Wind Noise Detection: Signal Processing Concepts for Speech Communication

Christoph Nelke, Peter Jax, Peter Vary
Institut für Kommunikationssysteme, Muffeter Weg 3a, 52074 Aachen, Deutschland
Email: {nelke, jax, vary}@iks.rwth-aachen.de

Abstract

The acoustic signal induced by wind during speech recordings can be a severe problem, e.g., for mobile phones, video recordings or hearing aids. Due to the dimension and design constraints, many devices for these applications do not offer space for the use of mechanical windscreens. Therefore, it is necessary to combat the acoustic noise in the captured signal by digital signal processing techniques. The first step towards the reduction of an undesired noise component in a speech signal is a detection of segments with wind noise activity. The detector must be capable of adapting quickly to the non-stationary signal characteristics of wind noise. In this paper three new wind detection concepts are presented. The first algorithm is based on the short-term mean introduced by wind noise in a recorded signal while the characteristic spectral shape and energy distribution of wind noise are exploited in the second and third approach. All three proposed methods are compared with known approaches from literature in terms of their accuracy using real wind noise recordings. All considered algorithms are implemented as real-time processing schemes working on the short-term spectrum of signal frames of 20 ms. This is realized by an overlap-add structure, which is widely used for digital speech processing procedures.

Introduction

Nowadays, mobile communication takes place in nearly every environment. However, the advantage of a higher mobility leads to more challenging acoustic scenarios. Additive noise signals may impair the quality and intelligibility of the recorded speech signal. A particular noise is generated outdoors when the microphone is exposed to wind resulting in annoying low-frequency rumbling sounds. Since the size of devices such as mobile phones, camcorders, or hearing aids does not allow the usage of windscreens, the recorded signals can suffer from great noise magnitudes. For reducing the wind noise influence, digital signal processing approaches can be very helpful (see, e.g., [1], [2]). A sufficiently precise detection of wind noise is the first step towards a suppression of noise in the captured signals. Furthermore, a specific detection method for wind noise is very helpful for outdoor recordings and videos, where otherwise a degradation of the recorded signal by wind might not be noticed during the recording process. Several algorithms for wind noise detection can be found in the field of signal processing for hearing aids. In this paper new approaches are presented which exploit special characteristics of speech and wind noise. Based on the found features accurate detection methods are realized.

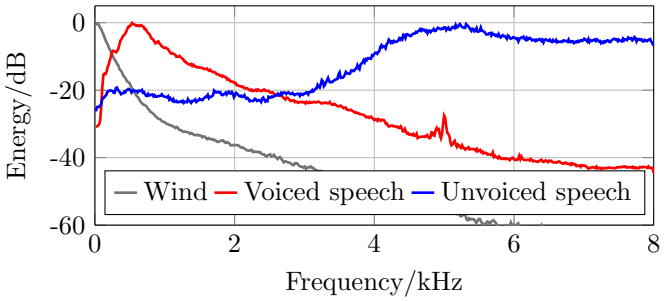

Figure 1: Energy distribution of speech and wind.

Signal Statistics

Wind noise is mainly generated by a turbulent air flow around obstacles which induces transient acoustic signals. Many conventional noise reduction algorithms exploit the temporal statistics of speech and stationary background noise but fail for instationary wind noise. In order to detect wind noise segments, features are required which 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 very low frequencies (< 10 Hz) and a $1/f$-behavior for higher frequencies [3]. The spectrum of speech signals differ greatly from this low pass characteristic. Segments of speech signals 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. 1. The curves show the averaged spectra of speech segments from [4] and wind noise signals from [5]. Wind noise exhibits a clearly visible low pass characteristic. Most of the energy of voiced speech is located around 1000 Hz whereas unvoiced speech is distributed in the frequency range above 3000 Hz.

System Overview

The structure of a typical speech enhancement system is depicted in Fig. 2. The noisy input signal is given in a digital representation $x(k)$ with the discrete time index k as superposition of the speech signal $s(k)$ and the wind noise signal $n(k)$. The aim of the speech enhancement part is to estimate the clean speech signal. This is commonly realized by an adaptive filtering either of short signal frames $x_\lambda(\kappa)$ with frame index λ and κ determines the sample position within the frame. Or the discrete Fourier transform (DFT) $X(\lambda, \mu)$ of the frame is used, where μ is the discrete frequency bin using a DFT

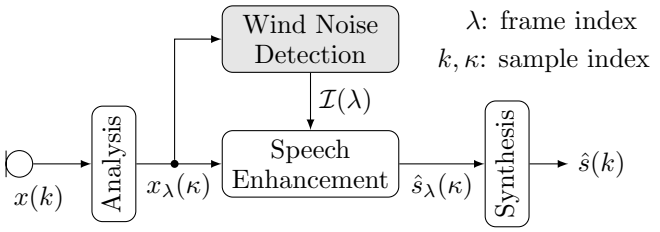

Figure 2: Speech processing system.

length of $N = 512$. Therefore the analysis stage segments the input signal into frames of $L = 320$ samples at a sampling frequency of 16 kHz. The estimate $\hat{s}_\lambda(\kappa)$ of the speech signal is then combined by the synthesis stage into the enhanced output signal $\hat{s}(k)$. Further literature on the speech enhancement part can be found in general in [6] or for wind noise reduction, e.g., in [2] and references therein. In this contribution a frame-wise detection of wind noise is considered which is crucial for many wind noise reduction approaches. The resulting wind indicator $\mathcal{I}(\lambda) \in [0, 1]$ represents a soft decision whether wind is present or not in the current frame.

Wind Detection

In this section, three new algorithms (STM,SSC, and TSC) for wind noise detection are proposed based on a frame-wise processing. In addition, two reference approaches from literature are introduced.

STM: Short-Term Mean

The low frequency characteristic of wind noise can be investigated by the short-term mean (STM) of the signal. Usually, the digital representation of an acoustic signal can be assumed to be zero-mean (see, e.g., [6]). However, the zero-mean property is only valid in a long-term sense, while shorter signal segments can show a direct component (DC) depending on their frequency components. The DC or mean value of short segments can be used to detect low frequency parts in a signal and is here defined in a normalized way as

$$\mathcal{I}_{\mathrm{STM}}(\lambda) = \left| \frac{\sum_{\kappa=0}^{L-1} x_\lambda(\kappa)}{\sum_{\kappa=0}^{L-1} |x_\lambda(\kappa)|} \right|. \tag{1}$$

The normalization with the sum of the absolute values of the frame $x_\lambda(\kappa)$ leads to values close to 0 for high frequency components. For DC dominated signals, such as wind noise, the two sums in Eq. 1 will be identical and thus the STM will be 1.

SSC: Signal Sub-band Centroids

In [2] a method for wind noise estimation is proposed that investigates the energy distribution of a given spectrum. In this context the so-called signal sub-band centroids SSC were introduced. They depict the center-of-gravity in a given frequency range from f_1 to f_2 and are defined for a spectrum $X(f)$ by

$$\Xi_{f_1,f_2} = \frac{\int_{f_1}^{f_2} |X(f)|^2 \cdot f df}{\int_{f_1}^{f_2} |X(f)|^2 df} \tag{2}$$

For a theoretical investigation of the SSC, this continuous frequency-domain representation is considered. As described previously, it is assumed that the wind noise magnitude spectrum can be approximated by an $1/f$ shape, which yields in the wind noise power spectrum approximation

$$|X(f)|^2 \approx \frac{\beta}{f^2}. \tag{3}$$

The parameter β scales the total signal energy of the wind noise. Inserting Eq. 3 in Eq. 2, β cancels out and the integrals can be solved, giving the following expression

$$\Xi_{f_1,f_2,\mathrm{wind}} = f_1 \cdot f_2 \cdot \left(\frac{\ln(f_2) - \ln(f_1)}{f_2 - f_1} \right) \tag{4}$$

as a function of the frequency limits f_1 and f_2. An interesting feature is that $\Xi_{f_1,f_2,\mathrm{wind}}$ tends towards zero, if $f_1 \to 0$. For the implementation in a digital signal processing system, the discrete frequency-domain representation from Eq. 5 is used beginning at low frequencies ($\mu_1 = 0$) up to the discrete frequency bin μ_2 corresponding to f_2.

$$\Xi_{\mu_1,\mu_2}(\lambda) = \frac{\sum_{\mu=\mu_1}^{\mu_2} |X(\lambda,\mu)|^2 \cdot \mu}{\sum_{\mu=\mu_1}^{\mu_2} |X(\lambda,\mu)|^2}, \tag{5}$$

Again, a wind indicator is desired, which takes only values in the range between 0 and 1. Setting $\mu_1 = 0$ leads to SSC values close to zero for wind noise, whereas speech will generate higher values with a theoretical maximum of μ_2. The SSC-based wind indicator is finally defined as

$$\mathcal{I}_{\mathrm{SSC}}(\lambda) = \frac{\mu_2 - \Xi_{\mu_1,\mu_2}(\lambda)}{\mu_2} \quad \in [0, 1]. \tag{6}$$

TSC: Template Spectrum Combination

A different approach for the detection of wind noise is derived from a concept using pre-trained codebooks containing speech and noise templates (see, e.g., [7]). The basic idea is that the noisy spectral magnitude $|X(\lambda,\mu)|$ can be decomposed into the speech template $|\widetilde{S}_i(\mu)|$ with the index i and a noise template $|\widetilde{N}_j(\mu)|$ with index j. Then, the template spectrum combination (TSC) of the noisy magnitude spectrum is approximated by

$$|\widehat{X}(\lambda,\mu)| = \sigma(\lambda) \cdot |\widetilde{S}_i(\mu)| + (1 - \sigma(\lambda)) \cdot |\widetilde{N}_j(\mu)|. \tag{7}$$

Because all signals in Eq. 7 tagged with the $\widetilde{}$-operator are normalized to a frame-energy of 1, the codebook

weight $\sigma(\lambda)$ takes values between 0 and 1. An extensive search is applied using all combination of codebook entries $\widetilde{S}_i(\mu)$ and $\widetilde{N}_j(\mu)$ and discrete values for the codebook weight σ for an estimation of the noise spectrum in [7]. Here, a simplified procedure is applied to detect wind noise by using only a single representative for the speech and wind noise component. For the speech component $\widetilde{S}(\mu)$ the standardized long-term average speech spectrum (LTASS) as defined in [8] is used, while the $1/f$-approximation from Eq. 3 represents the wind noise component $\widetilde{N}(\mu)$. The vector notation describes the spectral magnitudes of the DFT coefficients in each frame λ

$$\mathbf{X}(\lambda) = [|X(\lambda,0)|,\ldots,|X(\lambda,\mu)|,\ldots,|X(\lambda,N/2)|]^T. \quad (8)$$

By minimizing the mean square error between a given input signal $\mathbf{X}(\lambda)$ and the estimate $\widehat{\mathbf{X}}(\lambda)$ defined in Eq. 7

$$\begin{aligned} e(\lambda) &= ||\mathbf{X}(\lambda) - \widehat{\mathbf{X}}(\lambda)||^2 \\ &= ||\mathbf{X}(\lambda) - \sigma(\lambda)\cdot\widetilde{\mathbf{S}}(\lambda) - (1-\sigma(\lambda))\cdot\widetilde{\mathbf{N}}(\lambda)||^2 \stackrel{!}{=} \min \end{aligned} \quad (9)$$

an optimal template weight $\sigma_{\mathrm{opt}}(\lambda)$ can be derived by taking the derivative with respect to $\sigma(\lambda)$ and setting the result to zero yielding

$$\sigma_{\mathrm{opt}}(\lambda) = \frac{||\widetilde{\mathbf{N}}(\lambda)||^2 - \widetilde{\mathbf{S}}^T(\lambda)\widetilde{\mathbf{N}}(\lambda) + \mathbf{X}^T(\lambda)\cdot(\widetilde{\mathbf{S}}(\lambda) - \widetilde{\mathbf{N}}(\lambda))}{||\widetilde{\mathbf{S}}(\lambda) - \widetilde{\mathbf{N}}(\lambda)||^2}, \quad (10)$$

Since all quantities in Eq. 7 are normalized to a frame-energy of 1, the template gain $\sigma_{\mathrm{opt}}(\lambda)$ indicates the amount of the speech component and $1 - \sigma_{\mathrm{opt}}(\lambda)$ the amount of the wind noise component. Thus, the template weight can be used as wind detector according to

$$\mathcal{I}_{\mathrm{TSC}}(\lambda) = 1 - \sigma_{\mathrm{opt}}(\lambda). \quad (11)$$

Reference 1: Zero Crossing Rate

The zero crossing rate (ZCR) is defined as the number of sign-changes of a given signal within a fix duration, i.e., the rate at which the signal changes from positive to negative amplitudes or back and is defined as

$$\mathrm{ZCR}(\lambda) = \frac{1}{L-1}\sum_{\kappa=1}^{L-1} \mathbf{I}\{x_\lambda(\kappa)\cdot x_\lambda(\kappa-1) < 0\} \ \in [0,1] \quad (12)$$

where the indicator function $\mathbf{I}\{A\}$ is 1 if its argument A is true and 0 otherwise. The ZCR is dependent on the frequency components and is a well known feature in the field of voice activity detectors (VAD). Low frequency signals result in slow changes of the time signal and thus a low number of sign-changes is generated resulting in a ZCR close to zero. Higher frequencies in the considered signal will produce more sign-changes, which leads to ZCR-values closer to one. The spectral component of a signal with the highest amplitude will mainly affect the ZCR. To detect wind segments, it was proposed in [9] to measure the ZCR in each signal frame, as the their low frequency behavior will also generate a low ZCR. For a soft decision in terms of an indicator in the range between zero and one for the two conditions *wind inactive* and *wind active* the ZCR based indicator is simply defined as

$$\mathcal{I}_{ZCR}(\lambda) = 1 - \mathrm{ZCR}(\lambda). \quad (13)$$

Reference 2: Negative Slope Fit

A further detector presented in [9] is based on the idea that the magnitude of the spectrum of wind noise can be roughly approximated by a linear decay over the frequency, which can be expressed as

$$\widehat{\mathbf{X}}(\lambda) = a_1 \cdot \boldsymbol{\mu} + a_0 \quad (14)$$

with the frequency vector

$$\boldsymbol{\mu} = [0,1,\ldots,N/2]^T \quad (15)$$

The parameters a_0 and a_1 control the DC and the slope of the approximation and will be denoted by $\boldsymbol{a} = [a_0, a_1]^T$. Combining $\boldsymbol{\mu}$ with a vector $\mathbf{1} = [1,1,\ldots,1]^T$ containing N/2+1 ones as a $(N/2+1)\times 2$ matrix

$$\mathbf{M} = [\mathbf{1}, \boldsymbol{\mu}] \quad (16)$$

Eq. 14 can be written as

$$\widehat{\mathbf{X}}(\lambda) = \mathbf{M}\cdot\boldsymbol{a}. \quad (17)$$

Because for wind noise a negative slope is expected, the approach is named negative slope fit (NSF). A least square analysis can be applied to compute $\mathbf{a}$ by minimizing the squared error $e(\lambda) = ||\mathbf{X}(\lambda) - \widehat{\mathbf{X}}(\lambda)||^2$ leading to the optimal solution

$$\boldsymbol{a}_{\mathrm{opt}}(\lambda) = (\mathbf{M}^T\mathbf{M})^{-1}\cdot\mathbf{M}^T\cdot\mathbf{X}(\lambda). \quad (18)$$

According to [9], two conditions must be fulfilled to classify the current frame as wind noise. Firstly, the slope of the approximated spectrum must be negative ($a_1 < 0$) and secondly the squared error $e(\lambda)$ must be smaller than a certain threshold. Normalizing the error to the energy of the observed spectrum the two conditions can be combined to the wind indicator

$$\mathcal{I}_{\mathrm{NSF}}(\lambda) = \begin{cases} 1 - \dfrac{e(\lambda)}{||\mathbf{X}(\lambda)||^2} & \text{, for } a_1 < 0, \\ 0 & \text{, else.} \end{cases} \quad (19)$$

A closer investigation of this algorithm has shown that an increased performance can be achieved by applying the indicator only on a limited frequency range between 0 and 1000 Hz, where most wind energy is expected.

Evaluation

All described algorithms for wind noise detection are compared in the following by means of two measures. Firstly, the quality of the wind noise detection is measured by the wind detection rate

$$\mathcal{P}_{\mathrm{w}}(\zeta) = \frac{\#\{\mathcal{I}(\lambda) > \zeta\}}{\#\{\mathcal{M}_{\mathrm{w}}\}}, \quad \lambda \in \mathcal{M}_{\mathrm{w}} \quad (20)$$

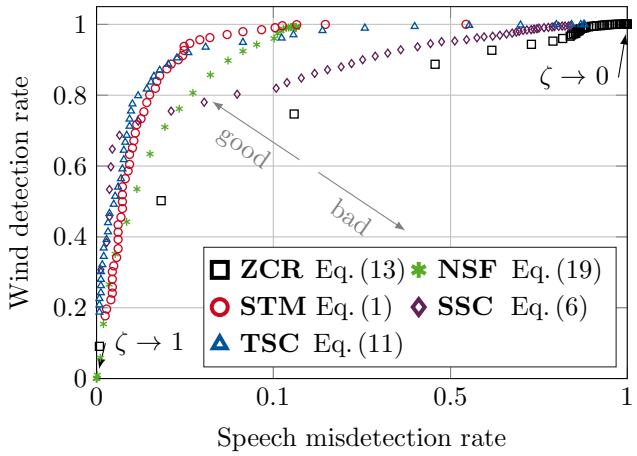

Figure 3: Evaluation results.

where $\#\{\cdot\}$ denotes the cardinality, i.e. for the numerator in Eq. 20 the number of elements in the considered set of frames in which the wind indicator $\mathcal{I}(\lambda)$ is greater than a threshold ζ. In a similar way the speech misdetection rate is defined by

$$\mathcal{P}_{\bar{\mathrm{s}}}(\zeta) = \frac{\#\{\mathcal{I}(\lambda) > \zeta\}}{\#\{\mathcal{M}_{\mathrm{s}}\}}, \quad \lambda \in \mathcal{M}_{\mathrm{s}}, \tag{21}$$

and counts the amount of clean speech, which is erroneously detected as wind noise. The sets $\mathcal{M}_{\mathrm{s}}$ of clean speech activity and $\mathcal{M}_{\mathrm{w}}$ of wind activity were labeled manually previously. An evaluation was carried out taking randomly chosen speech sentences from the TSP database [4]. The clean speech is mixed with wind noise from [5] and the corresponding noisy speech signal is segmented into frames of 20 ms. In 70% of the frames wind is active, in 50% of the frames speech is active, and speech and wind are active in about 30% of the frames. The global signal-to-noise-ratio (SNR) of the signal was -5 dB, which reflects a realistic situation.

Both measures describe important performance properties of the wind detection. On the one hand, a high detection rate of wind noise is desired for a sufficient removal of the noise in a subsequent step. But on the other hand, no clean speech segments should be detected as wind, which results in a low speech misdetection rate. Both rates defined in Eqs. 20 and 21 are dependent on a threshold ζ, which is applied to the wind indicator. Since all wind detection methods result in an indicator between 0 (no wind) and 1 (wind active), a good comparison between the algorithms is given by passing through values between 0 and 1 and measuring the resulting detection rates. Taking both the speech misdetection rate and the wind detection rate at different thresholds into account, the so-called receiver operating characteristic (ROC) can be generated as depicted in Fig. 3.

For each algorithm, a curve displays different operating points, which belong to certain values of the threshold ζ applied to the corresponding wind indicator. A good detection results in a high $\mathcal{P}_{\mathrm{w}}(\zeta)$ value and a low $\mathcal{P}_{\bar{\mathrm{s}}}(\zeta)$ value, as indicated by the arrows. The upper right corner of Fig. 3 represents thresholds close to zero, while the lower left corner depicts thresholds close to one. Because some of the above mentioned approaches only take discrete values, e.g., a discrete frequency bin or a discrete number of zero-crossings, some of the curves show partially large gaps between the working points. The ROC can be roughly separated into two parts:

- $\zeta \to 1$: All algorithms are characterized by low misdetection rates and the SSC and TSC methods show the highest detection rates.
- $\zeta \to 0$: The detection rate rises slowly, but the misdetection increases. In this range, the detector resulting from the STM and the TSC method gives the best results.

The remaining two methods, zero crossing rate (ZCR) and negative slope fit (NSF), give only inaccurate results for most of the operating points.

Summary

In this contribution three different methods for the detection of wind noise in a digital signal are presented and evaluated. The focus is set on the detection of wind noise in a speech signal which is a frequently arising task in many speech communication devices. Therefore, the detection rate is investigated along with a misdetection of clean speech frames as wind. All studied methods exploit special characteristics of the wind noise signal in a digital representation either in the time or discrete frequency domain. In conclusion, the STM and the TSC methods present the best trade-off between a low misdetection rate of speech and a high wind noise detection rate. If extremely low speech misdetection rates are required the SSC concept outperforms the two aforementioned methods.

References

[1] 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*, Aachen, Germany, September 2012.

[2] C. Nelke and P. Vary, "Wind noise short term power spectrum estimation using pitch adaptive inverse binary masks," in *Proc. of IEEE Intern. Conf. on Acoustics, Speech, and Signal Process.*, Brisbane, Australia, April 2015.

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

[4] P. Kabal, "TSP speech database," Tech. Rep., McGill University, Montreal, Canada, September 2002.

[5] C. Nelke and P. Vary, "Measurement, analysis and simulation of wind noise signals for mobile communication devices," in *Proc. of Intern. Workshop on Acoustic Echo and Noise Control*, Sophia-Antipolis, France, September 2014.

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

[7] F. Heese, C. Nelke, M. Niermann, and P. Vary, "Self-learning codebook speech enhancement," in *ITG-Fachtagung Sprachkommunikation*, Erlangen, Germany, September 2014.

[8] ITU-P, "P.50: Artificial voices," September 1999.

[9] E. Nemer, W. LeBlanc, M. Zad-Issa, and J. Thyssen, "Single-microphone wind noise suppression," Patent 2010/00209, 2010.