1 Lecture Outline

Reading: Chapter 13

Some material in this lecture is from *Speech Enhancment Theory and Practice* by P. Loizou, Chapter 5.

- Introduction
- Basic Principles of Spectral Subtraction
- Filtering View of Spectral Subtraction
- Shortcomings of Spectral Subtraction
- Spectral Subtraction using Oversubtraction
- Performance of Spectral Subtraction

2 Introduction

Speech enhancement is concerned with improving quality and intelligibility of speech degraded by additive noise. Most speech enhancement algorithms improve only quality. It is possible to reduce background noise but often, algorithms introduce speech distortion, which in turn may impair intelligibility. The main challenge is to suppress noise without introducing perceptual distortion in the signal.

Speech enhancement algorithms can be divided into three main classes:

1. Spectral Subtraction: Estimate the noise spectrum and subtract it from the noisy speech spectrum
2. Statistical (Wiener filter): Compute a statistical estimator of the clean speech signal
3. Subspace: Decompose noisy speech space into speech subspace and noise subspace; estimate the clean signal by nulling the component of the noisy speech residing in the noise subspace

3 Basic Principles of Spectral Subtraction

The basic problem that we are trying to deal with in spectral subtraction is that of having a noisy speech signal $y[n]$ that is composed of a clean speech signal $x[n]$, corrupted by an additive noise source $d[n]$:

$$y[n] = x[n] + d[n]. \tag{1}$$

When we take the Fourier transform of both sides we get

$$Y(\omega) = X(\omega) + D(\omega) \tag{2}$$

We can rewrite these in polar form as

$$Y(\omega) = |Y(\omega)|e^{j\phi_y(\omega)}, \quad X(\omega) = |X(\omega)|e^{j\phi_x(\omega)}, \quad D(\omega) = |D(\omega)|e^{j\phi_d(\omega)} \tag{3}$$

so that

$$|Y(\omega)|e^{j\phi_y(\omega)} = |X(\omega)|e^{j\phi_x(\omega)} + |D(\omega)|e^{j\phi_d(\omega)} \tag{4}$$

In the usual speech enhancement problem, we are given $Y(\omega)$ and do not have direct access to $X(\omega)$. We may have limited access to $D(\omega)$ during non-speech segments.

Normally, the phase spectrum has only limited impact on speech quality. Therefore, we assume

$$\phi_x(\omega) \approx \phi_y(\omega). \tag{5}$$

From this assumption, we can estimate the spectrum of the speech spectrum as

$$\begin{aligned} |Y(\omega)|e^{\,j\phi_y(\omega)} &\approx \underbrace{|X(\omega)|e^{\,j\phi_y(\omega)}}_{\hat{X}(\omega)} + \underbrace{|D(\omega)|e^{\,j\phi_y(\omega)}}_{\hat{D}(\omega)} \\ &\downarrow \\ \hat{X}(\omega) &= \left[|Y(\omega)| - |\hat{D}(\omega)|\right] e^{\,j\phi_y(\omega)} \end{aligned} \tag{6}$$

where the noisy phase is given by

$$\phi_y(\omega) = \arg Y(\omega)/|Y(\omega)|. \tag{7}$$

We obtain the estimate of $D(\omega)$ by averaging during nonspeech (silence) segments. This requires a Voice Activity Detector (VAD). From $\hat{X}(\omega)$, we use an overlap-add method or least-squares estimate to compute $\hat{x}[n]$.

Note that we have used the phase of the noisy speech spectrum, $\phi_y(\omega)$ as the phase estimate for the clean speech because we do not know $\phi_x(\omega)$. Equation (6) is the basic form from which all spectral subtraction algorithms are derived.

3.1 Spectral Subtraction

It is possible that the estimated noise magnitude spectrum, $|\hat{D}(\omega)|$ may be greater than the input spectrum, $|Y(\omega)|$ which would lead to a nonsensical negative magnitude spectrum if (6) is carried out. One solution to deal with this is to apply a half-wave rectifier in the frequency domain

$$|\hat{X}(\omega)| = \begin{cases} |Y(\omega)| - |\hat{D}(\omega)|, & |Y(\omega)| > |\hat{D}(\omega)| \\ 0, & \text{otherwise.} \end{cases} \tag{8}$$

3.2 Power Spectrum Subtraction

Rather than working with the magnitude spectrum we can also use the power spectrum which, in some cases, performs better magnitude-based spectral subtraction. If we multiply (2) by its conjugate $Y^*(\omega)$ we have

$$\begin{aligned} |Y(\omega)|^2 &= |X(\omega)|^2 + |D(\omega)|^2 + X(\omega)D^*(\omega) + X^*(\omega)D(\omega) \\ &= |X(\omega)|^2 + |D(\omega)|^2 + 2Re\{X(\omega)D(\omega)\} \end{aligned} \tag{9}$$

We can approximate the cross terms with their expectation,

$$\begin{aligned} X^*(\omega)D(\omega) &\approx E\left\{X^*(\omega)D(\omega)\right\} \\ X(\omega)D^*(\omega) &\approx E\left\{X(\omega)D^*(\omega)\right\}. \end{aligned} \tag{10}$$

If we assume that the noise is uncorrelated with the speech, then the above expectations are zero then (9) becomes

$$|Y(\omega)|^2 = |\hat{X}(\omega)|^2 + |D(\omega)|^2. \tag{11}$$

The clean speech power spectrum is then estimated as

$$|\hat{X}(\omega)|^2 = |Y(\omega)|^2 - |\hat{D}(\omega)|^2 \tag{12}$$

and the clean speech spectrum is formed as

$$\hat{X}(\omega) = |\hat{X}(\omega)|e^{j\phi_y(\omega)}. \tag{13}$$

Half-wave rectification is also applied to ensure a non-negative power spectrum.

3.3 Generalized Spectral Subtraction

We can generalize the magnitude- and power-spectral subtraction algorithms as

$$|\hat{X}(\omega)|^p = |Y(\omega)|^p - |\hat{D}(\omega)|^p. \tag{14}$$

The estimate of the clean speech spectrum is thus

$$\hat{X}(\omega) \quad = \quad |\hat{X}(\omega)|e^{j\phi_y(\omega)}. \tag{15}$$

Typically the noisy speech signal is processed on a frame-by-frame basis with 20-30 ms windows. The process performed on each frame is shown in Fig. 3

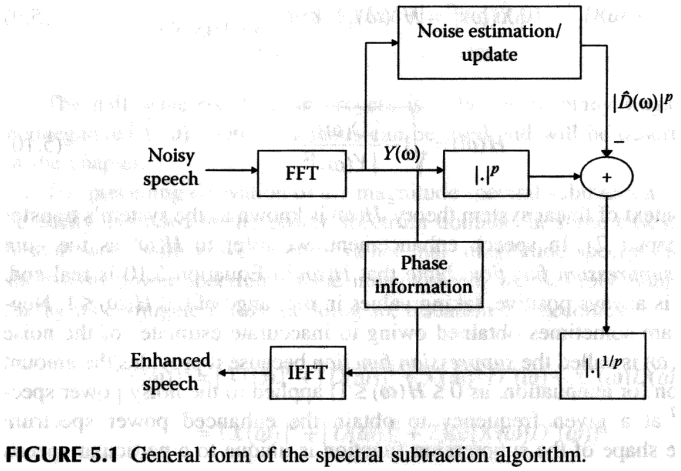

FIGURE 5.1 General form of the spectral subtraction algorithm.

Figure 1: Generalized Spectral Subtraction Block Diagram.

Demo: Audio demonstrations of spectral subtraction.

4 Spectral Subtraction as Non-Causal, Linear Filtering

We can view spectral subtraction as a filtering operation:

$$\begin{aligned} |\hat{X}(\omega)|^2 &= |Y(\omega)|^2 - |\hat{D}(\omega)|^2 \\ \underbrace{|\hat{X}(\omega)|^2}_{\text{output}} &= \underbrace{\left(1 - \frac{|\hat{D}(\omega)|^2}{|Y(\omega)|^2}\right)}_{\text{filter}} \underbrace{|Y(\omega)|^2}_{\text{input}} \end{aligned} \tag{16}$$

where filter or suppression function is

$$H(\omega) = \sqrt{1 - \frac{|\hat{D}(\omega)|^2}{|Y(\omega)|^2}} \tag{17}$$

5 Shortcomings of the Spectral Subtraction Method

The major shortcoming of spectral subtraction is presence of musical noise (distortion) in the enhanced speech. This musical noise results from half-wave rectification of the spectrum (to ensure a nonnegative clean

speech spectrum). The rectification results in small peaks in the spectrum occurring at random frequencies in each frame. In the time-domain, these peaks sound like tones whose frequency changes randomly from frame to frame

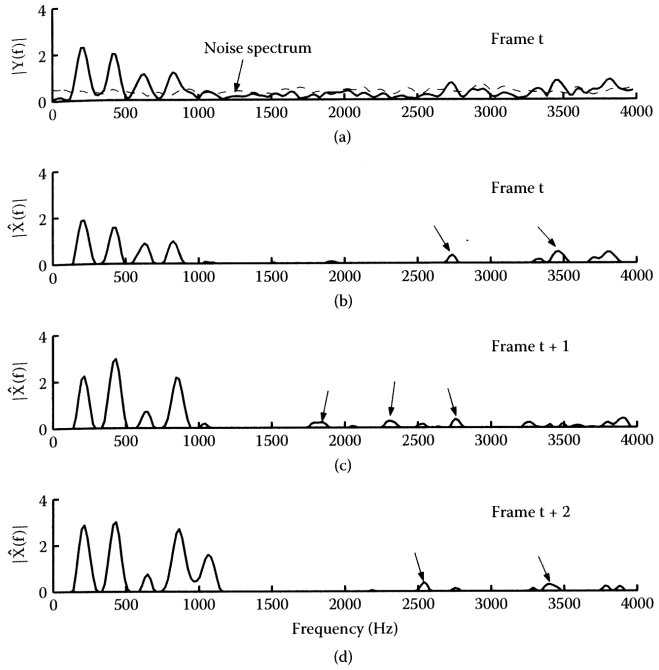

FIGURE 5.7 Example magnitude spectra of the enhanced signal following half-wave rectification in the subtraction process. Arrows show the isolated peaks responsible for musical noise.

Figure 2: Generalized Spectral Subtraction Block Diagram.

Demo: Musical noise in enhanced (spectral subtraction) speech.

Much of the research in the area of spectral subtraction has gone towards reducing musical noise. The major contributors to musical noise are

- Nonlinear processing of the negative subtracted spectral components, i.e. half-wave rectification
- Inaccurate estimate of noise estimate, we are typically forced to use an averaged estimate.
- Large variance in the estimates of the noisy and noise signal spectra (even if long windows are used)

Typically we must perform a trade-off between the amount of noise reduction and distortion in the resulting speech signal.

6 Spectral Subtraction Using Oversubtraction (Berouti et. al.)

One method that has been proposed to combat musical noise actually subtracts an overestimate of the noise power spectrum, while at the same time preventing the resulting spectral components from going below the noise floor. The basic algorithm by Berouti is given below.

$$|\hat{X}(\omega)|^2 = \begin{cases} |Y(\omega)|^2 - \alpha|\hat{D}(\omega)|^2, & |Y(\omega)|^2 > (\alpha+\beta)|\hat{D}(\omega)|^2 \\ \beta|\hat{D}(\omega)|^2, & \text{otherwise} \end{cases} \tag{18}$$

where α is the oversubtraction parameter and β is the spectral floor parameter. If we specify $\alpha = 1$ and $\beta = 1$ the oversubtraction algorithm reduces to the power spectral subtraction algorithm in (11).

When we specify $\alpha > 1$, this method tends to help reduce the presence of broadband noise which tends to have smaller peaks. By having $\beta > 0$, we also create a noise floor to help mask the remaining narrower peaks in the spectrum. It has been shown that Berouti's method (18) tends to have less musical noise than power spectral subtraction (11).

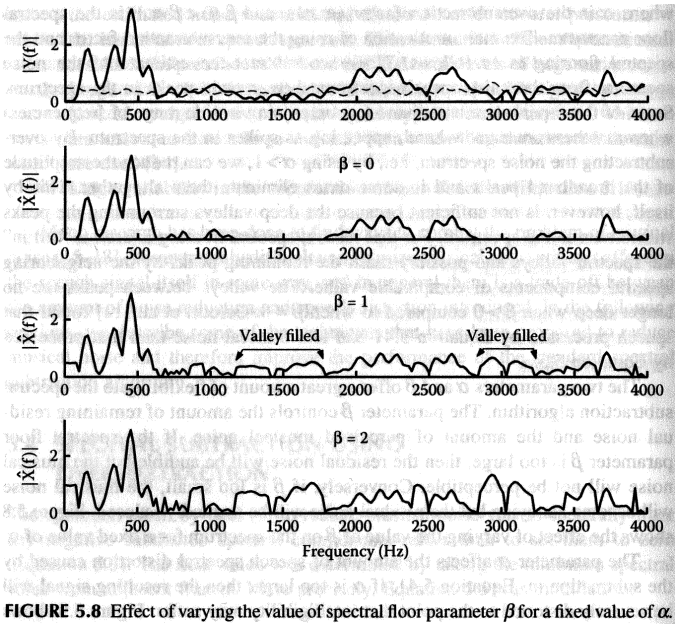

FIGURE 5.8 Effect of varying the value of spectral floor parameter β for a fixed value of α.

Figure 3: Generalized Spectral Subtraction Block Diagram.

Berouti suggests that the paramter α should vary from frame to frame according to

$$\alpha = \alpha_0 - \frac{3}{20}\text{SNR} \quad \text{-5 dB} \leq \text{SNR} \leq \text{20 dB} \tag{19}$$

Experiments by Berouti have shown that the "best" α_0 should be set in the range of 3 to 6 while the suggested value of β should be in the range $[0.02, 0.06]$ for high noise levels (SNR $\leq$ 0 dB) and $[0.005, 0.02]$ of low noise levels. Also, it is recommended that windows of length 25–35 ms should be used.

Eq. (18) can be rewritten the form

$$|\hat{X}(\omega)| = H(\omega)|Y(\omega)| \tag{20}$$

where $H(\omega)$ is a time-varying filter given by

$$H(\omega) = \left(\frac{|Y(\omega)|^2 - \alpha|\hat{D}(\omega)|^2}{|Y(\omega)|^2} \right)^{1/2}. \tag{21}$$

This can also be written as

$$H(\omega) = \left(\frac{\gamma(\omega) - \alpha}{\gamma(\omega)} \right)^{1/2} \tag{22}$$

where $\gamma(\omega)$ is the *a posteriori* SNR defined by

$$\gamma(\omega) = \frac{|Y(\omega)|^2}{|\hat{D}(\omega)|^2} \tag{23}$$

Clearly when $\gamma(\omega)$ (the SNR) is large, $H(\omega)$ will be close to one and have minimal impact on the spectrum. On the otherhand, when $\gamma(\omega)$ is small, $H(\omega)$ will tend towards zero and will attenuate the spectrum.

7 Performance of Spectral Subtraction

Most studies show spectral subtraction improves speech quality but not speech intelligibility. Intuitively, one would expect that by improving speech quality speech intelligibility would improve but this is not the case. Listeners prefer quality of speech that has minimal noise present without paying much attention to the fact that there might be missing speech information caused by spectral subtraction. Speech that is less noisy seems to be preferred in general by most listeners. Spectral subtraction does an excellent job in removing noise but at the expense of eliminating occasionally low-energy speech information, which explains the lack of improvement in speech intelligibility.