Single Acoustic–Channel Speech Enhancement based on Glottal Correlation using Non-Acoustic Sensor

Rongqiang Hu, David V. Anderson

Center for Signal and Image Processing
Georgia Institute of Technology, Atlanta, GA, USA
(rqhu, dva) @ece.gatech.edu

Abstract

This paper describes a single acoustic–channel speech enhancement, utilizing an auxiliary non-acoustic sensor. Unlike classical algorithms, which make use of the knowledge from acoustic signal alone, the glottal correlation (GCORR) algorithm takes advantage of non-acoustic throat sensors such as the general electromagnetic motion sensor (GEMS). The non–acoustic sensor provides a measure of the glottal excitation function that is relatively immune to background acoustic noise. Thus, inspired by human speech production mechanisms, the GCORR algorithm extracts the desired speech signal from noisy acoustic mixture using statistical correlation between the speech and its excitation. The algorithm leads to a significant reduction of wide–band noise, even when the SNR is very low. The improvement in the quality of the speech is demonstrated in terms of an objective evaluation.

1. Introduction

It is often essential to process speech acquired in environments having high ambient white or colored noise. This is true for consumers (as in cell phone usage), industry (e.g. communications in factory environments), and military. The presence of noise causes deterioration in both the quality and the intelligibility of speech. For most applications, some sort of speech enhancement is employed to reduce the average background noise level by 4–15 dB. However, this amount of noise suppression is still insufficient for many applications. Thus, speech enhancement remains an important research problem for improving the performance of speech processing systems in noisy environments.

One well-known family of speech enhancement algorithms is the subtractive-type algorithms. These algorithms attempt to estimate the short-time spectral magnitude of speech by subtracting estimated noise from noisy speech. The estimated speech spectrum is then used directly or as the first step in further processing such as Wiener filtering.

These algorithms rely on the knowledge from a single acoustic channel to try to derive an estimate of the noise and speech statistics. In ultra low SNR cases (<0 dB), these algorithms have difficulties in removing background noise while retaining quality speech. More recently, researchers have sought to use auxiliary non-acoustic sensors [2][3] for speech enhancement. These sensors provide measurements of speech-related processes from human body, such as the movements of the vocal tract articulators, and are significantly less susceptible to acoustic disturbance.

In this paper, we propose a single acoustic–channel speech enhancement method using a non-acoustic sensor, the general electromagnetic motion sensor (GEMS), to obtain a measurement of the vocal excitation. The GEMS signal is statistically correlated to the acoustic speech and independent of the ambient noise. These properties of the GEMS signal enable the design of a glottal correlation (GCORR) algorithm that is effective in cleaning up the noisy speech, thus improving speech quality.

In section 2, we formulate the GCORR filter and compare it to classical algorithms. In section 3, the implementation of the improved GCORR algorithm is introduced. In section 4 the performance of this algorithm is evaluated in a variety of noisy environments.

2. Glottal Correlation Approach

2.1. Classical Speech Enhancement Approach

Consider a speech signal $s(k)$ corrupted by statistically independent noise $n(k)$, the noisy mixture $y(k)$ can be represented as:

$$y(k) = s(k) + n(k).$$

The speech enhancement objective is to estimate $s(k)$, based solely on knowledge derived from $y(k)$, in such a way that the mean square error (MSE) between the estimated and original clean signals is minimized. In the frequency domain, Equation 1 can be expressed as:

$$Y(f) = S(f) + N(f)$$

where $Y(f)$, $S(f)$, and $N(f)$ are the fourier representations of the corresponding signals. A well-known algorithm, the
Wiener filter, was derived as the optimal minimum mean square error (MMSE) solution for estimating $S(f)$:

$$H(f) = \frac{\hat{P}_{ss}(f)}{\hat{P}_{ss}(f) + \hat{P}_{nn}(f)}$$  (3)

where $\hat{P}_{ss}(f)$ and $\hat{P}_{nn}(f)$ are respectively the estimated power spectra of clean signal and background noise and $\hat{S}(f) = H(f)Y(f)$. The processing is done on a short-term frame-by-frame basis, $\hat{P}_{nn}(f)$ is initially estimated in speech pauses, and smoothed out by magnitude averaging [6]. A voice activity detector is needed to detect the noise–only periods and control the smoothing. This method yields a zero–phase enhancement filter, therefore the phase of the $\hat{S}(f)$ is equal to the phase of the noisy unprocessed signal. In high noise conditions, the solution is not perceptually optimal since the phase noise is audible [1].

2.2. Glottal Correlation Approach

It is generally believed that, in human speech production, the acoustic speech signal is generated by the frequency shaping of the glottal excitation signal by the vocal tract. The description of this sound generation in the frequency domain is

$$S(f) = T(f)G(f)$$  (4)

where $T(f)$ and $G(f)$ are the vocal tract and excitation functions respectively.

Several sensors may be used to obtain a measure of $G(f)$, such as throat accelerometers, contact microphones that are impedance–matched to the skin, electro-glottal graph (EGG), and the GEMS [3]. The GEMS is an active sensor that can measure conditions of many of the internal vocal articulators, in real-time, as the speech signal is generated. It is attached on the throat at the laryngeral notch and measures electromagnetic signals reflected from the glottis. This sensor, combined with the corresponding acoustic microphone, enables an alternative formulation of the observed noisy mixture:

$$Y(f) = T(f)G(f) + N(f)$$  (5)

where $G(f)$ can be assessed by GEMS.

The glottal correlation approach to speech enhancement using the GEMS was originally proposed by Burnett [2]. Since the cross–correlation of the GEMS and speech signal is:

$$P_{gy}(f) = T(f)P_{gg}(f) + P_{gn}(f).$$  (6)

From Equation 6, an estimate of $T(f)$ [2] is derived

$$\hat{T}(f) = \frac{P_{gy}(f) - P_{gn}(f)}{P_{gg}(f)}.$$  (7)

The glottal excitation is a human internal glottis movement, the source of clean speech, it is uncorrelated to the external disturbance, $P_{gn}(f) = 0$. So, equation 7 becomes

$$\hat{T}(f) = \frac{P_{gy}(f)}{P_{gg}(f)}.$$  (8)

Using the result of Equation 8, enhanced speech can be obtained from the glottal measure and the estimated vocal tract filter

$$\hat{S}(f) = \hat{T}(f)G(f).$$  (9)

This approach yields a solution with both magnitude and phase reconstruction. From the cross-correlation between the noisy signal and the glottal excitation function, the phase and magnitude of vocal tract are constructed.

3. Proposed Implementation

3.1. Scenario

We propose an improved approach to implementing glottal correlation, which differs from that of Burnett’s [2], in that it uses an improved GEMS–speech correlation estimator and improved synchronization and pitch estimation. The scenario of the speech enhancement algorithm is shown in Figure. 1.

3.2. Synchronization

The GEMS signal is measured at or near the point of vocal tract excitation. Before the same signal is measured by the acoustic microphone, it must travel through the vocal tract, out the mouth, and some distance to the microphone. Thus, the acoustic signal is delayed relative to the excitation signal and this condition may be exacerbated by delays in the signal acquisition systems. An example of the resulting misalignment is shown in Figure. 2 but it should be noted that this is only an example and the actual misalignment can vary significantly for different setups. Thus, a synchronization is done to eliminate the delay between them by evaluating cross-correlation of the LPC residual [7] of the speech and the derivative of the GEMS.

3.3. Pitch Estimation

As demonstrated in Figure. 2, the GEMS signal is quasi–periodic and relatively immune to acoustic disturbance, which makes an accurate pitch estimation easier. A zero–crossing algorithm was introduced by Burnett that is effective in estimating pitch if a very high quality GEMS signal is obtained. However, we have observed that in practice the GEMS signal is sometimes non-ideal, especially when the talker moves. A more robust pitch estimation is applied
by carefully finding points of glottal closure using points of maximum negative slope.

### 3.4. Dynamic Windowing

Speech signals can only be considered stationary for a short time; however, the Wiener filter is based on an assumption of stationarity. Therefore, in the noise suppression algorithm, the window length should be small enough to make the stationarity assumption valid. However, to achieve good harmonic separation in the spectrum, it is desirable to have as long a window as possible. By selecting a dynamic window that is exactly equal to an integer multiple of the pitch period, it is possible to keep the window short and to improve the harmonic separation sufficiently. A dynamic window with the length of two pitch periods, from the precise start to the harmonic separation in the spectrum, is desirable. It is possible to keep the window short and to improve the analysis accuracy. The enhanced output is smoothed by overlapping adjacent windows by one pitch period.

![Figure 2: Analysis of GEMS signal alignment with acoustic speech.](image)

#### 3.5. GEMS-Speech Correlation Estimation

For short-time Fourier analysis, Equation 8 becomes

\[
\hat{T}(m, k) = \frac{P_{yy}(m, k)}{P_{yy}(m, k + 1)}
\]

where \( m \) represents the frame number and \( k \) represents the index of the DFT bin. The cross-correlation of noisy speech and excitation, \( P_{yy}(m, k) \), is crucial for estimating \( \hat{T}(m, k) \). Generally, the speech signal only remains stationary for less than 30ms. So, the estimation of \( P_{yy}(m, k) \) is a challenge in short-time processing. This is especially true in frequency bins with low \( P_{yy}(m, k) \), where the estimation is more sensitive [4]-[7].

Note that the excitation function provides fine harmonic information better than that from the noisy acoustic signal.

The power spectra in corresponding harmonics are the local peaks. Using this observation, a refined estimation for \( \hat{T}(m, k) \) is achieved as follows. Let \( K_m \) be the set of harmonic bins in \( m \)th frame. For \( k \in K_m \), \( \hat{T}(m, k) \) is estimated as a function of all the energy in the DFT peak. For \( k \notin K_m \), \( \hat{T}(m, k) \) is set to zero as shown below.

\[
\hat{T}(m, k) = \sum_{i=-\sigma_k}^{\sigma_k} \beta(k, i) \frac{P_{yy}(m, k + i)}{P_{yy}(m, k + 1)}
\]

for \( k \in K_m \) where \( \sigma_k \) is adapted based on the width of the spectral peak represented by \( k \), and \( \beta(k, i) \) is an adaptive weighting function for smoothing (typically implementing a triangular window of width \( 2\sigma_k \)).

The final temporally smoothed vocal tract function, \( \tilde{T}(m, k) \), is given by

\[
\tilde{T}(m, f) = \begin{cases} 
\alpha \hat{T}(m - 1, k) + (1 - \alpha) \hat{T}(m, k) & k \in K_m \\
0 & \text{else}
\end{cases}
\]

\( \alpha \) is a forgetting factor, typically near 0.85.

Utilizing the harmonics information extracted from the GEMS signal, synthesis is performed in frequency domain to obtain the response.

\[
\tilde{T}(m, k) = W(m) \otimes \hat{T}(m, k)
\]

where \( W(m) \) is a hamming window with the length of one harmonic period, and the operator \( \otimes \) represents convolution.

### 4. Performance Evaluation

#### 4.1. Speech Spectrograms

To indicate the structure of residual noise and enhanced speech, we present speech spectrograms in Figure. 3–Figure. 4. The noise conditions are white noise (SNR=0dB) and tank noise (SNR=–5dB).
The figures reveal that the GCORR algorithm produces a fine spectral shape in the enhanced speech even in ultra low SNR cases, especially in low–frequency bands, as seen from the marked region in the figures. The gain comes from the fact that the non-acoustic sensor yields a better measure in those bands.

4.2. SNR Improvement

The amount of noise suppression is demonstrated by the segmental SNR improvement in the voiced segments of speech. The gains were calculated as

$$SNR_{imp} = \frac{1}{L} \sum_{m=0}^{L-1} 10 \cdot \log_{10} \left( \frac{\frac{1}{N} \sum_{k=0}^{N-1} n_m^2(k)}{\frac{1}{N} \sum_{k=0}^{N-1} [s_m(k) - \hat{s}_m(k)]^2} \right)$$

where $L$ represents the number of frames in the voiced signal and $N$ is the number of samples in $m$th frame.

Experiments were performed with the data from arcorpus in various noise types and levels conditions. The results show significant improvement in speech quality as indicated in Figure 5. As a comparison, the well–known EMSR algorithm was used as the baseline system. This version of EMSR is the standard pre–processor in MELP coder with well–tuned–up parameters. The proposed algorithm shows performance superior to EMSR, especially in very low SNR and colored noise conditions. For some cases, the segmental SNR gain can almost reach 25 dB.

5. Conclusion

An enhanced glottal correlation (GCORR) algorithm is presented for speech enhancement in noisy environments using a non–acoustic sensor. This method is different from classical algorithms in that it removes magnitude and phase noise simultaneously and makes use of the knowledge of the glottal excitation function in addition to the single acoustic channel signal. The algorithm is highly effective in extracting correlated speech signals out of noise corrupted mixtures, especially in low frequency bands and in low noise cases. More than 20dB improvement in segmental SNR measure was achieved for some environment.

6. References