COHERENT MODULATION COMB FILTERING FOR ENHANCING SPEECH IN WIND NOISE

Brian King and Les Atlas

University of Washington, Department of Electrical Engineering, Seattle, Washington, USA

ABSTRACT

Enhancing single-channel speech corrupted by wind noise has proven to be a difficult topic due to the complex characteristics of wind noise. Methods that assume a stationary or quasistationary noise source are ineffective against wind noise due to its nonstationarity and unpredictability. In contrast, the new method proposed works by finding the elements of the signal that are speech-like and suppressing the noise. This method takes advantage of the harmonic nature of speech by using a coherent modulation comb filter. Traditionally, very high-order IIR filters have potentially crippling stability constraints, but the proposed method bypasses these constraints by using coherent demodulation to filter harmonic subsets with lower-order filters. Potential applications for this research include mobile phones, audio production software, and as a front-end for automatic speech recognition (ASR) systems.

1. INTRODUCTION

Much work has been done in the area of signal processing for speech enhancement. Two standard methods of noise reduction are Wiener filtering [1] and spectral subtraction [2]. Both of these algorithms work on the assumption that the noise is stationary or quasistationary and perform well when the noise fits these characteristics. Wind noise, however, is highly nonstationary and unpredictable, causing such methods to perform poorly [3]. Other methods include a hidden Markov model with Gaussian mixture model [4], vector quantization [5], and non-negative sparse coding [3]. The methods above require training sets of either speech, wind noise, or both in order to develop models. In contrast, the method proposed introduces a new type of wind noise removal, one that doesn't rely on building models from training data. Since this method is fundamentally different from the above modeling methods, it does not directly compete with them and can potentially be combined with a previous method to perform better than either one separately.

2. PROPOSED FILTER MODEL

The filter model proposed takes advantage of the characteristics of the speech and noise for speech enhancement and noise suppression, so before discussing the filter model, the speech and wind characteristics of interest will be presented. The noise model for wind is that of a nonstationary noise source with the energy concentrated in the lower frequencies and rolling off at approximately $\frac{1}{f}$ as the frequency increases [6] (see Figure 1). The nonstationarity occurs because the bursts of wind are statistically dynamic and unpredictable. The speech model used [7] categorizes speech as being either voiced or unvoiced. Speech having significant harmonic content is modeled as voiced, while inharmonic components of speech are unvoiced. It has been observed that voiced speech's energy is concentrated in the lower frequencies (<4 KHz) while the majority of unvoiced speech is in the higher frequencies (>4 KHz).

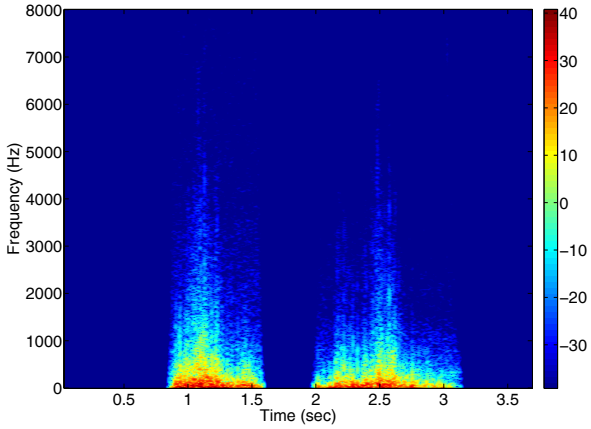

Fig. 1. Spectrogram of Wind Noise Bursts

The proposed filter model is divided into two primary components, one for high frequencies (>4 KHz) and one for low frequencies (<4 KHz). The high frequency module is a simple FIR highpass filter, which is possible because only a small percentage of noise energy is present in the higher frequencies. In the lower frequencies, however, the SNR can often be less than 0 dB. The filter model takes advantage of the fact that the majority of low-frequency content in speech is harmonic in nature. A coherent modulation comb filter is used to extract the harmonics from the signal. These harmonic signals contain both the energy of the speech and the noise at the harmonic intervals. But within the harmonic frequencies, the speech has a much higher SNR than the original signal because the noise energy is spread more evenly throughout the spectrum, while the speech energy is concentrated at the harmonics.

Other components of the wind noise removal system include a pitch tracker, wind detector, and voiced speech detector. For more details on the complete wind removal system, please refer to [8].

3. CONVENTIONAL COMB FILTERING

The two types of comb filtering are FIR and IIR. FIR comb filtering for speech enhancement was first explored by Shields [9] and Frazier *et al.* [10]. FIR comb filters enhance the periodic nature of a signal by placing evenly-spaced nonzero filter coefficients at the estimated pitch periods and setting all other coefficients to zero. This

This work was funded by the Air Force Office of Scientific Research Grant No. FA95500610191 and Adobe Inc.

Fig. 2. IIR Comb Filter Stability. Green represents stable parameters and red represents unstable parameters.

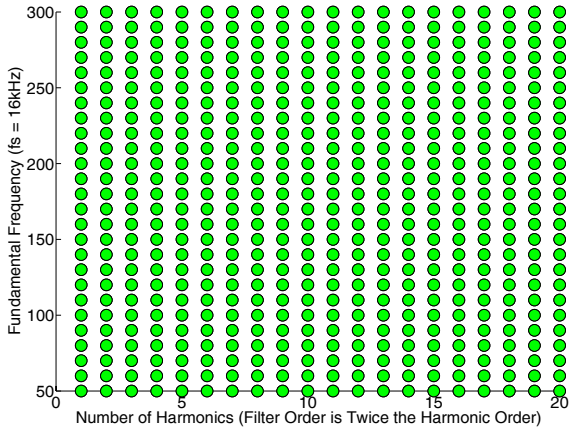

Fig. 3. Coherent Modulation Comb Filter Stability. Green represents stable parameters and red represents unstable parameters.

FIR approach produces a precisely synchronously averaged waveform of several periods. Aperiodic noise is thus attenuated. This FIR approach would work well if the speech was exactly periodic. However, since speech is not perfectly periodic, there is known distortion with this FIR method. Due to the temporal blurring caused by comb filtering the quasiperiodic speech, this method decreases intelligibility despite its reduction of perceived noise [11, 12].

The other method, IIR comb filtering, was later developed by Nehorai and Porat [13] and improves upon many of the undesirable characteristics of the FIR comb filter. The method cascades a number of second-order IIR bandpass filters to create a high-order comb filter. The IIR version is able to achieve a more ideal magnitude frequency response with a smaller order, and is less prone to the temporal blurring that plagues the FIR version.

Despite the significant advantages that the IIR comb filter has over its FIR counterpart, high-order IIR comb filters are often infeasible due to instability constraints, which can be seen in Figure 2. For example, an IIR filter containing fifteen harmonics becomes unstable below approximately 225 Hz (for $f_s = 16$ kHz), which is in the middle of the frequency range for female speech and about an octave above male speech. At lower sampling rates, the issue is even greater because the fundamental frequency at the stability boundary scales with the sampling frequency. For example, the fifteen-harmonic (30th order) filter mentioned above becomes unstable at only 125 Hz for sampling rates of 8 kHz. This means that such a filter would be impractical for many speech applications.

High-order IIR filters become extremely sensitive to quantization error, with just a small error causing a pole to jump outside the unit circle, making the filter unstable. In the work presented here, 64-bit double-precision floating point numbers were used. In hardware systems, such as mobile devices, where 32-bit or smaller words sizes are used, stability constraints are even more of an issue. In the following section we will present how coherent demodulation sidesteps these constraints to allow for IIR comb filters of arbitrary order (see Figure 3).

4. PROPOSED COHERENT MODULATION COMB FILTERING

This section will begin by giving some background for coherent demodulation and contrasting it with a previous version called incoherent demodulation. An explanation will be given why only coherent demodulation works in this case. Next, we will present how coherent demodulation is used to get around the constraints of traditional high-order IIR comb filters.

4.1. Coherent Demodulation

Coherent demodulation has been recently developed as a new way of representing a signal as a set of carriers and modulators. It had been developed to address the shortcomings of incoherent modulation filtering [14], which breaks an analytic signal into an envelope $m(n)$ and carrier $c(n)$:

$$\begin{aligned} \hat{x}_k(n) &= x_k(n) - jH\{x_k(n)\} \\ &= |\hat{x}(n)|e^{j\phi(n)} \end{aligned} \quad (1)$$

$$m_k(n) = |\hat{x}_k(n)| \quad (2)$$

$$c_k(n) = e^{j\phi(n)} \quad (3)$$

where $H\{\}$ is the Hilbert transform and the k's denote that signals may be divided into a set of bandlimited analytic signals using subbands or other methods if desired. Some of the shortcomings of incoherent demodulation are that the bandwidth of the Hilbert carrier $c_k(n)$ is typically larger than that of the original analytic signal $\hat{x}_k(n)$ [15]. Also, filtering the envelope introduces significant artifacts. Atlas *et al.* [16] have proposed that the Hilbert envelope $m_k(n)$ is incorrect for subsequent processing. By not restraining this modulation envelope to be nonnegative and real, both the carrier and modulator enjoy better characteristics, such as a reduced-bandwidth carrier and artifact-free filtering of $m_k(n)$.

Coherent demodulation is similar to the incoherent approach in that they both break up an analytic signal into a single carrier/modulator product pair. The difference, however, is how the

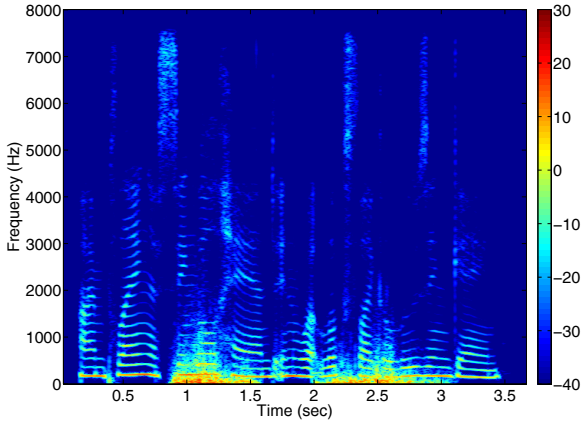

Fig. 4. Speech in Wind Noise (-6 dB SNR)

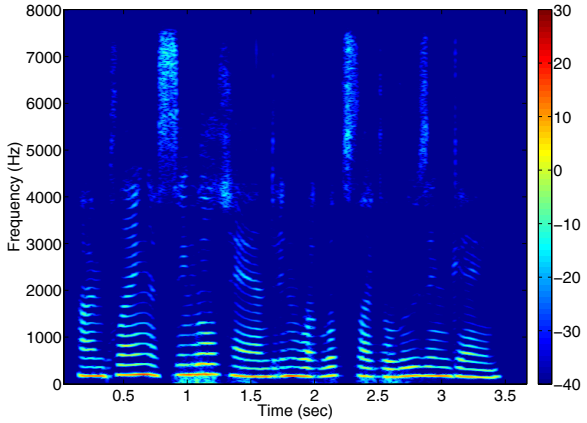

Fig. 5. Speech Processed with CMCF

carrier and modulator signals are estimated. The previously described incoherent approach simply splits a signal into a polar magnitude and phase component, which are the modulator and carrier. The coherent approach, in contrast, first estimates the carrier of the signal and then multiplies the signal by the carrier's complex conjugate to determine the potentially complex modulator:

$$\begin{aligned} \hat{x}_k(n) &= m_k(n) \cdot c_k(n) \\ &= m_k(n) \cdot e^{j\phi_k(n)} \end{aligned} \tag{4}$$

$$\begin{aligned} m_k(n) &= \hat{x}_k(n) \cdot c_k^*(n) \\ &= \hat{x}_k(n) \cdot e^{-j\phi_k(n)} \end{aligned} \tag{5}$$

After the modulator is isolated, it may be filtered as desired. Coherent modulation filtering, thus, consists of splitting a signal into a series of carriers (with or without the use of subbands), finding the coherent modulator for each carrier, filtering the modulator, recombining the filtered modulator with the original carrier, and adding the signals back together.

For coherent modulation filtering, carrier estimation is the key step. Several different estimation techniques have been proposed. One popular technique is to first split a broadband signal into subbands and to employ carrier estimation and demodulation on each subband. The estimation technique commonly employed calculates the first moment of the subband's spectral energy [17]. The carrier estimation algorithm used in this work is the least squares harmonic model [18] because it doesn't require splitting the signal into subbands. It also performs better in lower SNR's by taking advantage of the harmonic nature of speech.

Most of the research in this area to date (e.g. [19]) has used simple linear time-invariant FIR filtering on the modulator. In this case, the modulators can be thought of in another way. A lowpass modulation filter can also be considered a time-varying bandpass filter with a fixed bandwidth and a center frequency that tracks the estimated carrier frequency found by the desired carrier tracking method. The new method proposed here, coherent modulation comb filtering, modulates an adaptive comb filter so that it can filter any desired consecutive harmonics, such as harmonics 6 through 10 and higher, as seen in Figure 4. Such processing is not possible using traditional comb filters.

4.2. Coherent Modulation Comb Filter

The coherent modulation comb filter (CMCF) uses coherent demodulation techniques to extend the capabilities of a traditional IIR comb filter. Since the pitch of speech is time-varying, the time signal is divided into short frames where the pitch is assumed constant. In each frame, the signal used for the CMCF is a summation of a harmonic component $s(n)$ and an inharmonic component $v(n)$:

$$\begin{aligned} x(n) &= s(n) + v(n) \\ &= \sum_{k=1}^{P} c_k(n) \sin(k\omega_0 n + \phi_k) + v(n) \end{aligned} \tag{6}$$

The goal is to remove the inharmonic component and keep the harmonic component. To do this, a notch filter $h(n)$ is created and then subtracted from the original signal in order to emphasize the harmonics:

$$y(n) = x(n) - \frac{h(n) * x(n)}{K} \tag{7}$$

where K is the DC gain of the filter. The comb filter is created by cascading a series of identical bandpass filters together:

$$\begin{aligned} H(z) &= \frac{A(z)}{A(\rho z)} \\ &= \frac{\prod_{k=1}^{P}(1 + \alpha_k z^{-1} + z^{-2})}{\prod_{k=1}^{P}(1 + \rho\alpha_k z^{-1} + \rho^2 z^{-2})} \end{aligned} \tag{8}$$

where

$$\alpha_k = -2\cos(k\omega_0) \tag{9}$$

and the parameter ρ is the magnitude of the poles [13]. As the poles' magnitudes approach 1, the bandwidth of the filter tightens around the defined harmonics. The filter outlined above becomes unstable for certain frequencies when the harmonic count exceeds 5, thus introducing the need for a new method of comb filtering. Coherent modulation comb filtering extends the stability constraints of traditional IIR comb filters to allow filtering of any number of harmonics. By coherently demodulating the original signal, normal comb filters can be used to filter up to five consecutive harmonics anywhere in the signal. As an example, the following steps are used to filter harmonics N_1 through N_2:

1. Compute the analytic signal of the real-valued signal:

$$\hat{x}(n) = x(n) - jH\{x(n)\} \tag{10}$$

2. Demodulate the signal by $\frac{N_1+N_2}{2} f_0$. This centers the harmonics of interest around DC.
3. Lowpass filter signal by $\frac{N_2-N_1}{2} f_0$. This filters out all frequency content outside the range of harmonics N_1 through N_2.
4. Modulate the signal by $1+\frac{N_2-N_1}{2} f_0$ to lines up the harmonic positions into $1f_0$ through $5f_0$.
5. Use a normal 5-harmonic time-varying comb filter on the signal.
6. Remodulate signal by $(1+N_1)f_0$ to return harmonics to original frequencies.

For a more information on CMCF and wind noise removal, please refer to [8].

5. EXPERIMENTS

In order to test the system, speech signals from Carnegie Mellon University's ARCTIC Corpus [20] were mixed with samples of wind noise bursts. Such an example can be seen in Figure 4, where the wind bursts can clearly be seen at 0.8-1.5 seconds and 2.0-3.0 seconds. The signal is then processed using an easy-to-make-stable 20-harmonic CMCF. Figure 5, the processed signal, shows significant noise reduction with very little artifact. Informal listening tests indicate that the coherent modulation comb filtered signal is strongly preferred to the unprocessed signal.

6. FUTURE WORK

The CMCF has opened up several areas of future research. One topic is how to optimally vary filter bandwidth across both time and harmonics. Currently, the pole magnitudes are fixed, meaning that the comb filter is always "on." It would be more desirable to be able to control the filter's bandwidth according to the speech's pitch and SNR. Also, fewer artifacts will be perceived if the bandwidth widened for higher harmonics, providing a more gradual transition between the higher and lower frequency components. Finally, much more work is necessary in testing, including formalized listening tests, using recordings of speech with wind instead of mixing them in the studio, testing the system in cases of clipped audio, and testing automatic speech recognition systems preprocessed with CMCF.

7. REFERENCES

[1] Norbert Wiener, *Extrapolation, Interpolation, and Smoothing of Stationary Time Series*, The MIT Press, 1964.

[2] S. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Transactions on Signal Processing*, vol. 27, pp. 113–120, 1979.

[3] M.N. Schmidt, J. Larsen, and Fu-Tien Hsiao, "Wind noise reduction using non-negative sparse coding," in *Machine Learning for Signal Processing, 2007 IEEE Workshop on*, 2007, pp. 431–436.

[4] Sam T. Roweis, "One microphone source separation," 2000, pp. 793–799.

[5] D.P.W. Ellis and R.J. Weiss, "Model-based monaural source separation using a vector-quantized phase-vocoder representation," in *Acoustics, Speech, and Signal Processing. ICASSP 2006 Proceedings*, 2006, vol. 5, p. V.

[6] Scott Morgan and Richard Raspet, "Investigation of the mechanisms of low-frequency wind noise generation outdoors," *The Journal of the Acoustical Society of America*, vol. 92, pp. 1180–1183, 1992.

[7] Thomas F. Quatieri, *Discrete-Time Speech Signal Processing: Principles and Practice*, Prentice Hall PTR, Nov. 2001.

[8] Brian King, *Enhancing Single-Channel Speech in Wind Noise Using Coherent Modulation Comb Filtering*, MSEE Thesis, University of Washington, 2008.

[9] U. C. Shields, *Separation of added speech signals by digital comb filtering*, SM thesis, M. I. T., 1970.

[10] R. Frazier, S. Samsam, L. Braida, and A. Oppenheim, "Enhancement of speech by adaptive filtering," in *Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP '76.*, 1976, vol. 1, pp. 251–253.

[11] Jae Lim, A. Oppenheim, and L. Braida, "Evaluation of an adaptive comb filtering method for enhancing speech degraded by white noise addition," *IEEE Transactions on Signal Processing*, vol. 26, pp. 354–358, 1978.

[12] Y. Perlmutter, L. Braids, R. Frazier, and A. Oppenheim, "Evaluation of a speech enhancement system," in *Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP '77.*, 1977, vol. 2, pp. 212–215.

[13] A. Nehorai and B. Porat, "Adaptive comb filtering for harmonic signal enhancement," *IEEE Transactions on Signal Processing*, vol. 34, pp. 1124–1138, 1986.

[14] Rob Drullman, Joost M. Festen, and Reinier Plomp, "Effect of temporal envelope smearing on speech reception," *The Journal of the Acoustical Society of America*, vol. 95, pp. 1053–1064, Feb. 1994.

[15] Oded Ghitza, "On the upper cutoff frequency of the auditory critical-band envelope detectors in the context of speech perception," *The Journal of the Acoustical Society of America*, vol. 110, pp. 1628–1640, 2001.

[16] L. Atlas, Qin Li, and J. Thompson, "Homomorphic modulation spectra," in *Acoustics, Speech, and Signal Processing. ICASSP 2004 Proceedings*, 2004, vol. 2, pp. ii–761–4 vol.2.

[17] Patrick J. Loughlin and Berkant Tacer, "On the amplitude- and frequency-modulation decomposition of signals," *The Journal of the Acoustical Society of America*, vol. 100, pp. 1594–1601, 1996.

[18] N. Abu-Shikhah and M. Deriche, "A robust technique for harmonic analysis of speech," in *Acoustics, Speech, and Signal Processing, 2001. Proceedings. (ICASSP '01). 2001 IEEE International Conference on*, 2001, vol. 2, pp. 877–880 vol.2.

[19] S.M. Schimmel, L.E. Atlas, and K. Nie, "Feasibility of single channel speaker separation based on modulation frequency analysis," in *Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on*, 2007, vol. 4, pp. IV–605–IV–608.

[20] Alan Black and Kevin Lenzo, "Festvox: Cmu_artic databases," May 2008.