STFT-Based Speech Enhancement by Reconstructing the Harmonics

Iman Haji Abolhassani¹, Sid-Ahmed Selouani², Douglas O'Shaughnessy¹

¹ INRS-Energie-Matériaux-Télécommunications, Montréal, Canada
² Université de Moncton, Campus de Shippagan, Canada

imanhaji@emt.inrs.ca, sid-ahmed.selouani@umcs.ca, dougo@emt.inrs.ca

Abstract

A novel Short Time Fourier Transform (STFT) based speech enhancement method is introduced. This method enhances the magnitude spectrum of a noisy speech segment. The new idea that is used in this method is to basically reconstruct the harmonics at the multiples of the fundamental frequency (F₀) rather than trying to improve them. The harmonics are produced, in the magnitude spectrum, using the knowledge of the window function we are using for the STFT. These harmonics are then scaled and laid on multiples of F₀.

Experimental results prove the effectiveness of this enhancement method in various noisy conditions and various SNR ratios.

Index Terms: speech enhancement, speech recognition, speech magnitude spectrum enhancement

1. Introduction

Enhancement of degraded speech is useful in aircraft, mobile, military and commercial communication, and in aids for the handicapped. The objectives of speech enhancement vary widely: reduction of noise level, increased intelligibility, reduction of auditory fatigue, etc. [1]. The speech enhancement algorithm that is introduced in this paper works in the spectral domain. Before explaining the proposed method, we review how the magnitude spectrum of a voiced speech segment is formed.

The Fourier transform of a voiced speech signal is the Fourier transform of the glottal excitation X(ω) multiplied by the Fourier transform of the vocal tract filter V(ω). By windowing the time-domain speech signal, in the frequency domain, the Fourier transform of the window function W(ω) gets convolved with the Fourier transform of the speech signal. So the overall frequency representation of the STFT speech segment S(ω) can be written as follows:

\[ S(\omega) = (X(\omega)V(\omega)) * W(\omega). \]  (1)

Assuming the glottal excitation to be an ideal impulse train in the time domain and assuming the window length to be big enough (e.g., more than two times the fundamental period), we can rewrite Equation 1 for the magnitude spectra as:

\[ |S(\omega)| = |X(\omega)||V(\omega)|*|W(\omega)|. \]  (2)

Since we assumed the excitation to be an ideal impulse train in the time domain, its magnitude spectrum also becomes an impulse train in the frequency domain |X(\omega)|. The multiplication of |X(\omega)| by the vocal tract filter |V(\omega)| scales impulses in |X(\omega)| according to |V(\omega)|. So basically, |X(\omega)||V(\omega)| is still an impulse train, but with scaled impulses. Thus, the convolution of |X(\omega)||V(\omega)| and |W(\omega)| results in shifted scaled instances of |W(\omega)| at the locations of these impulses. These shifted scaled instances of |W(\omega)| are called “harmonics”. As a result, assuming that we have the ideal excitation, if we have the places of harmonics in the magnitude spectrum (i.e., by having an estimate of the fundamental frequency), by adding up shifted scaled versions of the magnitude spectrum of the window function that we use for STFT, we can reconstruct the speech segment magnitude spectrum.

In [2] it is statistically shown that the magnitude and phase parts of the STFT of the speech segment show a form of independence. This is the assumption we have in our method. Our enhanced STFT of the speech segment is the aforementioned reconstructed magnitude spectrum combined with the phase of the noisy speech. In [3], it is shown that the use of the phase spectrum of the noisy speech as the estimate for the phase of the clean speech minimizes the enhancement error while in [4], it is shown that improving this phase estimate does not lead to a noticeable improvement in speech enhancement.

It should be mentioned that our proposed method has similarities with the concept of using comb filtering in speech enhancement [5,6]. The principle of that approach is enhancing the speech magnitude spectrum by improving the harmonics at multiples of the fundamental frequency and suppressing the noise for the intervals between the harmonics. Both the comb filtering method and our method exploit the quasi-periodicity characteristic of speech. However, the main idea of our approach consists of reconstructing the harmonics of the magnitude spectrum rather than improving them or suppressing the noise in between them.

2. Proposed method

The core idea of the proposed method is laying shifted scaled instances of the magnitude spectrum of the window function (the harmonics) on multiples of F₀. To do this, we need to know F₀, the shape of the harmonics, and the scaling factor for each of the harmonics.

2.1. Model parameters

We first need to have an estimate of the fundamental frequency. For this purpose, the autocorrelation (AC) pitch estimation method is used. The autocorrelation method is chosen because it is known to be one of the best pitch estimators under noisy environments [7]. By having F₀, we know the center frequencies of the harmonics in the magnitude spectrum since they are multiples of F₀.
Figure 1: a) Magnitude spectrum of a real clean speech segment with $F_0 = 183$ Hz for 0 to 1 kHz windowed by a 20 msec Bartlett-Hann window, b) magnitude spectrum of the same segment with the same window under the SNR of 0 dB Additive White Gaussian Noise (AWGN), c) the reconstructed magnitude spectrum using the proposed method, d, e, and f) same as a, b, and c, except that the window function in these three magnitude spectra is a 50 msec Hanning window. The additional lines in b, c, e, and f show the magnitudes at the multiples of $F_0$ (harmonic locations). These magnitudes represent the harmonic scaling factors.

For the shape of the harmonics the magnitude spectrum of the window function is used. We denote the magnitude spectrum of the harmonics by $H(\omega)$. Since we are going to scale the harmonics later, for the moment, we normalize the maximum of the harmonics to 1. Thus, we have:

$$|H(\omega)| = |W(\omega)| / \max|W(\omega)|.$$  \hfill (3)

We get the scaling factors of the harmonics from the magnitude spectrum of the noisy segment. Scaling factors are the magnitudes corresponding to the multiples of $F_0$ (locations of the harmonic peaks) in the magnitude spectrum of the noisy speech segment.

Now that we have all of the parameters for our model, we can produce the reconstructed magnitude spectrum. The process is straightforward: We shift the harmonics on multiples of the fundamental frequency, scale them according to the scaling factors and add them up. In Figure 1 the process is demonstrated.

An important issue that can be seen in Figure 1 is that the shape of the harmonics of the reconstructed speech magnitude spectrum is a function of the window function that is being used. As a result, this method automatically adapts itself to different windows (different in type or length). This is why the bandwidth of the harmonics in Figure 1-c and Figure 1-f are different. Another interesting point is that using this method we come up with a reconstructed magnitude spectrum, which is very close to the clean speech magnitude spectrum, by only using a few scaling factors and the magnitude spectrum of the window function.

The underlying assumption of our method is that the glottal excitation is an ideal impulse train, which is not the case in real speech. Speech is quasi-periodic rather than ideally periodic. As a result, the pitch period in our segments is not constant, especially for lengthy windows. This can pose two problems for our method:

- Bandwidth of the harmonics can change slightly because of the varying pitch [5].
- Periodicity in high frequencies (more than 3 kHz) degrades. This means that the locations of the harmonics change for high frequencies [5].

These two phenomena may slightly affect our enhancement but experimental results will prove the effectiveness of our method for our chosen window length (30 msec) despite this approximation.

2.2. Implementation

Figure 2 shows the block diagram of the proposed method. The first component is the pitch estimation module. For this purpose the time-domain autocorrelation method is used. The autocorrelation method is chosen because of its genuine performance in low SNR values [7]. The output of this module ($F_0$) will be used to localize the harmonics (on multiples of $F_0$) for calculating the scaling factors and for shifting the reconstructed harmonics in the coming modules.

After calculating the Fast Fourier Transform of the windowed speech segment, the magnitude and phase parts are separated. The magnitude part is used to extract the scaling factors and the phase part is combined with the reconstructed magnitude spectrum.

The Scaling Factors Extractor module, using the fundamental frequency, picks the magnitudes corresponding to the multiples of the fundamental frequency in the magnitude spectrum of the noisy speech. The resulting values are the harmonic scaling factors. Since we are using the FFT,
which is discrete in the frequency domain, interpolation is performed to achieve the scaling factors:

$$\lambda(k) = \left| S(k \frac{2\pi f_k}{f_s}) \right| \quad k = 0, 1, 2, 3, \ldots, k < \frac{f_s}{F_0} \quad (4)$$

In Equation 4 $\lambda(k)$ is the $k$th scaling factor which corresponds to the $k$th harmonic, $S(\omega)$ is the magnitude spectrum of the windowed noisy segment, $F_0$ is the fundamental frequency (Hz), and $f_s$ is the sampling frequency (Hz).

The Harmonic Scaler, Shifter, and Adder module is where the magnitude spectrum is reconstructed. The shape of the harmonics $[H(\omega)]$ is the magnitude spectrum of the window function. The harmonics are scaled according to the scaling factors and shifted to multiples of the fundamental frequency. These steps are done in the frequency domain as follows:

$$R(\omega) = \sum_k \lambda(k) \left| H(\omega - k \frac{2\pi F_0}{f_s}) \right| \quad k = 0, 1, 2, 3, \ldots, k < \frac{f_s}{F_0} \quad (5)$$

$R(\omega)$ is the reconstructed magnitude spectrum, and $[H(\omega)]$ is defined in Equation 3. Considering the time domain, we define $h[n]$ as:

$$h[n] = \text{IFFT} \left[ [H(\omega)] \right]$$

$h[n]$ is basically a scaled version of the window function $w[n]$. The frequency shifting characteristic of the Fourier transform states that:

$$h[n]e^{-j\omega_0 n} \xrightarrow{\text{FFT}} H(\omega - \omega_0) \quad (7)$$

Now using this characteristic and the linearity characteristic of the FFT we change Equation 5 to:

$$R(\omega) = \left| \text{FFT} \left( \sum_k \lambda(k) h[n] e^{j\frac{2\pi k}{F_0}} \right) \right|$$

$$k = 0, 1, 2, 3, \ldots, k < \frac{f_s}{F_0} \quad (8)$$

Implementing Equation 8 is much easier than Equation 5 and also results in much less computational load for this method.

Now that we have the reconstructed magnitude spectrum, we combine it with the phase of the noisy speech segment and finally perform an inverse FFT to extract the enhanced time-domain segment.

The last point for this section is that this design enhances voiced speech segments. In order to enable it to also accept unvoiced speech segments as input, we only need to add a Voiced/Unvoiced Detector (VUD) block prior to the pitch estimation module to track the unvoiced segments and keep them from being enhanced and simply leave them unchanged. The same idea was used in [5].

### 3. Experiments

#### 3.1. Experimental details

For assessing the capability of the proposed method in enhancing speech, a series of experiments were carried out. Voiced samples from both male and female speakers were selected from the DARPA TIMIT acoustic-phonetic Continuous Speech Corpus database. The sampling frequency was 16000 Hz. Speech samples were windowed using a 30 msec Hanning window and the time between frames was 20 msec. The continuous samples were enhanced at the overall SNR values of 10, 5, 0, and -5 dB with Babble, Factory, and Additive White Gaussian Noise (AWGN) patterns. As a result, a total of 9615 speech segments was treated. It should be mentioned that the SNR values mentioned above are the overall SNR values of the samples and as a result, the SNR values of the component segments of each sample had deviations from these values. Noise was not added segment by segment in order for the speech samples not to lose their continuity; otherwise, the process of pitch estimation would be influenced by these discontinuities.

#### 3.2. Experimental results

After the enhancement process, the Segmental SNR values (SSNR) of the noisy and enhanced speech segments were calculated. After having all of the SSNR values of the noisy
Figure 3: a) The Segmental SNR of the enhanced speech versus the Segmental SNR of the noisy input speech for AWGN, Babble Noise, and Factory Noise. b) the Segmental SNR enhancement gains for AWGN, Babble Noise, and Factory Noise.

and enhanced speech, polynomials of the order 4 that best fitted the data were calculated. In Figure 3-a, using these polynomials, the SSNR of the enhanced speech is plotted for different SSNR values of input noisy speech for AWGN, Babble, and Factory Noise patterns. In Figure 3-b, we can find the corresponding enhancement gains.

4. Discussion

The results show that the proposed method is efficient in enhancing noisy speech segments of various noise types. An issue left for discussion is the effect of pitch estimation errors on the enhancement process. The Gross Pitch Estimation Errors (GPEs) from the pitch estimation module would lead to the wrong identification of harmonic locations and as a result, the enhancement procedure would be degraded. However, since a substantial portion of the Gross Pitch Estimation Errors are doubling and halving errors, we may still be able to reconstruct a good amount of the harmonics and may even end up having an improved Segmental SNR eventually, despite the GPEs. Consequently, although the GPEs lead to a less effective enhancement, their overall effect on the SSNR proves not to be catastrophic.

A prominent point about this enhancement method is that in this method since instead of suppressing the noise we basically reconstruct the harmonics, there will be no residual noise in the magnitude spectrum. Musical noise is mainly because of noise residuals in the de-noised signal spectrum [8]. As a result, musical noise is minimal for this method. This can easily be realized by listening to the enhanced speech.

Another feature of this method is its characteristic of auto-adaptability to different windows (in shape or size). This is because this method automatically adapts the harmonic shapes to the magnitude spectrum of the window function (see Figure 1). As a result, while enhancing different segments from different window functions, we do not need to change any fundamental parameters or do any sort of re-training of the proposed enhancement method.

5. Conclusions

A new speech enhancement algorithm is introduced. It enhances speech segments by basically reconstructing the harmonics at multiples of the fundamental frequency. In order to locate the places of the harmonics on the frequency axis, the method uses the autocorrelation pitch estimation method. The effectiveness of the method for different noise types was experimentally evaluated using the objective Segmental SNR test and it was shown that this method is quite effective in enhancing noisy speech segments. The auto-adaptability to window function character and the minimal musical noise in the enhanced speech of this method are two of its prominent features. These advantages are confirmed by listening to the enhanced speech. We are continuing our efforts by assessing the performance of this technique in speech recognition configurations in adverse conditions.

6. References