Musical Source Separation

Name: Rajesh

Roll no: EE21M019

Department : Electrical-SPCOM,Mtech 1st year

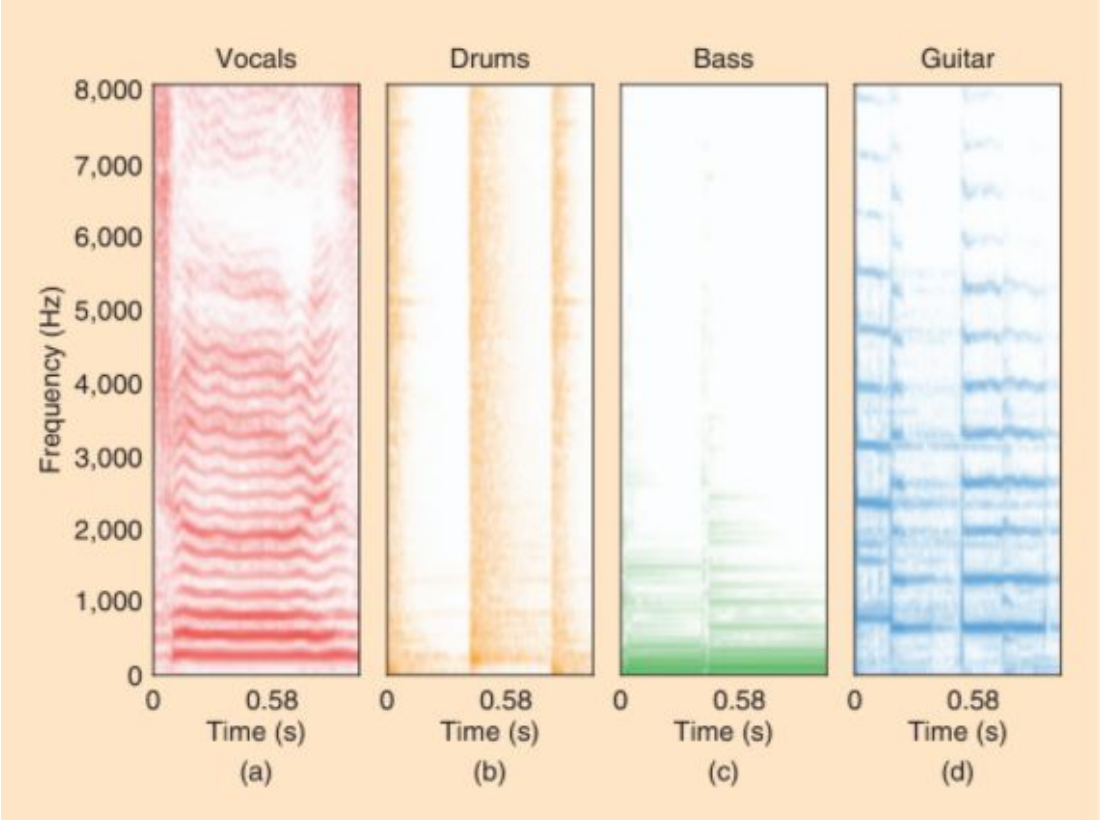
Guide : Dr. Rama Krishna Gorthi.

Summary

- Introduction
- Short time fourier transform - spectrogram
- Harmonic and percussive sounds
- Median filtering
- Masking
- Reconstruction
- References

INTRODUCTION

- Separate harmonics and precursive sources form a given musical signal.
- Every musical source has some characteristics,
- For example, violin has harmonic nature in frequency domain,whereas drums has percussive.
- Our voice also has such same nature
- Understand the spectrograms of different musical sources.



Source: musical source separation ,Estefanía Cano, Derry FitzGerald, Antoine Liutkus,Mark D. Plumbley, and Fabian-Robert Stöter

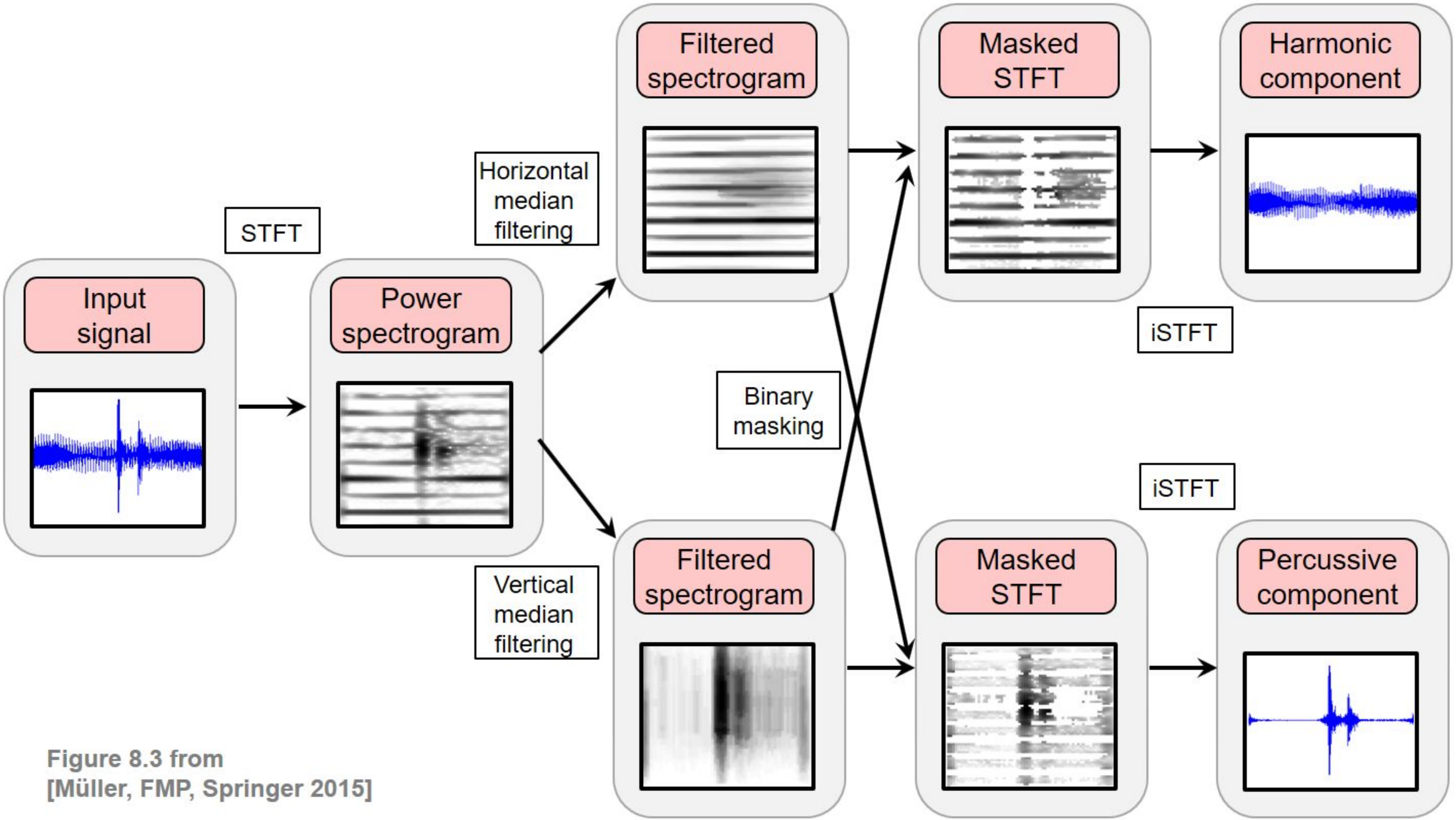


Figure 8.3 from
[Müller, FMP, Springer 2015]

Short time fourier transform- spectrogram

- Let x(n) is the signal, then STFT of x(n) is given by

$$\mathcal{X}(n,k) := \sum_{r=0}^{N-1} x(r+nH)w(r)\exp(-2\pi ikr/N),$$

- From X(n,k) we derive the (power) spectrogram Y

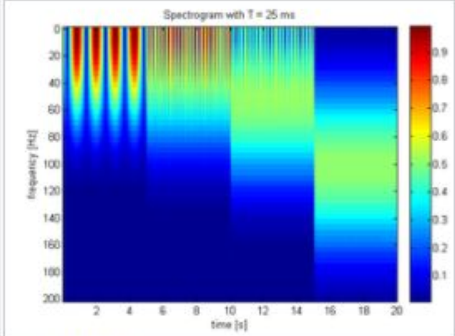
$$\mathcal{Y}(n,k) := |\mathcal{X}(n,k)|^2.$$

- A spectrogram can be understood as a mixture, in which each line represents a different frequency and each column represent a different instance of time.

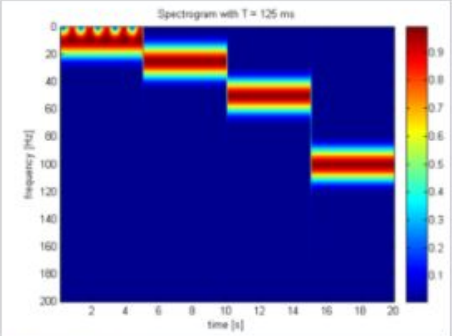
Example spectrogram

$$x(t) = \begin{cases} \cos(2\pi 10t) & 0\,\mathrm{s} \le t < 5\,\mathrm{s} \\ \cos(2\pi 25t) & 5\,\mathrm{s} \le t < 10\,\mathrm{s} \\ \cos(2\pi 50t) & 10\,\mathrm{s} \le t < 15\,\mathrm{s} \\ \cos(2\pi 100t) & 15\,\mathrm{s} \le t < 20\,\mathrm{s} \end{cases}$$

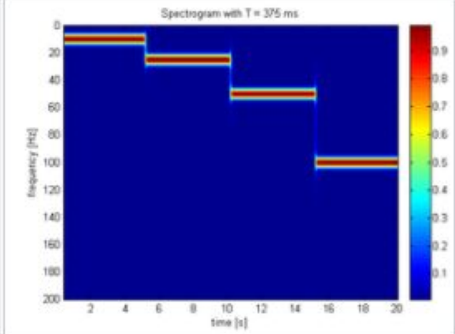
Then it is sampled at 400 Hz. The following spectrograms



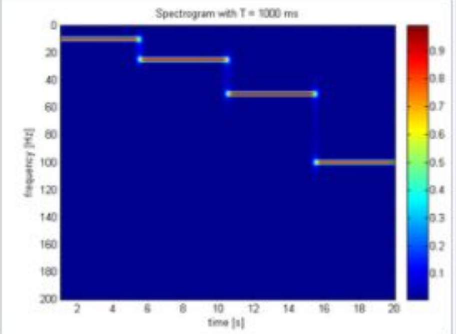
25 ms window



125 ms window

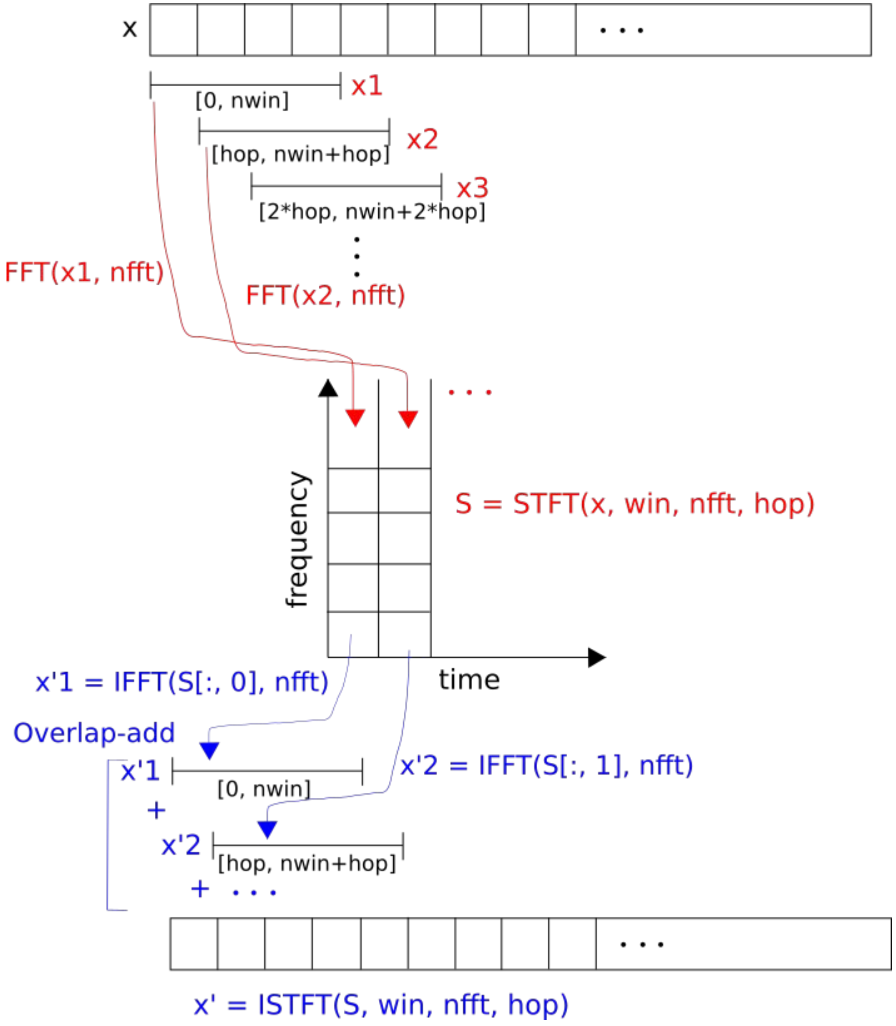


375 ms window



1000 ms window

Source: https://en.wikipedia.org/


x
. . .
x1
[0, nwin]
x2
[hop, nwin+hop]
x3
[2*hop, nwin+2*hop]
FFT(x1, nfft)
FFT(x2, nfft)
frequency
time
S = STFT(x, win, nfft, hop)
x'1 = IFFT(S[:, 0], nfft)
Overlap-add
x'1
[0, nwin]
+
x'2 = IFFT(S[:, 1], nfft)
x'2
[hop, nwin+hop]
+ . . .
x' = ISTFT(S, win, nfft, hop)

Harmonic and percussive sounds.

Harmonic

- Well defined in frequency
- Occurs in small frequency interval
- Containing lots of different instance of time.

Percussive

- Well defined in time
- Occurs in small time interval
- Containing lots of different frequencies.

Median

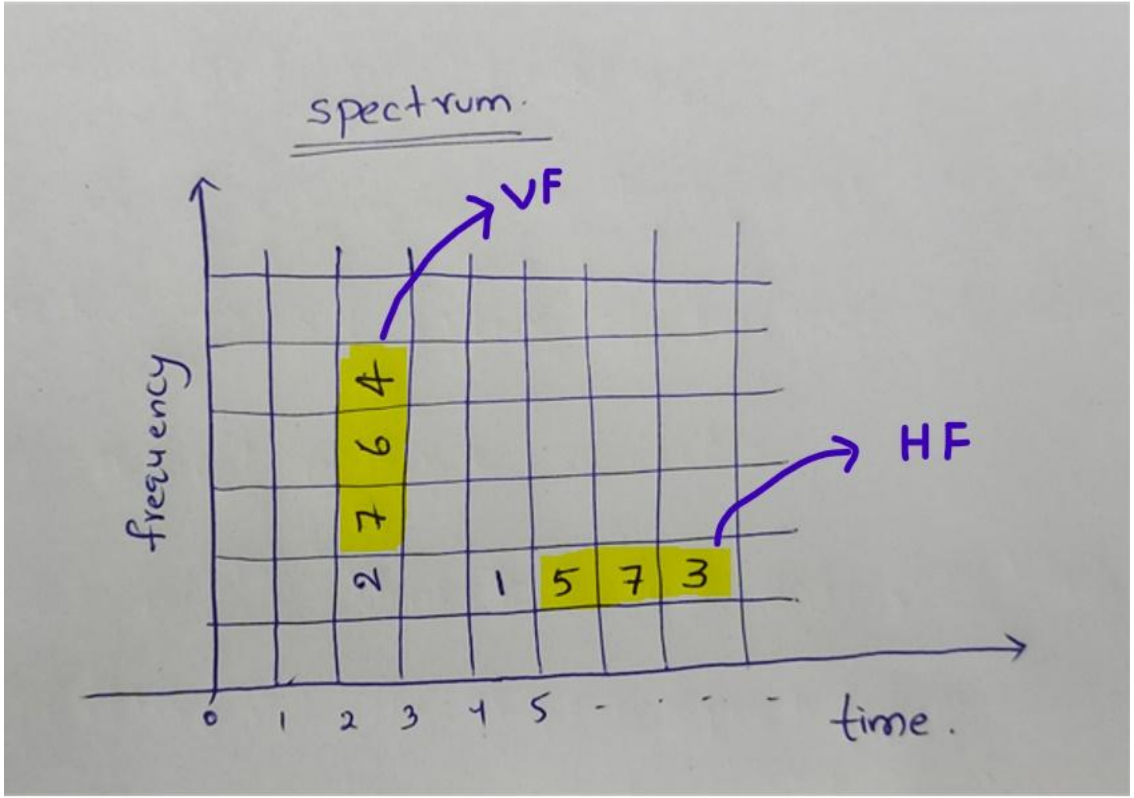
- The median is the value separating the higher values of the data samples from the lower values.
- For example ,

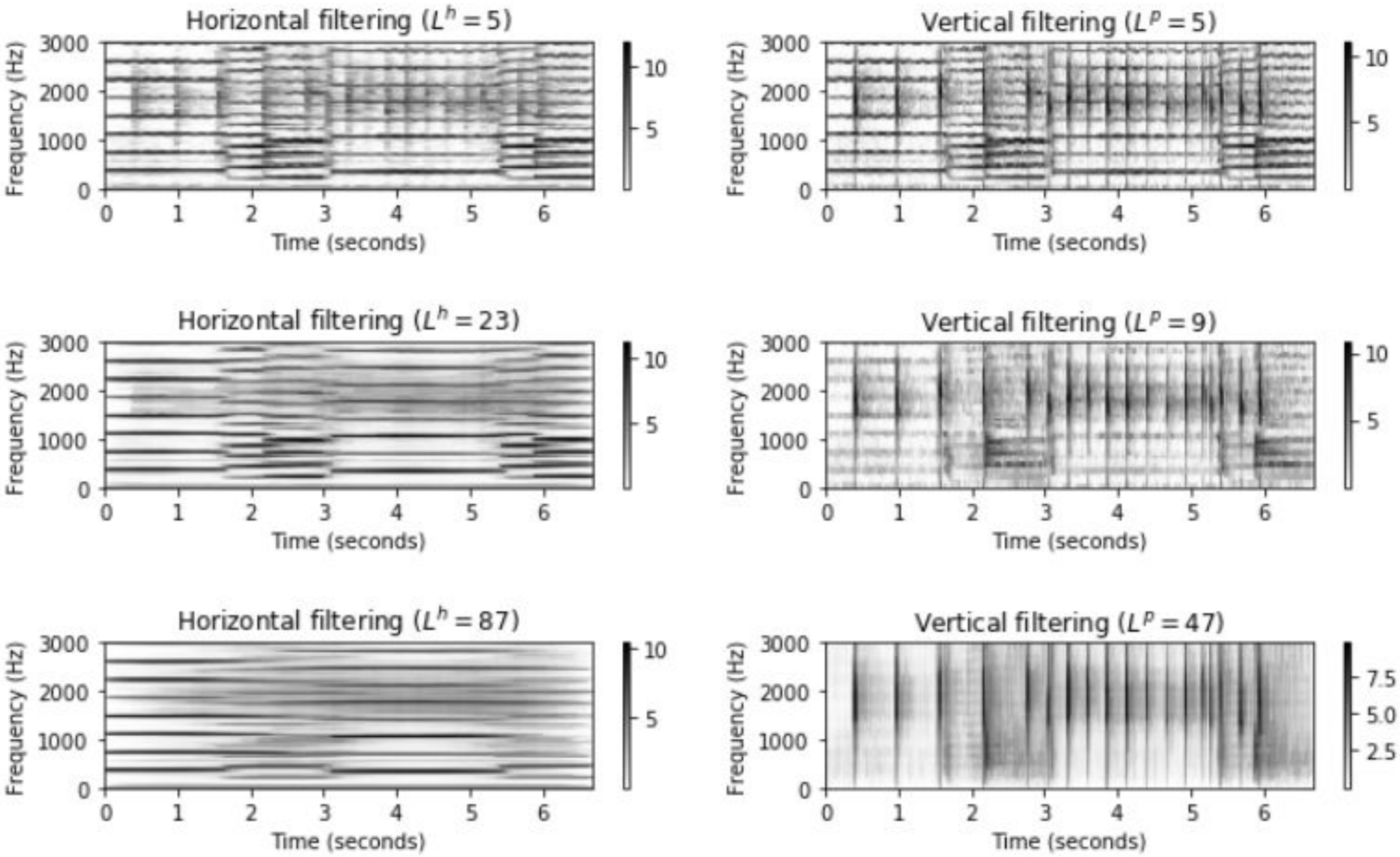
  Given vector x=[8 5 3 2 7]

  Put the elements in increasing order ,x=[2 3 5 7 8]

  median(x)=5

horizontal and vertical median filter





Horizontal filtering ($L^h = 5$)
Vertical filtering ($L^p = 5$)
Horizontal filtering ($L^h = 23$)
Vertical filtering ($L^p = 9$)
Horizontal filtering ($L^h = 87$)
Vertical filtering ($L^p = 47$)
Frequency (Hz)
Time (seconds)

Binary Masking

- each time–frequency bin is assigned either the value one or the value zero.

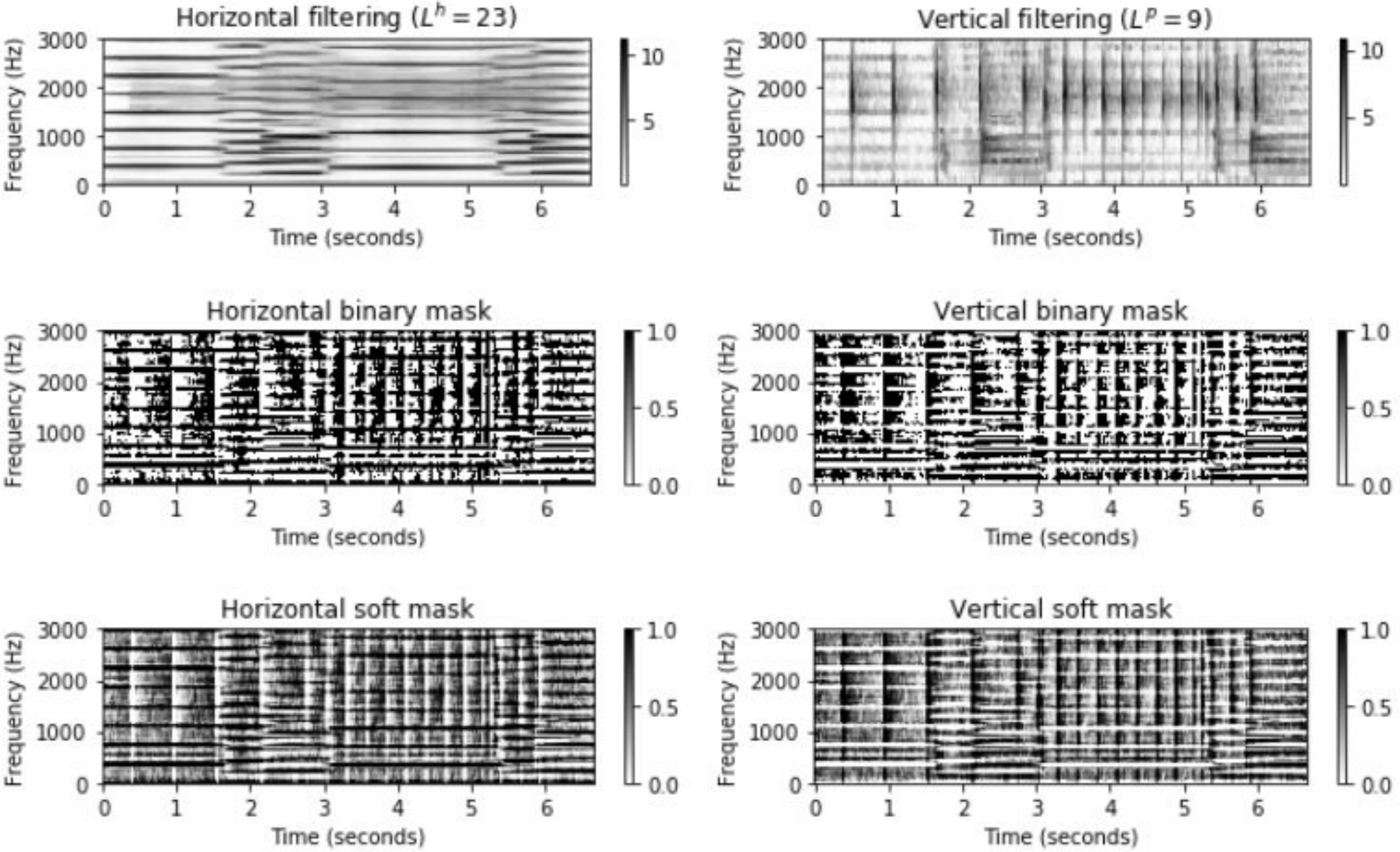
$$\mathcal{M}^{\mathrm{h}}(n,k) := \begin{cases} 1, & \text{if } \tilde{\mathcal{Y}}^{\mathrm{h}}(n,k) \geq \tilde{\mathcal{Y}}^{\mathrm{p}}(n,k), \\ 0, & \text{otherwise,} \end{cases}$$

$$\mathcal{M}^{\mathrm{p}}(n,k) := \begin{cases} 1, & \text{if } \tilde{\mathcal{Y}}^{\mathrm{h}}(n,k) < \tilde{\mathcal{Y}}^{\mathrm{p}}(n,k), \\ 0, & \text{otherwise} \end{cases}$$

- Instead of a binary (hard) decision, one can consider a relative weighting when comparing the magnitudes of spectral coefficients, this is called **soft masking.**

- To obtain the component, the mask is applied to the original spectrogram by pointwise multiplication.

$$\mathcal{Y}^{\mathrm{h}}(n,k) := \mathcal{M}^{\mathrm{h}}(n,k) \cdot \mathcal{Y}(n,k),$$
$$\mathcal{Y}^{\mathrm{p}}(n,k) := \mathcal{M}^{\mathrm{p}}(n,k) \cdot \mathcal{Y}(n,k)$$



Horizontal filtering ($L^h = 23$)
Vertical filtering ($L^p = 9$)
Horizontal binary mask
Vertical binary mask
Horizontal soft mask
Vertical soft mask
Frequency (Hz)
Time (seconds)

Signal Reconstruction (Inverse STFT)

- Apply the two masks directly to the original STFT X yielding two complex-valued masked STFTs X_h and X_p.

$$\mathcal{X}^{\mathrm{h}}(n,k) := \mathcal{M}^{\mathrm{h}}(n,k) \cdot \mathcal{X}(n,k),$$
$$\mathcal{X}^{\mathrm{p}}(n,k) := \mathcal{M}^{\mathrm{p}}(n,k) \cdot \mathcal{X}(n,k)$$

- Then apply Inverse STFT of X_h and X_p , we get the separated harmonic and percussive

Result

[MATLAB]

References

1) J. Driedger, M. M¨uller, and S. Ewert, “Improving time-scale modification of music signals using harmonic-percussive separation,” Signal Processing Letters, IEEE, vol. 21, no. 1, pp. 105–109, 2014

2) Driedger, J., M. Muller, and S. Disch. "Extending harmonic-percussive separation of audio signals." Proceedings of the International Society for Music Information Retrieval Conference. Vol. 15, 2014.

3) https://www.audiolabs-erlangen.de/resources/MIR/FMP/C8/C8S1_HPS.html

Thank you