Informal listening tests confirm this results, though the morphological approach tends to provide a more aggressive noise reduction and induce some speech distortions. The right plot in Fig. 6 exemplifies the performance of the proposed method in the shown spectrograms of a noisy signal (SNR = -5 dB) and the enhanced output signal. It can clearly be seen that the wind noise is reduced by a great amount in speech pauses as well as during speech activity. For reasons of clarity, only the 0-4 kHz range is depicted.

5.3. Complexity

As initially mentioned one advantage of the proposed system is the low computational complexity. The morphological operations in [9] are used to find connected regions in the time-frequency plane. Based on their shapes, these regions are then further classified as voiced speech or wind noise Both the morphological operations and the classification has a high computational effort. In the method from [8] the codebook search is rather complex and the algorithm requires memory for the storage of the codebook entries. In contrast, the operations of our approach are less complex and have only little memory consumption. In a MATLAB implementation the computation time of our method is about 5 times lower than for the approaches of [8] and [9].

6. CONCLUSIONS

In this contribution a system was proposed which exploits the short-term spectral energy distributions to detect and reduce wind noise in noisy speech signal. Based on the signal centroids a procedure to estimate the wind noise PSD in a single microphone signal is presented. An evaluation with wind noise recordings shows that the proposed method outperforms the state-of-the-art SPP approach [3] for background noise estimation and leads to similar results as other approaches especially designed for wind noise reduction with a significantly lower computational complexity.

7. REFERENCES

[1] 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.

[2] R. 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.

[3] T. Gerkmann and R. Hendriks, "Noise power estimation based on the probability of speech presence," in *Proc. of IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, New Paltz, NY, USA, 2011.

[4] G. Elko, "Reducing noise in audio systems," US Patent US7 171 008, 2007.

[5] S. Franz and J. Bitzer, "Multi-channel algorithms for wind noise reduction and signal compensation in binaural hearing aids," in *Proc. of Intern. Workshop on Acoustic Echo and Noise Control (IWAENC)*, Tel Aviv, Israel, 2010.

[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] S. Kuroiwa, Y. Mori, S. Tsuge, M. Takashina, and F. Ren, "Wind noise reduction method for speech recording using multiple noise templates and observed spectrum fine structure," in *Intern. Conf. on Communication Technology*, 2006.

[9] C. Hofmann, T. Wolff, M. Buck, T. Haulick, and W. Kellermann, "A morphological approach to single-channel wind-noise suppression," in *Proc. of Intern. Workshop on Acoustic Echo and Noise Control (IWAENC)*, Aachen, Germany, Sept. 2012.

[10] C. Nelke, N. Nawroth, M. Jeub, C. Beaugeant, and P. Vary, "Single microphone wind noise reduction using techniques of artificial bandwidth extension," in *Proc. of European Signal Processing Conf. (EUSIPCO)*, Bucharest, Romania, August 2012.

[11] S. Bradley, T. Wu, S. Hünerbein, and J. Backman, "The mechanisms creating wind noise in microphones," in *Audio Engineering Society, 114th Convention*, 2003.

[12] P. Kabal, "TSP speech database," Dep. of Electrical & Computer Engineering, McGill University, Montreal, Canada, Tech. Rep., 2002.

[13] K. Paliwal, "Spectral subband centroid features for speech recognition," in *Proc. of IEEE Intern. Conf. on Acoustics, Speech, and Signal Process. (ICASSP)*, Seattle, USA, 1998.

[14] J. Chen, Y. Huang, Q. Li, and K. Paliwal, "Recognition of noisy speech using dynamic spectral subband centroids," *Signal Processing Letters, IEEE*, vol. 11, no. 2, pp. 258 – 261, Feb. 2004.

[15] S. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 27, no. 2, pp. 113 – 120, 1979.

[16] S. Gustafsson, R. Martin, P. Jax, and P. Vary, "A psychoacoustic approach to combined acoustic echo cancellation and noise reduction," vol. 10, no. 5, pp. 245–256, Jul. 2002.

[17] ANSI S3.5-1997, *Methods for the Calculation of the Speech Intelligibility Index*, American National Standards Inst. Std., 1997.