Multi-Sensor Spectro-Temporal Comb Filtering for Speech Enhancement

Cenk Demiroglu, David V. Anderson, Mark. A. Clements, and Thomas Barnwell

Department of Electrical and Computer Engineering
Georgia Institute of Technology, USA
demirogc,dva,clements,tom@ece.gatech.edu

Abstract

Comb filters are popular in the speech enhancement field because of their ability to suppress noise in the voiced speech spectrum without degrading the speech quality. Adaptive comb filters (ACF) can suppress significant amounts of noise in voiced speech using the quasiperiodic properties of the voiced speech sounds. However, they rely on accurate voicing and pitch epoch detection, which is challenging at low SNRs and in nonstationary noise conditions. Here, we propose using a non-acoustic auxiliary sensor for detecting the pitch epochs and voicing. Experiment results with the ACF filter showed practically no noise reduction when noisy speech is used for pitch epoch and voicing detection. However, when the auxiliary sensor signal is used, significant improvements are obtained. Therefore, it is shown that the ACF system becomes useful, at least under the considered noise conditions, only if the auxiliary data is available. In addition to the dramatic gain with the auxiliary sensors, performance of the ACF system is further boosted by using it in tandem with a frequency-domain ACF system proposed here. Objective measures, spectrogram analysis, and subjective listening test results clearly show substantial improvement with the tandem system compared to the time-domain ACF system.

Index Terms— comb filter, multi-sensor, constant pitch transform, speech enhancement

1. Introduction

Comb filters are popular in the speech enhancement field because of their ability to suppress noise in the voiced speech spectrum without degrading the speech quality [1]. The general idea of comb filtering is to leverage the quasiperiodic structure of voiced speech for suppressing the background noise. For example, the basic comb filters work by detecting the average pitch within a short-time noisy voiced speech, and then using the pitch to estimate the harmonic locations in the speech spectrum. Noise signal between the speech harmonics can be suppressed while the speech energy that is concentrated at the harmonic locations is preserved.

There are several problems with the basic comb filtering technique. The first problem is that voiced speech is quasistationary even within a small analysis window. Therefore, periodicity assumption of the comb filter is not valid in general. Adaptive comb filters (ACF) have been developed to partly address this problem [2]. The ACF system requires the pitch epochs in a speech frame without a periodicity assumption. However, finding the correct pitch epochs becomes challenging at low SNRs and in nonstationary noise environments.

In this work, a non-acoustic sensor, a glottal electromagnetic sensor (GEMS) device, is used to detect the pitch epochs in the noisy speech signal. Such auxiliary sensors are becoming increasingly popular in speech processing applications because of their immunity to acoustic noise. Speech experiments conducted using the baseline ACF system showed practically no improvement over the noisy speech. When the GEMS signal is used for voicing and pitch epoch detection, instead of the noisy speech signal, significant noise reduction could be achieved.

Even if the correct pitch epochs and voicing values are used, the ACF system still suffers from the quasiperiodicity and the time resolution problems. Because of these problems, significant amount of noise is left between the speech harmonics after noisy speech is enhanced with the ACF filter. To clean that residual noise, a frequency-domain adaptive comb filter is proposed in Section 3. The proposed system uses constant pitch transform (CPT) to handle the quasiperiodicity of speech. When the CPT system is cascaded with the ACF system, dramatic improvement in noise reduction is observed especially at low SNRs. The results were similar both for Babble and Factory noises. Spectrogram analysis and listening tests also confirmed the improved performance with the tandem system that uses a cascade of the temporal and the spectral ACF systems.

2. Adaptive Comb Filter

Voiced speech is typically modeled with a linear filter $h(t)$ driven by a periodic impulse train $e(t)$. Thus, voiced speech

$$s(n) = h(n) * e(n). \tag{1}$$

Since $e(n)$, and hence $s(n)$, is assumed to be periodic, $s(n)$ has a perfectly harmonic structure in the frequency domain. Comb filters remove significant amount of noise from voiced speech by suppressing the noise between the speech harmonics.

One of the major problems with the comb filter is that $e(n)$ is quasiperiodic, and the duration between the impulses (pitch period) changes with time. The pitch jitter is addressed by time warping the signal in the ACF system. Let's assume that the k^{th} pitch cycle starts at time t_k and has a pitch period of T_k. Noisy speech samples within the pitch cycle k can be

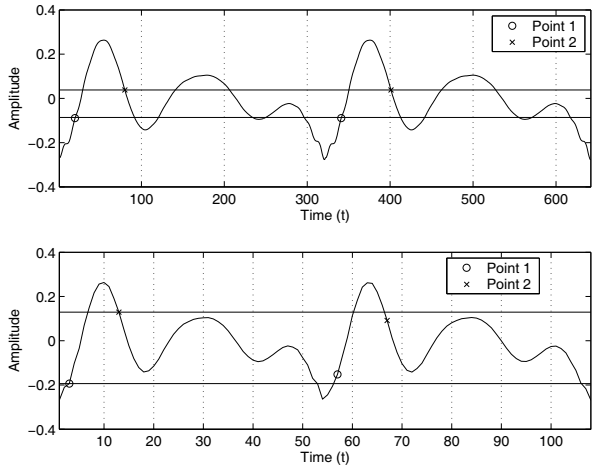

Figure 1: A perfectly periodic 48 kHz speech segment is shown in the top figure. The bottom figure shows the same speech segment downsampled to 8 kHz. Two points are selected in the first cycle, and their matching points are labeled in the second pitch cycle. The points in the first and second pitch cycles perfectly match in the first figure while they do not match in the second figure because of the reduced time resolution.

enhanced using

$$\hat{s}(t_k{+}t) = \sum_{i=-N}^{N} a_i y\left[t_{k-i} + rnd\left(t\frac{T_{k-i}}{T_k}\right)\right], t_k \le t < t_k{+}T_k \tag{2}$$

where $\hat{s}$ is the enhanced speech sample, a_i is the weighting factor, $rnd(x)$ maps the rational number x to the integer closest to x, and $y(n)$ is the noisy speech signal. The filter essentially outputs a weighted sum of the samples from $2N+1$ consecutive pitch cycles to enhance the speech samples at the k^{th} pitch cycle.

The time-warping approach of the ACF filter does not completely solve the problems related to quasiperiodicity. Besides the pitch jitter issue, there are other factors that creates quasiperiodicity. For example, neither the amplitude of the glottal impulses, nor the vocal tract shape remain constant between the pitch cycles. These two factors also introduce a source of dissimilarity between the pitch cycles, and they are not addressed with the time-warping approach.

Time resolutions is another factor that creates dissimilarity between the speech cycles. The time resolution problem is illustrated in Fig. 1. Two consecutive pitch periods with two points in each cycle in Fig. 1. When the resolution is 48 kHz, amplitude of the two points perfectly match. However, when the signal is downsampled to 8 kHz, the points have significantly different amplitudes. Thus, discretizing the signal further weakens the periodicity assumption of the ACF system.

The problems discussed above cause significant amount of residual noise in voiced speech after it is enhanced with the ACF system. To suppress the residual noise, a frequency-domain ACF system is proposed in the following section.

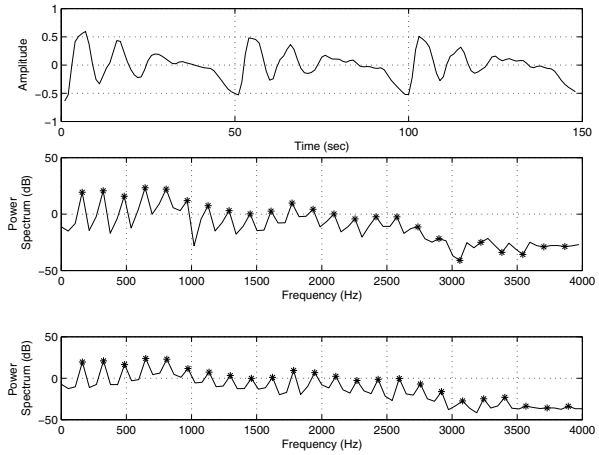

Figure 2: A voiced speech segment with three pitch cycles is shown in the top figure. Spectrum of the speech is shown in the middle figure. Spectrum of the speech after the CPT transform is shown in bottom figure. Peaks at the harmonic locations are labeled in the spectra.

3. Comb Filtering with Constant Pitch Transform

The basic idea of the constant pitch transform (CPT) is to warp the pitch cycles in the time-domain to generate consecutive pitch cycles that have the same duration. The CPT method that is used here is described below.

Let s_k denote the speech samples in the k^{th} pitch cycle of the speech signal s. Thus, s_k contains the samples t_k to $t_{k+1} - 1$. Let

$$v_k = [s_{k-N} ... s_k ... s_{k+N}] \tag{3}$$

denote the vector of speech samples that contains $2N + 1$ pitch cycles with s_k representing the middle cycle. The CPT of the vector v_k is

$$v_k^{'} = [s_{(k-N),r} ... s_k ... s_{(k+N),r}] \tag{4}$$

where $s_{(k-j),r}$ is the resampled version of $s_{(k-j),r}$ such that $s_{(k-j)}$ has the same length with s_k . A polyphase filter implementation can be used for resampling the pitch cycles.

Fig. 2 shows the spectra of v_k and $v_k^{'}$ for an example speech sample when $N = 1$. Two interesting observations can be made in Fig. 2. The first one is that the bandwidth around the harmonics get significantly smaller after the CPT transform. Thus, speech energy is concentrated more at the harmonic locations as expected from a periodic signal. The second observation is the significant improvement in periodicity at the high frequency region. The original speech signal loses its harmonicity property above 3000 Hz. After the CPT transform, however, the spectrum has high harmonicity even above 3000 Hz.

The CPT operation is an alternative to the time-warping method used in ACF for alleviating the quasiperiodicity problem. Once the pitch jitter problem is solved through the

CPT method, the harmonicity of the speech spectrum improves significantly as shown in Fig. 2. Since the harmonic locations are known, noise between the harmonics can be easily removed.

The ACF system leaves a significant amounts of residual noise because of the quasiperiodicity and time resolution problems described in the previous section. The residual noise located between the speech harmonics can be removed with the frequency-domain comb filter (CPT filter) described above. Therefore, a spectro-temporal comb filter is proposed here which cleans the noisy speech with the ACF system first, and removes the residual noise with the CPT filter. Experiment results of the ACF system, CPT system, and the tandem system are presented in Section 5.

4. Experiments

4.1. Experiment Setup

The state-of-the-art voicing and pitch detection algorithms described in [3] are used. Voiced speech is enhanced with the comb filter while the unvoiced speech is not enhanced. The sampling rate is 8 kHz. N is set to 1 in both ACF and CPT systems. A triangular window is used as the weighting function in the ACF filter with a weight vector (a_i) of $[0.25, 0.5, 0.25]$.

Two sets of experiments are performed. The goal of these first set of experiments was to measure the improvements gained by using the GEMS signal for voicing and pitch epoch detection. The ACF system is tested with and without the GEMS signal. When the ACF system is used without the GEMS signal, voicing and pitch epoch detection are done using the noisy speech signal. The second set of experiments are conducted to compare the performance of the ACF, CPT, and the tandem systems.

Performance comparisons are done using objective measures, spectrogram analysis, and subjective listening tests. Subjective listening tests and spectrogram analysis tests are done using the Harvard sentences in the ARCON speech database that was developed for the evaluations in the DARPA Advanced Speech Encoding (ASE) program. Four male, and four female speakers are used with two sentences from each speaker. Babble noise is added to noisy speech at -5 dB SNR and 8 kHz.

Objective measure tests are performed using six male and six female speakers with 50 sec of audio from each speaker. The audio is part of the ARCON database and contains Consonant-Vowel-Consonant syllables. Babble and Factory noises are added to clean speech at various SNRs between -5 and 10 dB SNRs. The noise data is obtained from the NOISEX-92 database. Segmental SNR (S-SNR) and modified Bark Spectral Distortion (MBSD) [4] measures are used as the objective measures.

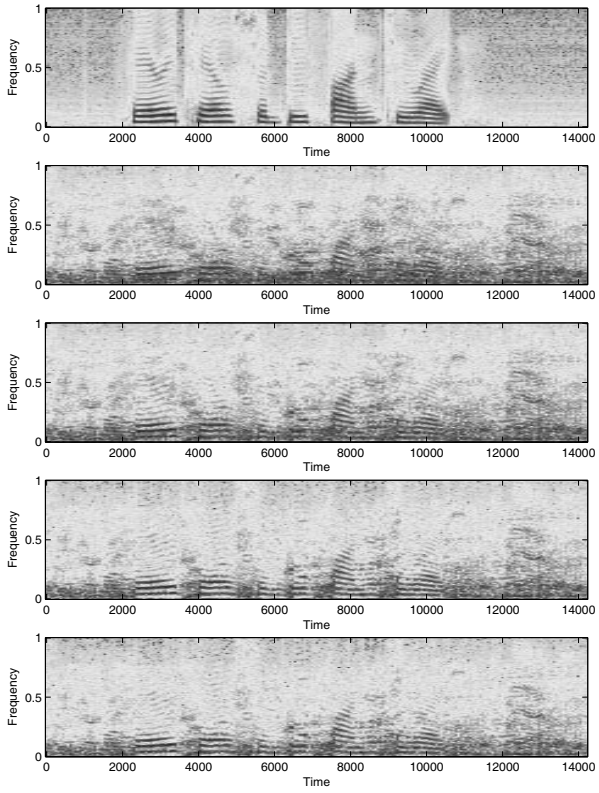

Figure 3: Top figure shows the spectrogram of a clean speech segment. The second figure shows the spectrogram of the noisy speech with the Babble noise. The three figures below them show the spectrograms of speech enhanced with ACF, Tandem, and CPT systems respectively.

5. Results and Discussion

The first set of experiments was conducted to measure the performance improvement obtained by using the GEMS signal for voicing and pitch detection as mentioned above. When noisy speech is used for detecting pitch epochs and voicing, the ACF system offered virtually no improvement over the noisy speech in S-SNR or MBSD measures at any SNR. The results were similar both for the Babble noise and the Factory noise. When the GEMS signal is used for voicing and pitch epoch detection, the ACF system significantly improved noisy speech in both measures as discussed below. Thus, the first set of results indicate the importance of accurate voicing and pitch epoch detection for the ACF system which is not always possible with a single sensor for many real-life scenarios.

The second set of experiments was performed to compare the ACF, CPT, and the tandem systems. The S-SNR and the MBSD measures are used to objectively compare the three systems. Results for the Babble noise are shown in Figs. 4 and 6, and the results for the Factory noise are shown in Fig 5 and 7. The ACF system that operates in the time-domain performed similar to the proposed CPT system that operates in the frequency domain. However, the tandem system that op-

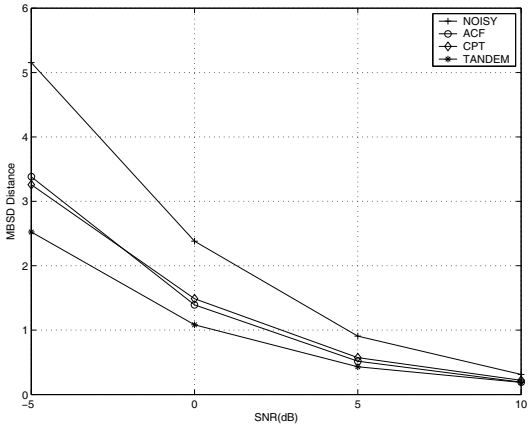

Figure 4: Comparison of ACF, CPT, and tandem systems with the noisy speech using the MBSD measure with the Babble noise.

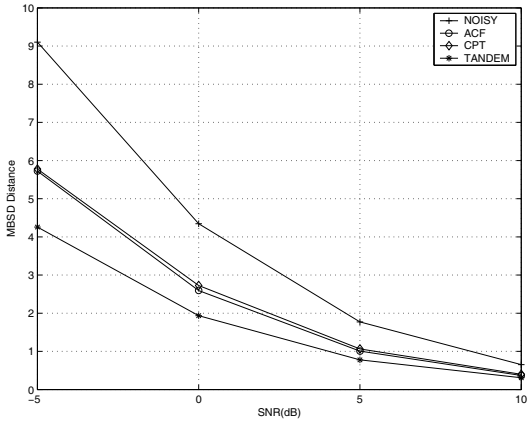

Figure 5: Comparison of ACF, CPT, and tandem systems with the noisy speech using the MBSD measure with the Factory noise.

erates in both domains significantly outperformed both systems. The performance gap is especially dramatic at -5 and 0 dB. Similar results are obtained both for Babble and Factory noises.

Spectrogram analysis is performed to further understand the operation of the three systems. One of the main observations was that the CPT system typically has a higher harmonic-to-noise ratio. This could be observed in many cases by looking at the residual noise in the spectrograms. An example case is shown in Fig. 3 where the noise between the harmonics has higher energy in the ACF spectrum compared to the CPT spectrum. Although the CPT system suppresses more noise than the ACF system in many cases, its overall noise suppression performance is similar to the ACF system. When they are concatenated, however, performance improves significantly compared to both systems.

Listening tests are performed by in-house expert listeners. Significant difference between the ACF system and the CPT system could not be observed. The tandem system significantly outperformed the ACF and CPT systems for 9 out of 16 speech samples. The samples are found to sound similar for the remaining seven samples.

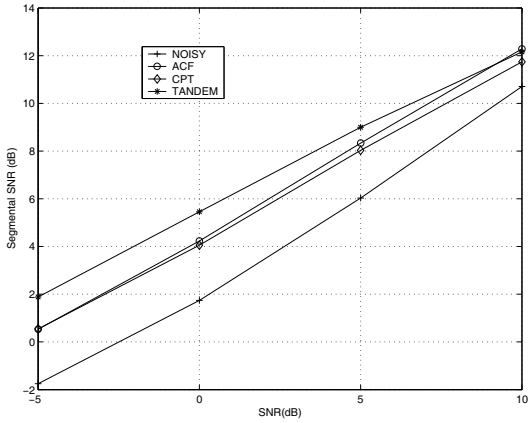

Figure 6: Comparison of ACF, CPT, and tandem systems with the noisy speech using the S-SNR measure with the Babble noise.

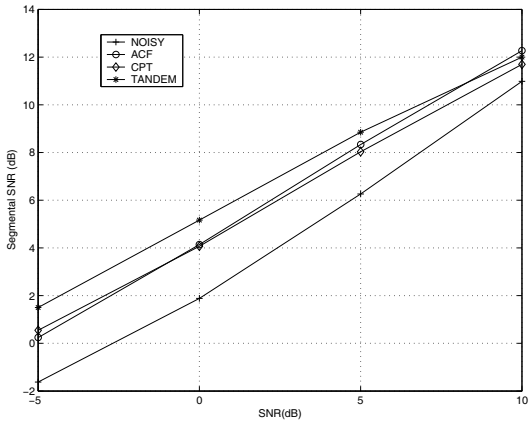

Figure 7: Comparison of ACF, CPT, and tandem systems with the noisy speech using the S-SNR measure with the Factory noise.

Comb filters are used commonly to denoise voiced speech segments and improve the speech quality. Here, a multi-sensor comb filter is proposed that can work at any SNR and any noise environment since pitch and voicing parameters are estimated with nonacoustic sensors. Moreover, a spectral comb filter is proposed that can significantly improve the performance of the time-domain ACF system when used in tandem.

6. References

[1] Y. Wang and K. Yoshida, "Speech and noise separations using comb filtering method for high quality speech coding," in *IEEE Workshop on Speech Coding*, Ibaraki, Japan, Oct 2002.

[2] R. H. Frazier, S. Samsamtt, L. D. Braida, and A. V. Oppenheim, "Enhancement of speech by adaptive filtering," in *ICASSP*, Philadelphia, 1976.

[3] A. E. Ertan, "Pitch synchronous analysis of speech signal for improving the quality of low bit rate speech coders," Ph.D. dissertation, Georgia Institute of Technology, 2003.

[4] W. Yang, M. Dixon, and R. Yantorno, "A modified bark spectral distortion measure which uses noise masking threshold," in *IEEE Speech Coding Workshop*, Pocono Manor, 1997.