Technical Disclosure Commons

Defensive Publications Series

September 2020

Holographic Detection and Reduction of Wind Noise

Anonymous

Holographic Detection and Reduction of Wind Noise

ABSTRACT

Many devices that include built-in microphone(s) are used in windy situations. Wind noise degrades the quality of audio detected by the microphone(s), causes microphone signal saturation at high wind speeds, causes nonlinear acoustic echo, and reduces the performance of acoustic echo cancellation (AEC). Applications such as voice-trigger, automatic speech recognition (ASR), and voice over internet protocol (VoIP) communication are negatively impacted by such degradation.

This disclosure describes cost-effective and robust techniques to detect and reduce wind noise. The described techniques deliver optimum removal and detection results by processing the audio signal in a holographic way by dealing with all related domains including time, frequency, and 3D space. This approach can improve the audio detection performance of any device that incorporates the techniques and can thereby improve the user experience of various applications such as voice-trigger, speech recognition, voice communication, event detection, etc. even on devices that have limited computational capability.

KEYWORDS

- Holographic processing
- Wind noise
- Noise reduction
- Noise removal
- Speech enhancement
- Acoustic echo cancellation (AEC)
- Adaptive noise cancellation (ANC)
- Automatic speech recognition (ASR)
- Voice-trigger
- Wake word
- Activation word
- Multi-microphone device

BACKGROUND

Many devices that include built-in microphone(s) are used in windy situations. Examples of such devices include augmented reality (AR) devices, smartphones and other mobile devices, personal digital assistants, wearable devices, hearing aids, devices used in automobiles (e.g., in-car entertainment or navigation systems), home security monitoring devices, tablets, computers, etc. In some instances, e.g., when using a VR device that couples to a smartphone, the smartphone can serve as a microphone for the user's VR experience.

In such situations, wind noise degrades the quality of audio detected by the microphone(s), causes microphone signal saturation at high wind speeds, causes nonlinear acoustic echo, and reduces the performance of acoustic echo cancellation (AEC). Applications such as voice-trigger, automatic speech recognition (ASR), and voice over internet protocol (VoIP) communication are negatively impacted by the degradation. The presence of wind noise can reduce audio event detection performance of outdoor home security devices.

However, handling wind noise is a very challenging problem. Many different techniques are currently utilized to detect and remove wind noise. Such techniques include, e.g., negative slope fit (NSF) approach, neural network (NN) or machine learning (ML) based approaches, non-negative sparse coding approach, singular value decomposition (SVD) approach, and generalized SVD (GSVD) subspace method. However, the shortcomings of these current techniques prevent them from being suitable in many practical applications in hardware devices in which wind noise detection and removal features are required.

Wind noises are statistically complex and have highly non-stationary characteristics, causing the traditional background noise reduction techniques listed above to fail to work properly. The negative slope fit (NSF) approach of wind noise detection assumes that wind noise

can be approximated as linear decay in frequency domain. However, this assumption can cause inaccuracies in detection. Neural network and machine learning based wind noise detection approaches require extensive network or model training. To support various types of wind and voice signals, noise-aware training is needed and requires a consistent estimate of noise, which is often difficult with highly non-stationary wind noise.

Non-negative sparse coding approach of wind noise reduction converges very slowly when delivering stable results and only works for signal-to-noise ratio (SNR) larger than 0.0 dB, which is not the case in many practical situations. Singular value decomposition (SVD) and the generalized SVD (GSVD) subspace methods are too complicated and unsuitable for implementation on low power devices. Importantly, these existing techniques fail to provide sufficient detection and reduction accuracies when the wind noise is of high speeds and/or when the desired voice and the high speed winds are simultaneously present which is the case in many practical applications.

Cost-effective and robust techniques that accurately detect and reduce wind noise can greatly improve the performance of various applications such as AEC, voice-trigger, ASR, voice communications, acoustic event detection, etc. For example, taking the application of adaptive noise cancellation (ANC) as used in headsets/headphones as a representative example, efficient and robust wind noise detection techniques can help in ensuring that the ANC filter adaptation can be frozen timely and accurately to avoid divergence in the presence of wind.

DESCRIPTION

Some key performance characteristics for wind noise detection and reduction (WNDR) techniques are the required processing power (which is a function of computational complexity in terms of computations and memory access) and the accuracy of detection, estimation, and

reduction. This disclosure describes cost-effective, robust, and holographic techniques to detect and reduce wind noise. The techniques are able to deliver optimum removal and detection results by processing the audio signal in a holographic way by dealing with all related domains including time, frequency, and 3D space. This approach can improve the audio detection performance of any device that incorporates the techniques and can thereby improve the user experience of various applications such as voice trigger, speech recognition, voice communication, event detection, etc. even on devices that have limited computational capability.

To understand the wind noise detection and removal techniques of this disclosure, first an introduction to the statistical and physical characteristics of wind noise is useful. For example, wind noise of 8 to 12 mph speeds (gentle breeze) can cause leaves and small twigs to move, light flags to extend; wind noise of 13 to 18 mph speeds (moderate breeze) can cause small branches to sway and dust and loose paper to blow about; wind noise of 19 to 24 mph speeds (fresh breeze) can cause small trees to sway and waves to break on inland waters; and wind noise of 25 to 31 mph speeds (strong breeze) can cause large tree branches to sway and can make umbrellas difficult to use. Applications such as automatic speech recognition (ASR) in severe wind conditions (wind speeds of 25 mph and above) may be infeasible for any wind noise reduction approaches.

Wind noise spectrum falls off inversely with frequency. Moreover, wind noise for wind speeds up to 12 mph speeds is strongly active in the low frequency region up to 500 Hz. At higher wind speeds, e.g., 13 to 24 mph, wind noise can affect frequencies of up to 2 kHz or thereabouts.

Taking into account the above wind noise characteristics, the described holographic wind noise detection (WND) and wind noise reduction (WNR) techniques are performed jointly in

time domain (TD), frequency domain (FD), and spatial domain (SD) in three dimensions. The specific implementations are user-configurable, e.g., to support different applications such as voice-trigger, ASR, and voice communications (e.g., VoIP or other calls where a human listener receives audio after WNR). For example, WNR can be configured to focus on reducing noise only in the low frequency range up to 2 kHz for voice-trigger and ASR applications such that voice signal remains uncorrupted from 2 kHz. For calls or other applications where a human listener is present, holographic WNR can be configured to reduce wind noise up to 3.4 kHz for narrowband voice calls and up to 7.0 kHz for wideband voice calls.

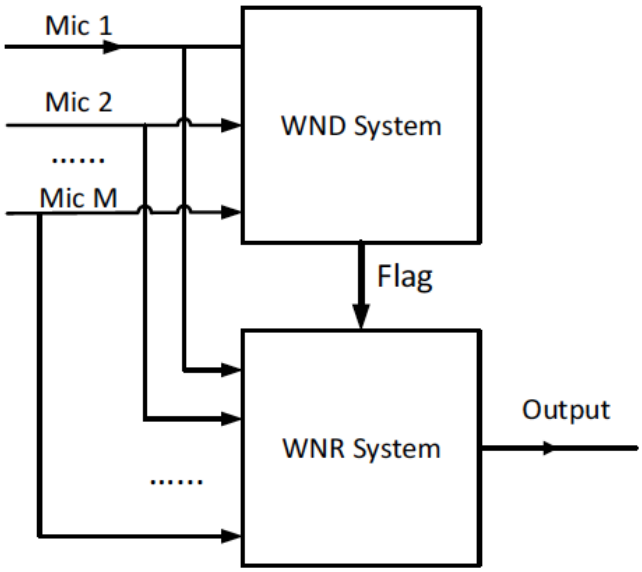

Figure 1: Multi-Microphone System for Holographic Removal and Detection of Wind Noise

Figure 1 illustrates an overview of a multi-microphone holographic wind noise detection and reduction (WNDR) techniques of this disclosure. As shown in Figure 1, *M* microphones are used to detect audio which is fed to a WND system and also to a WNR system. The WND system provides a flag that when "True" indicates that wind is present. Wind noise detection is described in detail with reference to Figure 2 below.

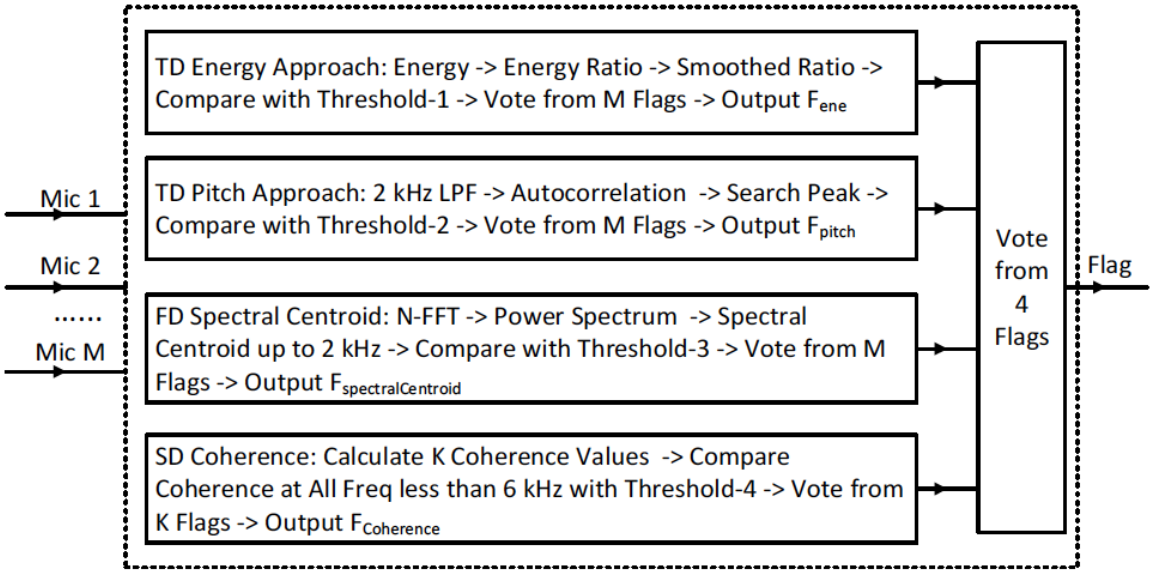

Figure 2: Multi-Microphone wind noise detection

As illustrated in Figure 2, wind noise detection is based on feature extraction in the low frequency region. These features include the energy, pitch, and spectral centroid features of each microphone signal, and coherence feature between *M* microphone signals. Other features can also be utilized.

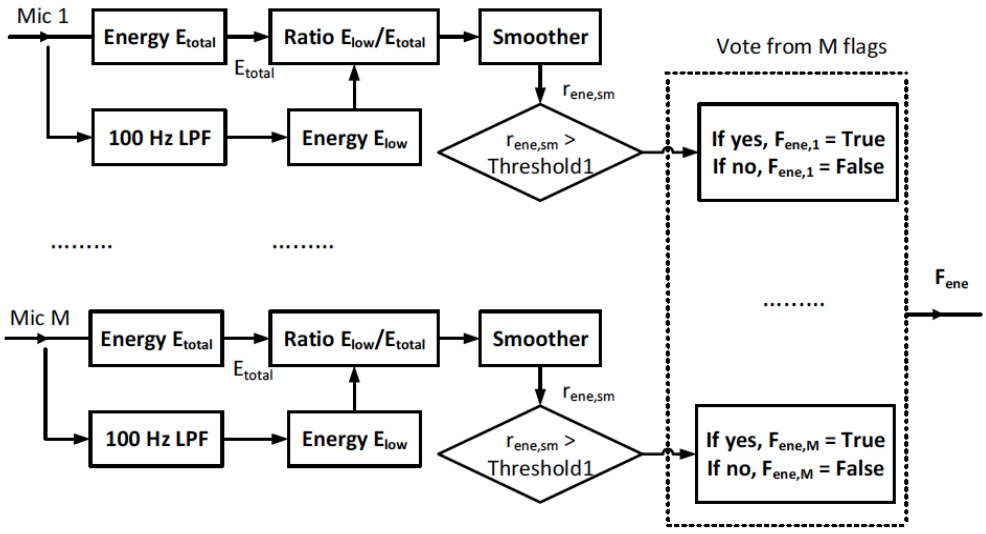

Figure 3: Energy Feature Extraction based WND Algorithm (TD Approach)

Wind noise detection

Time domain approach: Energy feature extraction

Wind noise energy dominates in frequencies lower than 100 Hz when both wind noise and voice are present together. Figure 3 illustrates energy extraction based wind noise detection.

The *m-th* frame of each microphone signal *[x(m, 0), x(m, 1), x(m, 2), ..., x(m, L-1)]* (where *L* is the frame length in unit of samples) is processed by a 100 Hz lowpass filter (LPF) and results in a frame of the filtered signal *[y(m, 0), y(m, 1), y(m, 2), ..., y(m, L-1)]*. The energies of the filtered signal and the original signal (i.e., E_{low} and E_{total}) can be calculated as follows:

$$E_{low}(m) = \frac{1}{L}\sum_{n=0}^{L-1}[y(m,n)]^2 \quad (1)$$

$$E_{total}(m) = \frac{1}{L}\sum_{n=0}^{L-1}[x(m,n)]^2 \quad (2)$$

Next, define the ratio $r_{ene}(m)$ between $E_{low}(m)$ and $E_{total}(m)$ as follows:

$$r_{ene}(m) = \frac{E_{low}(m)}{E_{total}(m)} \quad (3)$$

For robustness of this feature extraction, smooth the ratio $r_{ene}(m)$ as follows:

$$r_{ene,sm}(m) = r_{ene,sm}(m-1) + \alpha * (r_{ene}(m) - r_{ene,sm}(m-1)) \quad (4)$$

where α is a smoothing factor and ranges from 0.0 to 1.0.

If the smoothed ratio $r_{ene,sm}(m)$ is larger than a first threshold (e.g., 0.45), it can be determined that wind noise is present in this microphone since wind noise energy dominates in frequencies lower than 100 Hz when both wind noise and voice are present together. If more than half the microphones (*M/2*) indicate the presence of wind noise, the energy feature indicates wind noise is present in the system such that the flag F_{ene} is set as true, otherwise the flag F_{ene} is

set as false.

Time domain approach: Pitch extraction

Pitch extraction is performed mainly to make use of the property that the wind noise does not have any pitch. Each microphone signal is processed by a low-pass filter (e.g., a 2 kHz LPF), and the pitch f_0 is estimated by using autocorrelation approach and the filtered signal. To enhance the robustness of the feature extraction, the obtained autocorrelation values are smoothed over time. If all of the smoothed autocorrelation values are smaller than a second threshold (e.g., 0.40), it can be determined that the wind noise is present, because the wind noise dominated frame does not have the pitch f_0. If more than half the microphones ($M/2$) indicate the presence of the wind noise, then the pitch feature indicates wind noise is present and the flag *Fpitch* is set as true, otherwise the flag *Fpitch* is set as false.

Frequency domain approach: Spectral centroid extraction

Spectral centroid is correlated to sound brightness. Wind noise is of much lower spectral centroid than voice signal. Each microphone signal is of fs (Hz) sampling rate and is processed by an N-point fast Fourier transform (FFT) (e.g., fs = 16 kHz, N = 256). The frequency resolution Δf is fs/N (Hz). The 2.0 kHz frequency is around the *J-th* bin which can be obtained by the following equation:

$$J = \text{integer of } (2000.0/\Delta f) \qquad (5)$$

The frequency at the *J-th* bin is $f_J = J * \Delta f$ (Hz).

The spectral centroid $f_{sc}(m)$ in the *m-th* frame (in unit of Hz) is calculated as follows:

$$f_{sc}(m) = \frac{\sum_{k=0}^{J} f(k)X(m,k)}{\sum_{k=0}^{J} X(m,k)} \qquad (6)$$

where *X(m, k)* represents magnitude spectrum of time domain signal in the *m-th* frame at the *k-th* bin, and *f(k)* is the frequency (Hz) of the *k-th* bin such that $f(k) = k * \Delta f$ (Hz).

An alternative approach of calculating the spectral centroid *fsc* is to replace the magnitude spectrum by power spectrum in Equation (6).

For the robustness of this feature extraction, *fsc(m)* is smoothed as follows:

$$f_{sc,sm}(m) = f_{sc,sm}(m-1) + \beta * (f_{sc}(m) - f_{sc,sm}(m-1)) \quad (7)$$

where β is a smoothing factor and ranges from 0.0 to 1.0.

If the smoothed spectral centroid $f_{sc,sm}(m)$ is less than a third threshold (e.g., 40 Hz), then it is determined that the wind noise is present at the received audio samples. If more than *M/2* microphones indicate the presence of wind noise, then the spectral centroid feature indicates wind noise is present in the received signals and the flag $F_{spectralCentroid}$ is set as true; otherwise, the flag $F_{spectralCentroid}$ is set as false.

Spatial domain approach: Coherence extraction

Wind noise is of very low correlation in frequencies lower than 6 kHz. Wind noise is incoherent between 2 microphones separated by 1.8 cm to 10 cm. The coherence value of wind noise signal is close to 0.0 for frequencies up to 6 kHz, while the coherence of voice signal is much larger than 0.25.

To select two microphone signals from *M* microphone signals, there are *K* coherence values with *K* defined as follows:

$$K = \binom{M}{2} = \frac{M(M-1)}{2(2-1)} = \frac{M(M-1)}{2} \quad (8)$$

The coherence between two microphone signals (e.g., *x(t)* and *y(t)*) is calculated as follows:

$$C_{xy}(f) = \frac{|G_{xy}(f)|^2}{G_{xx}(f)G_{yy}(f)} \qquad (9)$$

where $G_{xy}(f)$ is the cross-spectral density (CSD) (or cross power spectral density (CPSD)) between microphone signals $x(t)$ and $y(t)$, and $G_{xx}(f)$ and $G_{yy}(f)$ are the auto-spectral density of $x(t)$ and $y(t)$, respectively. The CSD or CPSD is the Fourier transform of the cross-correlation function, the auto- spectral density is the Fourier transform of the autocorrelation function.

Values of coherence always satisfy $0.0 \leq C_{xy}(f) \leq 1.0$. If the coherence values are equal to zero, it is an indication that microphone signals $x(t)$ and $y(t)$ are completely unrelated.

If the coherence values are smaller than a fourth threshold (e.g., 0.25) for all the frequencies ranging from 0 Hz to 6 kHz, then it can be determined that the wind noise is present at the *m-th* frame. This is because wind noise is incoherent between 2 microphones separated by 1.8 cm to 10 cm. The coherence value of wind noise signal is close to 0.0 for frequencies up to 6 kHz, while the coherence of voice signal is much larger than 0.25. If more than $K/2$ coherences indicate the presence of the wind noise, then the coherence feature indicates wind noise is present and the flag $F_{coherence}$ is set as true; otherwise, the flag $F_{coherence}$ is set as false.

As illustrated in Figure 2, the above four features can further be combined in a statistically rigorous way so as to optimally detect the presence of the wind noise. For example, if two or more features indicate the presence of the wind noise, then the WND system indicates wind noise is present.

Once a frame is determined as the case having wind noise, an M-channel ramped dynamically sliding HPF, an M-channel adaptive beam-former (ABF), and a spectral shaping filter are applied to reduce the wind noise in TD, SD, and FD dimensions, respectively, which are illustrated in Figure 4.

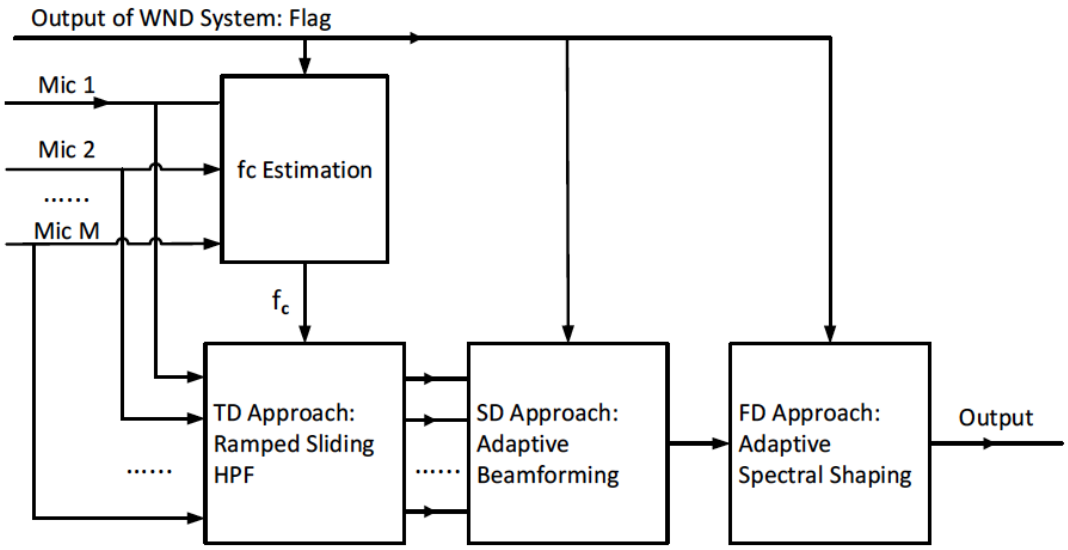

Figure 4: Multi-Microphone Holographic WNR System

Below are details of the three key blocks of Fig. 4- the TD, SD and FD components of the holographic processing.

Wind noise reduction

Time domain approach: ramped dynamic sliding HPF

The estimation algorithm for the cutoff-frequency f_c of the sliding HPF can be implemented as follows. If wind noise is not present according to the output of Figure 2, set f_c as 80 Hz. Otherwise, a cumulative energy is calculated from 80 Hz to 500 Hz for each microphone signal. To reduce computational complexity, either of the magnitude spectrum or the power spectrum obtained in the spectral centroid extraction process described above is used to calculate the cumulative energy.

If the cumulative energy of the *i-th* microphone signal ($i=1, 2, ..., M$) at frequency $f_{c,i}$ is larger than a fifth threshold (e.g., 200.0), then the $f_{c,i}$ is chosen as the potential cutoff frequency of the ramped sliding HPF. The final f_c is calculated as follows.

$$f_c = \frac{1}{M}\sum_{i=1}^{M} f_{c,i} \qquad (10)$$

Therefore, f_c is dynamically adjusted between 80 Hz and 500 Hz. These results are used for wind noise reduction time domain processing component (described below), in which each microphone signal is processed by a ramped sliding HPF with the same filter coefficients.

The ramped sliding HPF can be implemented by a second order infinite impulse response (IIR) filter whose design is given in the following steps. First, let $cs = cos(2\pi(f_c/f_s))$ and $\gamma = sin(2\pi(f_c/f_s))/(2Q)$ with Q being the quality factor (e.g., $Q = 0.7071$ for Butterworth filter).

$$b1 = -(1.0 + cs),\ b0 = -b1/2.0,\ b2 = b0$$

$$a0 = 1.0 + \gamma,\ a1 = -2.0 * cs,\ a2 = 1.0 - \gamma$$

Filter coefficients can then further be normalized as follows:

$$\text{HPF numerator } B = [b0/a0\ \ b1/a0\ \ b2/a0] \qquad (11)$$

$$\text{HPF denominator } A = [1.0\ \ a1/a0\ \ a2/a0] \qquad (12)$$

When wind noise is present, the filter coefficients are linearly ramped on each processed audio sample according to coefficient increments (e.g., 0.01). The original coefficients *A* and *B* vectors are kept unchanged. The increments and the ramping length are selected such that the filter coefficients reach their final value at the end of ramping function call. At the end of ramping, the ramping function is set as bypass mode so as to use the original coefficients *A* and *B* vectors and reduce the computational complexity. Each microphone signal is processed by the same ramped dynamic sliding HPF.

Spatial domain approach: Adaptive beamformer (ABF)

The spatial domain processing of Figure 4 is implemented by an adaptive beamforming

approach, such as, a differential beamformer or a minimum variance distortionless response (MVDR) beamformer. Differential beamformers can boost the signals that have low correlation between the microphone signals. This boosting mainly happens at low frequencies. In order to limit wind noises with having low correlation at low frequencies, a constraint or regulation rule is utilized for better determining the beamformer coefficients, such that the differential beams are of omni patterns below about 500 Hz.

Alternatively, the MVDR beamformer can be adopted to reduce wind noise in spatial domain. The SNR of beamformer outputs is described as follows:

$$SNR = \frac{E[|W^H S|^2]}{E[|W^H N|^2]} = \frac{\sigma_s^2 \, |W^H \boldsymbol{a}(\theta)|^2}{W^H R_n W} \qquad (13)$$

where W is a complex weight vector, H denotes the Hermitian transform, R_n is the estimated noise covariance matrix, and σ_s^2 is the desired signal power, $\boldsymbol{a}$ is a known steering vector at direction θ.

The beamformer output signal at the time instant n can be written as $y(n) = W^H x(n)$.

In the case of a point source, the MVDR beamformer is obtained by minimizing the denominator of the above SNR Equation (13) in the form of the following optimization problem:

$$\min_W(W^H R_n W) \text{ subject to } W^H \mathbf{a}(\theta) = 1 \qquad (14)$$

where $W^H \boldsymbol{a}(\theta) = 1$ is the distortionless constraint applied to the signal of interest. The solution of the optimization problem (14) can be found as follows:

$$W = \lambda \boldsymbol{R}_n^{-1} a(\theta) \qquad (15)$$

where $(.)^{-1}$ denotes the inverse of a positive definite square matrix, λ is the normalization constant that does not affect the output SNR Equation (13) and can be omitted in some implementation for simplicity.

Frequency domain approach: Spectral shaping

The frequency domain processing of Figure 4 is implemented by spectral filtering approach (spectral shaping). The spectral shape of the spectral filter is dynamically estimated from that frame having wind noise. This spectral shaping suppresses wind noise in frequency domain.

The spectrum of the estimated clean voice in FD is modeled as follows.

$$|X(m, k)|^2 = H(m, k) * |Y(m, k)|, k = 0, 1, ..., N/2 \quad (16)$$

where $H(m, k)$ and $|Y(m, k)|$ are the spectral weight and input magnitude spectrum at the *k-th* bin and in the *m-th* frame, N is the FFT length. Wind noise spectral shape $|W(m, k)|^2$ in the *m-th* frame at the *k-th* bin can be estimated from the input spectrum when the WND System indicates the presence of the wind noise. The frequency at the *k-th* bin is of $f_k = k*fs/N$ (Hz) and fs is the sampling rate.

Without loss of generality, assume that f_{Limit} is 2 kHz, 3.4 kHz, and 7.0 kHz for voice-trigger and ASR applications, narrowband voice calls, and wideband voice calls, respectively. Set $H(m, k) = 1.0$ under the condition of $f_k \geq f_{Limit}$, otherwise, $H(m, k)$ can be calculated through one of the following approaches.

$$Wiener\ Filtering: H(m,k) = 1 - \mu\frac{|W(m,k)|^2}{|Y(m,k)|^2} \quad (17)$$

$$Power\ Spectral\ Substraction: H(m,k) = \sqrt{1 - \mu\frac{|W(m,k)|^2}{|Y(m,k)|^2}} \quad (18)$$

$$Magnitude\ Spectral\ Subtraction: H(m,k) = 1 - \mu\frac{|W(m,k)|}{|Y(m,k)|} \quad (19)$$

where parameter μ is between 0.0 and 1.0. The values of spectral weight always satisfy $0.0 < H(m, k) \leq 1.0$.

While the foregoing discussion refers to the use of a multi-microphone system for holographic detection and removal of wind noise, the described techniques can also be utilized in devices with a single microphone with the modifications shown in Figure 5 below. In such implementations, the focus is on only the time domain and frequency domain.

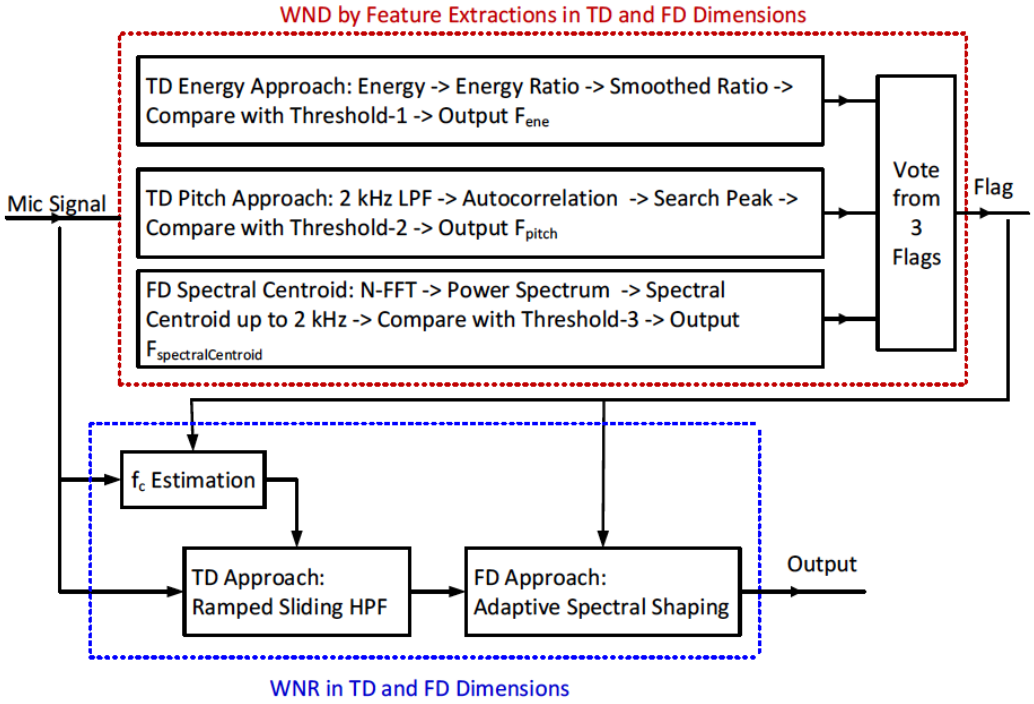

Figure 5: Single Microphone Holographic WNDR System

Some important features of the holographic wind noise detection and reduction system described in this disclosure include:

1. The holographic processing supports any sampling rate larger than or equal to 8 kHz.
2. The holographic processing works for any frame size larger than or equal to 8 msec.
3. The holographic processing works for any number of audio channels.
4. The different thresholds described in the above discussion are configurable in such a way as to support various applications and products. The thresholds can be predefined and can be tuned by the related corresponding training datasets.

CONCLUSION

This disclosure describes cost-effective and robust techniques to detect and reduce wind noise. The described techniques deliver optimum removal and detection results by processing the audio signal in a holographic way by dealing with all related domains including time, frequency, and 3D space. This approach can improve the audio detection performance of any device that incorporates the techniques and can thereby improve the user experience of various applications such as voice-trigger, speech recognition, voice communication, event detection, etc. even on devices that have limited computational capability.

REFERENCES

[1] Daniele Mirabilii, Emanuel A.P. Habets, “Multi-channel Wind Noise Reduction Using the Corcos Model,” ICASSP 2019, 12-17 May 2019, Brighton, United Kingdom.

[2] Jun Yang, Joshua Bingham, “Environment-aware Reconfigurable Noise Suppression,” ICASSP 2020 Virtual Conference, PP.3042-3046, 4-8 May 2020, Barcelona, Spain.

[3] Toya Kitagawa, Kazuhiro Kondo, “Evaluation of a Wind Noise Reduction Method Using DNN for Bicycle Audio Augmented Reality Systems,” 2018 IEEE International Conference on Consumer Electronics-Taiwan (ICCE-TW), 19-21 May 2018, Taichung, Taiwan, Electronic ISSN 2575-8284.

[4] Shiqiang Wang, Jorge Ortiz, "Non-negative matrix factorization of signals with overlapping events for event detection applications", Acoustics Speech and Signal Processing (ICASSP) 2017 IEEE International Conference on, pp. 5960-5964, 2017.

[5] Kisoo Kwon, Jong Won Shin, Nam Soo Kim, "NMF-Based Speech Enhancement Using Bases Update", Signal Processing Letters IEEE, vol. 22, no. 4, pp. 450-454, 2015.

[6] Omar Eldwaik, Francis F. Li, "Mitigating wind noise in outdoor microphone signals using a singular spectral subspace method", Innovative Computing Technology (INTECH) 2017 Seventh International Conference on, pp. 149-154, 2017.