Signal decomposition by a joint pitch, timbre and wideband model

Ricard Marxer

September 2011

1 Signal Decomposition Model

The main assumption of our spectrum decomposition method is that the short-term Fourier transform (STFT) of our audio signal, Y is a linear combination of N_C elementary spectra, also named basis components. This can be expressed as $Y = BG$ where $Y \in \mathbb{R}^{N_S \times 1}$ is the spectrum at a given frame m, N_S being the size of the spectrum. $B \in \mathbb{R}^{N_S \times N_C}$ is the matrix whose columns are the basis components, it is also referred to as the basis matrix. $G \in \mathbb{R}^{N_C \times 1}$ is a vector of component gains for the current frame.

Our focus is on low latency, unsupervised applications which require the decomposition of each spectrum frame to be done very quickly. Therefore, we will only consider solutions in which the basis components B are constant and fixed a priori.

It is obvious that the choice of the basis matrix has a large influence on the decomposition results. It is not in the scope of this article to study the effect of the basis matrix, but rather to propose a computationally cheap method to perform the decomposition given a suitable basis matrix.

As in many other NMF based [1, 2] approaches we set the basis matrix to be composed of a set of N_P single pitch multiple-harmonic spectra. However in order to model harmonic sources of different timbres we must allow different spectral envelopes. This is done by filtering the single pitch components with a filterbank of N_F filters. This results in a total of $N_P \cdot N_F$ harmonic basis components.

Modeling only harmonic sources is often not enough to explain all the possible observed spectra. In [3] the authors propose modelling wideband components to reconstruct transient sounds or background noise. We take a similar approach by adding to our basis matrix the spectra of the filters in our filterbank as wideband components. This results in a total of $N_C = (N_P + 1) \cdot N_F$.

The spectra components can be defined as:

$$\begin{aligned}
\varphi[i,n] &= 2\pi f_l H N_P \frac{2^{\frac{iH-F/2+n}{HN_P}} - 1}{S_r \ln(2)} \\
E_i[\omega] &= \sum_{n=0}^{F} w[n] \left(\sum_{h=1}^{N_h} sin\left(h\varphi[i,n]\right) \right) e^{-j\omega n} \\
B_{i,k}[\omega] &= \begin{cases} U_k[\omega] E_i[\omega] & \text{if } i \leq N_P \\ U_k[\omega] & \text{if } i = N_P + 1 \end{cases}
\end{aligned} \tag{1}$$

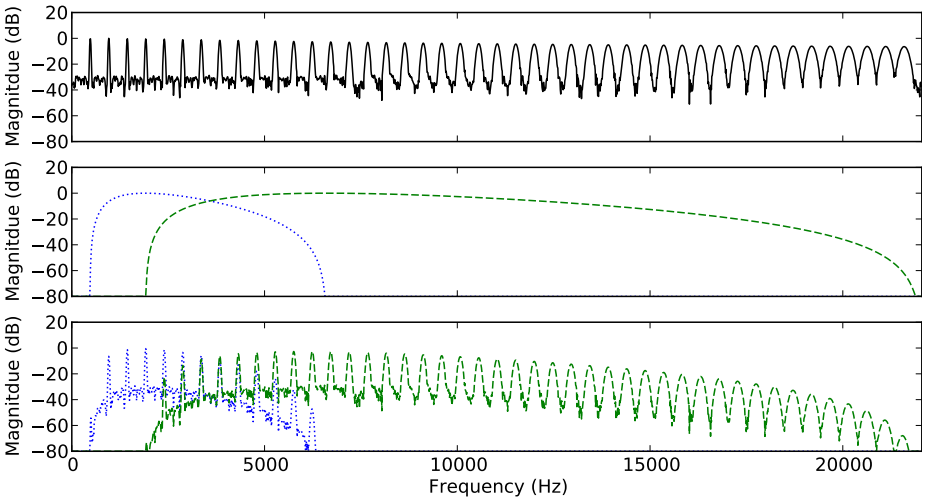

Figure 1: Two components of our basis matrix B. Top shows $E_i[\omega]$ for a frequency of 480Hz. Middle shows $U_k[\omega]$ for two consecutive values of k. Bottom shows $B_{i,k}[\omega]$ for the $E_i[\omega]$ and $U_k[\omega]$ shown above.

with $H = (1 - \alpha)F$. Where α is a coefficient to control the frequency overlap between the components, F is the frame size, S_r the sample rate, $w[n]$ is the analysis window, N_h is the number of harmonics of our components, $B_{i,k}$ is the spectrum of the component of i^{th} pitch filtered by k^{th} filter. U_k is the spectrum of the k^{th} filter in our filterbank. U_k is constructed as a sequence of N_F Hann windows, linearly distributed in the Mel scale and with a 50% overlap.

The column vectors $B_{i,k}$ are stacked horizontally to form the matrix B. This results in the spectrum $B_{i,k}$ of the component of i^{th} pitch and k^{th} filter being the column vector B_{iN_F+k}.

References

[1] T. Virtanen, "Monaural sound source separation by nonnegative matrix factorization with temporal continuity and sparseness criteria," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 3, pp. 1066 –1074, 2007.

[2] J.-L. Durrieu, G. Richard, B. David, and C. Févotte, "Source/filter model for unsupervised main melody extraction from polyphonic audio signals," *IEEE Transactions on Audio, Speech, & Language Processing*, vol. 18, no. 3, pp. 564 –575, 2010.

[3] J. Wu, E. Vincent, S. Raczynski, T. Nishimoto, N. Ono, and S. Sagayama, "Multipitch estimation by joint modeling of harmonic and transient sounds," in *IEEE Int. Conf. on Acoustics, Speech & Signal Processing (ICASSP)*, Prague, Czech Republic, May 2011.