Adaptive Time-Frequency Data Fusion For Speech Enhancement

Guangji Shi, Parham Aarabi, and Nevena Lazic
The Edward S. Rogers Sr. Department of Electrical & Computer Engineering
University of Toronto, Toronto, Ontario, Canada, M5S 3G4
guangji@comm.utoronto.ca, \{parham,lazic\}@ecf.utoronto.ca

Abstract – This paper proposes an adaptive time-frequency data fusion technique for the reduction of noise in speech signals using an array of two microphones. Recently, it has been shown that phase-error filtering can be a simple and effective method for "cocktail party" noise removal. However, while this technique is successful when the recorded speech signal is very noisy (less than 10dB signal-to-noise ratio (SNR)), it also tends to severely degrade the signal at higher SNRs. In this paper, the phase-error filtering technique is extended to dynamically estimate the SNR and adjust the filter parameters accordingly. Simulation results show an SNR gain of 11dB (with Gaussian noise) and an SNR gain of 17dB (Speech noise) at low SNRs without any signal degradation at higher SNRs. Speaker-independent speech recognition results using 5 speakers show that the proposed algorithm achieves a digit percent accuracy gain of 22% at 0dB and 15% at 20dB.

Keywords: Microphone arrays, data fusion, speech separation, speech enhancement, time-frequency analysis.

1 Introduction

In many speech communication applications such as automatic speech recognition, teleconferencing, and in hearing aids, the recorded speech signal may be corrupted by Gaussian noise, speech noise (unrelated conversations) and reverberations. This noise corruption is often strong enough to result in the failure of most applications. In order to enable such application to succeed in practical situations, noise removal techniques are required. Although single channel noise removal techniques such as spectral subtraction, Kalman filtering, and Wiener filtering exist, they often yield limited noise removal results. Multi-microphone based techniques such as beamforming [1,4,10] and independent component analysis (ICA) [5-9] typically offer superior performance when compared to the single microphone situation. The challenge with current multi-microphone noise removal algorithms is that they tend to have good performance without robustness (as in the case of ICA), or very limited performance with greater robustness (as in the case of beamforming) [1, 2].

In [3], it was shown that phase-error filtering can be a simple and effective microphone array based speech enhancement technique that results in large SNR gains at lower SNRs. Although this technique requires knowledge about the time-delay of arrival (TDOA) of the source of interest, this information can be obtained easily even in low SNR conditions [14-16]. For the Gaussian noise case, an SNR gain of up to 14dB was obtained at low SNRs (<20dB) although a large SNR loss was suffered at high SNRs (>20dB). In this paper, an adaptive phase-error filter is proposed which adjust the filter parameters according to the overall SNR.

In section 2, we briefly review two popular microphone array based speech enhancement methods: beamforming and ICA. In section 3, the ideas behind phase-error filtering and some simulation results are presented. In sections 4 and 5, the proposed adaptive phase-error filtering technique is presented and analyzed.

2 Previous speech enhancement methods

A number of microphone array based speech enhancement techniques have been explored in the past, including, among others, ICA and beamforming. In beamforming, signals coming from the direction of interest are amplified while signals coming from other directions are attenuated [10]. While beamforming is robust (no SNR losses), the SNR gains achieved are typically limited [10].
ICA is a speech enhancement technique that has recently gained popularity. One application of ICA is the cocktail-party problem (to identify the voice of each speaker from mixtures of the original speech signals). Given only the mixtures, ICA attempts to separate the mixtures by processing them to result in independent outputs. By the Darmois-Skitovich theorem, the independent outputs would be the separated source signals.

In [6], Bell and Sejnowski showed that ICA is able to successfully separate mixtures of independent signals. While their approach required several impractical assumptions (instantaneous mixtures, no reverberation, etc.), various versions of the ICA technique that deal with more practical mixtures have been proposed [9,11-13]. Nevertheless, while in limited situations ICA can result in very large SNR gains, in practical environments, with various noise sources and reverberations, ICA often runs into difficulty. The major difficulty with ICA is that when it fails, large signal degradations are obtained, making ICA unacceptable in many applications. In order to make microphone arrays an effective and affordable speech enhancement tool, new approaches that combine the robustness and generality of beamforming with the SNR gain of ICA are required.

3 Phase-error filtering

Phase error based filtering is a noise removal strategy originally proposed by [3]. The idea here is to minimize the mean-square phase error (phase variance), which can be defined as:

$$\psi_\tau = \frac{1}{m} \sum_\omega \theta_\tau^2(\omega) \tag{1}$$

where $\theta_\tau(\omega) = \angle X_1(\omega) - \angle X_2(\omega) - \omega \tau$ is the phase error, $\tau$ is a known TDOA for the source of interest, $m$ is the number of frequency components considered, and $X_1(\omega), X_2(\omega)$ are the Fourier transforms of the signals (at frequency $\omega$) received by the first and second microphones, respectively. The phase variance of equation (1) can also be extended to multiple segments (by summing the associated phase variances of the segments and then dividing by the number of segments). Ideally, $\psi_\tau$ should be zero. In reality, $\psi_\tau$ will often not be zero due to noise or reverberations. In fact, it can be shown that the magnitude of $\psi_\tau$ is directly related to (or upper bounded by) the intensity of the noises or reverberations [3]:

$$\psi_\tau \leq \frac{4}{m} \sum_\omega \arcsin^2 \left( \min \left( 1, \frac{1}{\sqrt{R(\omega)}} \right) \right) \tag{2}$$

where $R(\omega)$ is the SNR of the frequency block $\omega$, defined as:

$$R(\omega) = \frac{|S(\omega)|^2}{|N(\omega)|^2} \tag{3}$$

with $|S(\omega)|$ and $|N(\omega)|$ are the clean signal and noise magnitudes that make up the frequency block $\omega$ of both channels. Note that it is assumed, in the derivation of equation (3), that the signal and noise magnitudes of both microphones is identical (which is a valid assumption when the inter-microphone distance is not too large).

![Figure 1: Phase variance vs. SNR with speech noise](image1)

![Figure 2: Phase variance vs. SNR with Gaussian noise](image2)
sampling frequency, and window type are used in all of the simulations. The volume of the noise speaker is adjusted so that the input SNR ranges from -20dB to 100dB. Figure 1 shows the simulation result. The phase variance increases as the speech noise increases.

To investigate the relation between phase variance and noise level further, the phase variance is computed for different speakers. The TDOA is 5 samples and only independent Gaussian noise is used. Figure 2 illustrates the phase variance curves for two male and two female speakers.

In general, high noise levels lead to high phase variance, which is true for Gaussian noise, reverberations [3], and speech noise. In [3], a phase variance minimizing filter (based on the phase error of each frequency block) was proposed. It consisted of the following punish reward filter applied independently to all frequencies (or time-frequency blocks, in the case that multiple time-segments are used):

\[ H(\omega) = \left( 1 + \gamma \theta^2(\omega) \right)^{-1} \]  \hspace{1cm} (4)

where \( \gamma \) is an aggression constant (higher values correspond to more aggressive filters and vice versa). Depending on the value of \( \gamma \), the phase-error filter punishes each and every frequency block based on its phase error. Blocks with large phase errors are punished more and blocks with small phase errors are punished less. Figure 3 shows the results of applying this filter to a speech segment corrupted by Gaussian noise. The TDOA is 5 samples. Clearly, an SNR gain (up to 14dB) is achieved when the input SNR is low. However, when the input SNR is high, the signal is further degraded. The phase-error filter is also applied to a speech signal corrupted by speech noise. Figures 4 shows the corresponding simulation results. The simulation settings used for Figure 4 are the same as those used for Figure 1.

A common phenomenon exists in Figures 3 and 4: The output SNR saturates when the input SNR is high. This saturation is caused by an inherent error or uncertainty in the phase error reading due to the finite window size of the Fourier Transform. The finite window size blurs the true frequency contents, thereby resulting in erroneous phase readings. This effect becomes evident at higher SNRs where larger phase errors are mostly because of wrong phase readings than actual noise.

4 Adaptive phase-error filtering

One possible solution to this problem is to adjust the filter parameter to result in a more aggressive filter at lower SNRs and a more lenient filter at higher SNRs. In other words, we change based on the input SNR. The input SNR, however, is not known a-priori, and because it can dynamically change from segment-to-segment, it cannot be easily estimated. One way to estimate the input SNR is by using the relationship between SNR and phase variance. From simulations similar to those of Figures 3 and 4, we obtain the best \( \gamma \) value (the one that resulted in the highest output SNR) at each SNR. Using the SNR- relationship (similar to Figures 1 and 2), we empirically find the relationship between \( \psi_\tau \) and the best value of \( \gamma \). Figure 5 shows this relationship for four different speakers in the presence of Gaussian noise.

The curves shown in Figure 5 allow for the best filter parameter to be selected in an average sense. However, in practice, in order to ensure that the signal is not degraded at higher SNRs, it was found that a safety margin is required between the best \( \gamma \) value and the best \( \gamma \) value in an average sense. For the Gaussian noise situation, the following relationship was found to
approximate Figure 5 while having an adequate safety margin:

$$\gamma = (1 + (\psi_\tau / 2.5))^{10}$$  \hspace{1cm} (5)

This definition for the phase-error parameters results in the following filter:

$$H(\omega) = \left(1 + (\psi_\tau / 2.5)^{10} \theta_{r,k}^2 (\omega)\right)^{-1}$$  \hspace{1cm} (6)

The above adaptive relationship is derived from limited speech data of four different speakers. In order for such a relationship to be useful, it must give significant SNR improvements for other speech signals of the same speaker as well as for speech signals from a new speaker.

To verify the effectiveness of such a parameter adjustment approach, the proposed adaptive phase-error filter is applied to a different speech segment generated by one of male speakers and also to a speech segment of a new female speaker. Figure 6 shows the resulting SNR improvements. As shown, a better overall SNR gain is achieved compared with the results of using a single parameter value. An SNR gain of 11dB is achieved when the input SNR is about -20dB. When the input SNR is high, no signal degradation occurs.

The proposed adaptive technique can also be applied to the cases of speech noise. For speech noise, 4 different combinations of the 2 male and 2 female speakers were used. Using the same simulation settings as of Figure 1, the best values were obtained as shown in Figure 7. A new relation between the phase variance and filter parameter, shown below, was derived, again
Table 1: Digit recognition accuracy (%) for 5 different speakers (before and after phase-error filtering)

<table>
<thead>
<tr>
<th></th>
<th>S1</th>
<th>S2</th>
<th>S3</th>
<th>S4</th>
<th>S5</th>
</tr>
</thead>
<tbody>
<tr>
<td>0dB Noisy</td>
<td>22.3</td>
<td>23.3</td>
<td>3.3</td>
<td>0.0</td>
<td>1.7</td>
</tr>
<tr>
<td>0dB Filtered</td>
<td>40.7</td>
<td>45.0</td>
<td>21.7</td>
<td>25.0</td>
<td>28.3</td>
</tr>
<tr>
<td>10dB Noisy</td>
<td>57.7</td>
<td>50.0</td>
<td>41.7</td>
<td>38.3</td>
<td>30.0</td>
</tr>
<tr>
<td>10dB Filtered</td>
<td>63.0</td>
<td>68.3</td>
<td>58.3</td>
<td>66.7</td>
<td>56.7</td>
</tr>
<tr>
<td>20dB Noisy</td>
<td>67.0</td>
<td>73.3</td>
<td>53.3</td>
<td>85.0</td>
<td>71.1</td>
</tr>
<tr>
<td>20dB Filtered</td>
<td>78.0</td>
<td>81.7</td>
<td>91.7</td>
<td>91.7</td>
<td>81.7</td>
</tr>
</tbody>
</table>

with an adequate safety margin:

\[
\gamma = (\psi_T/0.7)^2
\]  

By applying the resulting adaptive phase-error filter to a new speech segment corrupted by speech noise, we obtain the SNR gain shown in Figure 8. An SNR gain of 17dB is achieved when the original SNR is -20dB.

One might ask why two separate filters were proposed for the different noise types. The reason is from Figures 5 and 7 it is clear that speech noise case requires much more aggressive filtering than the Gaussian noise case. Hence, for speech noise, the filtering strategy is changed by using the much more aggressive relationship of equation (7) instead of the more lenient relationship of equation (5).

5 Application to robust speech recognition

In order to test the usefulness of the proposed technique, an experiment was conducted with 4 male and 1 female speakers. A speaker-independent single digit recognition system (with no training) was built based on the Voice Extreme Module from Sensory Inc. The speech recognition system, which is small enough to be embedded in handheld applications, recorded 20-30 random set of digits (in the 0-9 range) from each speaker. The TDOA of each speaker was assumed to be known and the noise was artificially-added independent Gaussian. The filtering strategy of equation (5) was used to process the speech data (using half overlapped Hanning window sections that were added after filtering). The percent digit recognition accuracy is shown in Table 1 for all 5 speakers.

Figure 9 illustrates the percent accuracy of each of the speakers as a function of the SNR of the speech signal. As the SNR is increased, the percent accuracy is increased from an average of 12% at 0dB to an average of 70% at 20dB, without any filtering. After the adaptive phase-error filter is applied, a large recognition accuracy gain is achieved (Figure 10). The average percent accuracy gain ranges from 22% at 0dB down to 15% at 20dB.

6 Conclusions

Adaptive phase-error filtering represents a robust multi-microphone speech enhancement technique that can result in large SNR gains. Perceptually, the resulting noise-reduced signals obtained by phase-error filtering sound far better than the 10-20dB SNR gains indicated in the results. Nevertheless, the resulting gains are large enough to make phase-error filtering useful in large number of applications, including speech recognition, where the percent accuracy of the filtered result was significantly higher (especially at lower SNRs) after the filter was applied.
It should also be mentioned that phase-error filtering is applied to each of the microphone signals independently, hence leaving two signals at the end. As a result, multi-channel signal enhancement techniques such as ICA or beamforming can still be applied to the proposed filter outputs, resulting in further SNR gains.

References


