Speech Signal Processing
— Exercise 6 —
Speech Enhancement

Timo Gerkmann, Kristina Tesch, Danilo Oliveira

This exercise deals with single-channel speech enhancement techniques. Such methods usually apply the Fourier transform on short time-frames, which you already generated in the first and second exercise. Then, the noise components are suppressed in the spectral domain. The same holds for this exercise where the short-time spectra are used to estimate the power spectral density (PSD) of the background noise using a speech presence probability (SPP) based estimator. From this, the a posteriori signal-to-noise ratio (SNR) is determined with the decision-directed (DD) approach. Finally, both quantities are used to design the gain-function of the Wiener-filter which is applied to the noisy spectrum to achieve the desired noise reduction.

Download the file *Exercise6.zip* from *Moodle* which contains two speech signals which are corrupted once by white noise and once by babble noise. For both files the signal-to-noise ratio (SNR) is 10 dB.

1 Noise Power Estimation

In this part of the exercise, the noise power is estimated based on the speech presence probability. First, use your functions from the second exercise to create the STFT of the noisy input signals. The frame length should be 32 ms and the frame shift 16 ms. Use a $\sqrt{\text{Hann}}$-window as analysis window. From the STFT, compute the periodograms and determine the noise PSD by performing the following steps for each frame.

1. Compute the posterior probability of speech presence via

$$P(\mathcal{H}_1|Y[k,\ell]) = \left(1 + (1+\theta)\exp\left(-\frac{|Y[k,\ell]|^2}{\hat{\sigma}_n^2[k,\ell-1]}\frac{\theta}{1+\theta}\right)\right)^{-1}. \tag{1}$$

   Here, $|Y[k,\ell]|^2$ denotes the periodogram of the noisy input signal and $\hat{\sigma}_n^2[k,\ell]$ is the estimated PSD of the background noise. Furthermore, k indexes the frequency bins and ℓ the frames. The SNR between speech and noise which is assumed in speech presence is denoted by θ. This value should be set to 15 dB.

2. In order to avoid stagnations, a smoothed posterior probability $Q[k,\ell]$ is required which is given as

$$Q[k,\ell] = 0.9Q[k,\ell-1] + 0.1P(\mathcal{H}_1|Y[k,\ell])\,. \tag{2}$$

   For every frequency bin in the current frame ℓ for which $Q[k,\ell] > 0.99$ holds, set your computed posterior $P(\mathcal{H}_1|Y[k,\ell])$ from equation (1) to $\min(0.99, P(\mathcal{H}_1|Y[k,\ell]))$.

3. Estimate the noise periodogram via

$$|\hat{N}[k,\ell]|^2 = P(\mathcal{H}_1|Y[k,\ell])\,\hat{\sigma}_n^2[k,\ell-1] + \{1 - P(\mathcal{H}_1|Y[k,\ell])\}\,|Y[k,\ell]|^2 \tag{3}$$

and update the noise power density with

$$\hat{\sigma}_n^2[k,\ell] = 0.8\hat{\sigma}_n^2[k,\ell-1] + 0.2|\hat{N}[k,\ell]|^2. \tag{4}$$

Store the estimated noise PSD and the speech presence probability for each frame in a row of the matrices `m_noise_psd` and `m_spp` as it may simplify solving the following questions.

Questions

1.1 In equation (1), (3) and (4) the noise power estimate from the previous frame, $\hat{\sigma}_n^2[k,\ell-1]$, is required. However, there is none available for the first frame $\ell = 0$. The same problem occurs for the smoothed speech presence probability $Q[k,\ell]$ in (2).

- What would be appropriate initializations for the first noise estimate $\hat{\sigma}_n^2[k,-1]$ and the smoothed posterior probability $Q[k,-1]$? Explain briefly why you chose your initialization method.

1.2 Plot the speech presence probability $P(\mathcal{H}_1|Y[k,\ell])$ using `plt.imshow`.

a) Which values do you obtain for time-frequency points where speech is present?

b) What values do you get for time-frequency points where only noise is present?

c) If you compare the speech presence probability with the spectrogram of your input signal, can you see similarities?

1.3 Plot the estimated noise PSD as a spectrogram using `plt.imshow`.

a) How well is the background noise estimated?

b) Can you observe errors (e. g. components that do not belong to the background noise)?

c) What would be the consequence of such errors?

2 A priori SNR estimation and Wiener Filtering

In this part of the exercise, we will obtain an estimate of the a priori SNR $\hat{\xi}$ from our noise power estimate $\hat{\sigma}_n^2[k,\ell]$ using the decision-directed approach. It is given by

$$\hat{\xi}[k,\ell] = \alpha\frac{|\hat{S}[k,\ell-1]|^2}{\hat{\sigma}_n^2[k,\ell-1]} + (1-\alpha)\max\left(\frac{|Y[k,\ell]|^2}{\hat{\sigma}_n^2[k,\ell]} - 1, 0\right). \tag{5}$$

Here, $\hat{S}$ denotes the estimate of the clean speech spectrum S. The gain function G of the Wiener-Filter is

$$G[k,\ell] = \max\left(\frac{\hat{\xi}[k,\ell]}{1+\hat{\xi}[k,\ell]}, G_{\min}\right). \tag{6}$$

The enhanced spectra are finally obtained via

$$\hat{S}[k,\ell] = G[k,\ell]Y[k,\ell]. \tag{7}$$

Perform the steps in equations (5) – (7) for every frame of your input signal and store all enhanced speech spectra $\hat{S}[k,\ell]$ in a matrix, e. g. `m_enhanced_stft`.

Questions

2.1 In equation (5), the issue from Section 1 occurs again. An initialization for $\hat{S}[k,\ell]$ is required for the first frame.

- What would be a reasonable choice in this case?

2.2 Set $\alpha = 0.5$ and $G_{\min} = 0$. Plot the magnitude spectrogram of the noisy speech signal and the enhanced speech signal in dB and compare both. Make sure, that the color bar for both plots is the same. This can be achieved by manually setting the `vmin` and `vmax` parameters of `plt.imshow`.

a) How does the clean spectrogram differ from the spectrogram of the noisy input signal?

b) Can you see artifacts in the spectrogram of the enhanced signal?

3 Parameter tuning

Use the `compute_istft` function to synthesize the enhanced speech signal. For this, employ the same frame shift and FFT length as in Section 1. Further, use the $\sqrt{\text{Hann}}$-window also for the synthesis.

Listen to the noisy signal and the enhanced signal. For this you could use the `play` function from `sounddevice`.

Questions

3.1 Compare the noisy and the enhanced signal with each other.

a) How well is the background noise suppressed?

b) Can you hear any distortions of the speech signal in the enhanced signal?

c) What artifacts can you hear?

3.2 Vary α between 0 and 1 and listen to the synthesized signals.

a) What differences can you perceive?

b) How do the artifacts, speech signal and noise suppression change?

c) What is your favorite setting? Explain why.

d) Are the differences you heard also visible in the spectrogram of the enhanced speech signal?

3.3 Try different values for $G_{\min}$ which can be varied between 0 and 1, i. e. between $-\infty$ dB and 0 dB. Listen again to the synthesized signals.

a) What differences can be perceived now?

b) How does this parameter affect the artifacts, speech signal and noise suppression and what would be your favorite setting this time? Again, explain why.

c) How do the spectrograms change in this case?

In your report, include some spectrograms that support your reasoning for your choice of the parameters α and $G_{\min}$.