Modulation Domain Spectral Subtraction for Speech Enhancement

Kuldip Paliwal, Belinda Schwerin, Kamil Wójcicki

Signal Processing Laboratory, Griffith University, Queensland, Australia
{k.paliwal, b.schwerin, k.wojcickij}@griffith.edu.au

Abstract

In this paper we investigate the modulation domain as an alternative to the acoustic domain for speech enhancement. More specifically, we wish to determine how competitive the modulation domain is for spectral subtraction as compared to the acoustic domain. For this purpose, we extend the traditional analysis-modification-synthesis framework to include modulation domain processing. We then compensate the noisy modulation spectrum for additive noise distortion by applying the spectral subtraction algorithm in the modulation domain. Using subjective listening tests and objective speech quality evaluation we show that the proposed method results in improved speech quality. Furthermore, applying spectral subtraction in the modulation domain does not introduce the musical noise artifacts that are typically present after acoustic domain spectral subtraction. The proposed method also achieves better background noise reduction than the MMSE method.

Index Terms: speech enhancement, spectral subtraction, modulation domain, analysis-modification-synthesis (AMS)

1. Introduction

In the field of speech enhancement, we are interested in the reduction of noise from noise-corrupted speech in order to improve its intelligibility and quality. Many popular speech enhancement methods employ the analysis-modification-synthesis (AMS) framework to perform enhancement in the acoustic spectral domain [1]. In this paper, we investigate speech enhancement in the modulation spectral domain by extending the acoustic AMS framework to include modulation domain processing. We differentiate the acoustic spectrum from the modulation spectrum as follows. The acoustic spectrum is the short-time Fourier transform (STFT) of the speech signal, while the modulation spectrum at a given acoustic frequency is the STFT of the time series of the acoustic spectral magnitudes at that frequency.

Zadeh [2] was perhaps the first to propose a two-dimensional bi-frequency system, where the second dimension for frequency analysis was the transform of the time variation of the standard (acoustic) frequency. More recently, Atlas et al. [3] defined acoustic frequency as the axis of the first STFT of the input signal and modulation frequency as the independent variable of the second STFT transform. The short-time modulation spectrum is thus a function of time, acoustic frequency and modulation frequency.

There is growing psychoacoustic and physiological evidence to support the significance of the modulation domain for the analysis of speech signals. For example, neurons in the auditory cortex are thought to decompose the acoustic spectrum into spectro-temporal modulation content [4]. Drullman et al. [5, 6] investigated the importance of modulation frequencies for intelligibility by applying low-pass and high-pass filters to the temporal envelopes of acoustic frequency sub-bands. They showed frequencies between 4 and 16 Hz to be important for intelligibility, with the region around 4-5 Hz being the most significant. In a similar study, Arai et al. [7] showed that applying passband filters between 1 and 16 Hz does not impair speech intelligibility. While the envelope of the acoustic magnitude spectrum represents the shape of the vocal tract, the modulation spectrum represents how the vocal tract changes as a function of time. It is these temporal changes that convey most of the linguistic information (or intelligibility) of speech. In the above studies, the lower limit of 1 Hz stems from the fact that the slow vocal tract changes do not convey much linguistic information. In addition, the lower limit helps to make speech communication more robust, since the majority of noises occurring in nature vary slowly as a function of time and hence their modulation spectrum is dominated by modulation frequencies below 1 Hz. The upper limit of 16 Hz is due to the physiological limitation on how fast the vocal tract is able to change with time.

Modulation domain processing has grown in popularity finding applications in areas such as speech coding, speech and speaker recognition, objective speech intelligibility evaluation as well as speech enhancement. In the latter category, a number of modulation filtering methods have emerged. For example, Hermansky et al. [8] proposed to bandpass filter the time trajectories of cubic-root compressed short-time power spectrum for enhancement of speech corrupted by additive noise. More recently in [9, 10], similar bandpass filtering was applied to the time trajectories of the short-time power spectrum for speech enhancement.

There are two main limitations associated with typical modulation filtering methods. First, they use a filter design based on the long-term properties of the speech modulation spectrum while ignoring the properties of noise. As a consequence, they fail to eliminate noise components present within the speech modulation regions. Second, the modulation filter is fixed and applied to the entire signal, even though the properties of speech and noise change over time. In the proposed method, we attempt to address these limitations by processing the modulation spectrum on a frame-by-frame basis. In our approach, we compensate noisy speech for additive noise distortion by applying the spectral subtraction algorithm [11] in the modulation domain.

In this paper, we evaluate how competitive the modulation domain is for speech enhancement as compared to the acoustic domain. For this purpose, objective and subjective speech enhancement experiments are carried out. The results of these experiments demonstrate that the modulation domain is a useful alternative to the acoustic domain. Future work will investigate the use of other advanced enhancement techniques, such as MMSE, Kalman filtering, etc., in the modulation domain.

The rest of this paper is organized as follows. Section 2 details the traditional AMS-based speech processing. Section 3 presents details of the proposed modulation domain enhance-
ment method. Section 4 describes objective and subjective enhancement experiments. Experimental results along with a discussion are presented in Section 5. Conclusions are drawn in Section 6.

2. Acoustic analysis-modification-synthesis

The AMS framework consists of three stages: 1) the analysis stage, where the input speech is processed using the STFT analysis; 2) the modification stage, where the noisy spectrum undergoes some kind of modification; and 3) the synthesis stage, where the inverse STFT is followed by the overlap-add synthesis to reconstruct the output signal.

Let us consider an additive noise model

\[ x(n) = s(n) + d(n), \]

where \( x(n) \), \( s(n) \) and \( d(n) \) denote discrete-time signals of noisy speech, clean speech and noise, respectively. Since speech can be assumed to be quasi-stationary, it is analyzed framewise using the short-time Fourier analysis. The STFT of the corrupted speech signal \( x(n) \) is given by

\[ X(n, k) = \sum_{l=-\infty}^{\infty} x(l)w(n-l)e^{-j2\pi kl/N}, \]

where \( k \) refers to the index of the discrete acoustic frequency, \( N \) is the acoustic frame duration (in samples) and \( w(n) \) is an acoustic analysis window function. In speech processing, the Hamming window with 20–40 ms duration is typically employed. Using STFT analysis we can represent Eq. (1) as

\[ X(n, k) = S(n, k) + D(n, k), \]

where \( X(n, k), S(n, k), \) and \( D(n, k) \) are the STFTs of noisy speech, clean speech, and noise, respectively. Each of these can be expressed in terms of acoustic magnitude spectrum and acoustic phase spectrum. For instance, the STFT of the noisy speech signal can be written in polar form as

\[ X(n, k) = |X(n, k)|e^{j\angle X(n, k)}, \]

where \( |X(n, k)| \) denotes the acoustic magnitude spectrum and \( \angle X(n, k) \) denotes the acoustic phase spectrum.

Traditional AMS-based speech enhancement methods modify, or enhance, only the noisy acoustic magnitude spectrum while keeping the noisy acoustic phase spectrum unchanged. The reason for this is that the phase spectrum, at small window durations (20–40 ms), is considered unimportant for speech enhancement. Such algorithms attempt to estimate the magnitude spectrum of the clean speech. Let us denote the enhanced magnitude spectrum as \( \hat{S}(n, k) \), then the modified acoustic spectrum is constructed by combining \( \hat{S}(n, k) \) with the noisy phase spectrum, as follows

\[ Y(n, k) = |\hat{S}(n, k)|e^{j\angle X(n, k)}. \]

The enhanced speech, \( y(n) \), is reconstructed by taking the inverse STFT of the modified acoustic spectrum followed by synthesis windowing and overlap-add reconstruction [12].

3. Modulation

AMS-based spectral subtraction

The proposed speech enhancement method extends the traditional AMS-based acoustic domain enhancement to the modulation domain. To achieve this, each frequency component of the acoustic magnitude spectra, obtained using the AMS procedure outlined in Section 2, is processed framewise across time using a second AMS framework. Thus the modulation spectrum is computed using a second STFT as follows

\[ X(n, k, m) = \sum_{l=-\infty}^{\infty} |X(l, k)|e^{-j2\pi ml/M}, \]

where \( m \) refers to the index of the discrete modulation frequency, \( M \) is the modulation frame duration (in terms of acoustic frames) and \( v(n) \) is a modulation analysis window function. The resulting spectra can be expressed in polar form as follows

\[ X(n, k, m) = |X(n, k, m)|e^{j\angle X(n, k, m)}, \]

where \( |X(n, k, m)| \) is the modulation magnitude spectrum and \( \angle X(n, k, m) \) is the modulation phase spectrum.

We propose to replace \( |X(n, k, m)| \) with \( |\hat{S}(n, k, m)| \), where \( |\hat{S}(n, k, m)| \) is an estimate of clean modulation magnitude spectrum obtained using a spectral subtraction rule similar to the one proposed by Berouti et al. [11] and given by Eq. 8. In Eq. 8, \( \alpha \) denotes the subtraction factor that governs the amount of over-subtraction; \( \beta \) is the spectral floor parameter used to set spectral magnitude values falling below the spectral floor \( (\beta|\hat{D}(n, k, m)|) \) to that spectral floor; and \( \gamma \) determines the subtraction domain, e.g. for \( \gamma \) set to unity the subtraction is performed in the magnitude spectral domain, while for \( \gamma = 2 \), the subtraction is performed in the power spectral domain. The estimate of the modulation magnitude spectrum of the noise, denoted by \( |\hat{D}(n, k, m)| \), is obtained based on a decision from a simple voice activity detector (VAD) [1] applied in the modulation domain.

The modified modulation spectrum is produced by combining \( |\hat{S}(n, k, m)| \) with the noisy modulation phase spectrum as follows

\[ \hat{Z}(n, k, m) = |\hat{S}(n, k, m)|e^{j\angle X(n, k, m)}. \]

Note that unlike the acoustic phase spectrum, the modulation phase spectrum does contain useful information [8]. In the present work, we keep \( \angle X(n, k, m) \) unchanged, however future work will investigate approaches that can be used to enhance it. In the present study, we obtain the estimate of the modified acoustic magnitude spectrum \( |\hat{S}(n, k)| \), by taking the inverse STFT of \( \hat{Z}(n, k, m) \) followed by overlap-add with synthesis windowing. A block diagram of the proposed approach is shown in Fig. 1.
4. Speech enhancement experiments

4.1. Speech corpus

In our experiments we employ the Noizeus speech corpus [1]. Noizeus is composed of 30 phonetically-balanced sentences belonging to six speakers, three males and three females. The corpus is sampled at 8 kHz and filtered to simulate receiving frequency characteristics of telephone handsets. Noizeus comes with non-stationary noises at different SNRs. For our experiments we keep the clean part of the corpus and generate noisy stimuli by degrading the clean stimuli with additive white Gaussian noise (AWGN) at various SNRs. The noisy stimuli are constructed such that they begin with a noise only section long enough for noise estimation in both acoustic and modulation domains (approx. 500 ms).

4.2. Stimuli types

Modulation domain spectral subtraction stimuli (ModSpecSub) were constructed using the proposed speech enhancement method detailed in Section 3. The acoustic frame duration was set to 32 ms, with an 8 ms frame shift and the modulation frame duration was set to 256 ms, with a 32 ms frame shift. The Hamming window was used for both the acoustic and modulation analysis windows. The FFT-analysis length was set to \(2N\) and \(2M\) for the acoustic and modulation AMS frameworks, respectively. The value of the subtraction parameter \(\alpha\) was selected as described in [11]. The spectral floor parameter \(\beta\) was set to 0.002. Power spectral subtraction was used in the modulation domain, i.e. \(\gamma = 2\). Griffith and Lim’s method for windowed overlap-add synthesis [12] was used for both acoustic and modulation syntheses.

For our experiments we have also generated stimuli using two popular speech enhancement methods, namely the acoustic spectral subtraction (SpecSub) [11] and the MMSE method [13]. Clean and noisy stimuli were also included. Example spectrograms of the above stimuli are shown in Fig. 2.

4.3. Objective experiment

The objective experiment was carried out over the entire Noizeus corpus for AWGN at 0, 5, 10 and 15 dB SNR. Perceptual evaluation of speech quality (PESQ) measure was employed to predict mean opinion scores for stimuli types outlined in Section 4.2.

4.4. Subjective experiment

The subjective evaluation was in a form of AB listening tests that determine subjective method preference. Two Noizeus sentences, belonging to male and female speakers, were included. AWGN at 5 dB SNR was investigated. All five treatment types detailed in Section 4.2 were included. Fourteen English speaking listeners participated in this experiment. The listening tests were conducted in a quiet room. Total of 40 stimuli pairs were played back to the participants. The listeners were asked to make a subjective preference for each stimuli pair. The audio stimuli were presented in randomized order over closed circumaural headphones at a comfortable listening level.

5. Results and discussion

The results of the objective experiment, in terms of mean PESQ scores, are shown in Fig. 3. The proposed method performs consistently well across the SNR range, with particular improvements shown for stimuli with lower input SNR. MMSE showed the next best performance, with all enhancement methods achieving comparable results at 15 dB SNR.

The results of the subjective experiment, along with the standard error bars, are shown in Fig. 4. The subjective results are in terms of average preference scores. A score of one for a particular stimuli type, indicates that the stimuli type was always preferred. Conversely, a score of minus one means that the stimuli type was never preferred. Subjective results show that the clean stimuli were always preferred, while the noisy stimuli were the least preferred. Of the enhancement methods tested, ModSpecSub was preferred over MMSE, with SpecSub being the least preferred. Notably, the subjective results are consistent with the objective results.

The above results can be explained as follows. The acoustic spectral subtraction introduces spurious peaks scattered throughout the non-speech regions of the acoustic magnitude spectrum. At a given acoustic frequency bin, these spectral magnitude values vary over time (i.e. from frame to frame) causing audibly annoying sounds referred to as the musical noise. This is clearly visible in the SpecSub spectrogram of Fig. 2(c). On the other hand, the proposed method subtracts the modulation magnitude spectrum estimate of the noise from the modulation magnitude spectrum of the noisy speech along each acoustic frequency bin. While some spectral magnitude variation is still present in the resulting acoustic spectrum, the
residual peaks have much smaller magnitude. As a result, ModSpecSub stimuli do not suffer from the musical noise audible in SpecSub stimuli. This can be seen by comparing spectrograms in Fig. 2(c) and Fig. 2(d). While the MMSE method does not suffer from the problem of musical noise [1], it does not achieve as good noise suppression as the proposed method. This can be seen by comparing spectrograms in Fig. 2(d) and Fig. 2(e).

We would like to emphasize that the phase spectrum plays a more important role in the modulation domain than in the acoustic domain [8]. While in this preliminary study we keep the noisy modulation phase spectrum unchanged, in future work further improvements may be possible by also processing the modulation phase domain.

6. Conclusions

In this paper, we proposed to compensate noisy speech for additive noise distortion by applying the spectral subtraction algorithm in the modulation domain. To evaluate the proposed approach, both objective and subjective speech enhancement experiments were carried out. The results of these experiments demonstrate that the modulation domain processing is a useful alternative to acoustic domain processing for the enhancement of noisy speech.

7. References