Monaural Singing Voice Separation Using Non-Negative Matrix Factorization

Nicholas Esterer

April 29th, 2015

Applications:

- Melody extraction
- Song recognition
- Singer identification

Features of the Singing Voice

- Compared to harmonic instrumental accompaniment, the singing voice exhibits a spectrum whose frames vary over time whereas instrumental accompaniment remains static.
- Compared to percussive instrumental accompaniment, the singing voice spectrum does not vary as greatly over time as brief percussive impulses

Why?

- In Marin and McAdams (1991), studies showed that modulating the frequency and vowel of a synthesized tone increased its prominence among unmodulated vowels.
- This suggest that singers modulate frequency and amplitude of the voice to have prominence over other harmonic instruments.

Characteristics of Singing Voice Spectrograms

- Because the singing voice exhibits some frequency and amplitude modulation, taking a long window spectrogram of the voice will show excitation in different bins of adjacent frames, while a spectrogram of the more static harmonic accompaniment will show adjacent frames sharing prominent bins.

- Because the singing voice exhibits less frequency and amplitude modulation than percussive instruments, taking a short window spectrogram of the singing voice will show excitation of the same bin in adjacent frames whereas the spectrogram of a percussive sound will not share prominent bins among adjacent frames.

Technique

- These observations motivate a technique that removes those elements of the spectrogram that are not deemed to stem from the voice:
  - Remove spectra whose prominent bins are non-varying on a long-window spectrogram.
  - Remove spectra whose prominent bins are varying on a short-window spectrogram

- Of course, the singing voice and the instrumental accompaniment will often be simultaneous in the recording and their respective spectra also simultaneous.
- We would like to separate these spectra.
- Hennequin et al. (2011), among others, demonstrated that non-negative matrix factorization (NMF) is good at separating spectrograms into a set of basis vectors $\mathbf{B}$ and their time-varying gains $\mathbf{G}$.
- We use NMF on both the long- and short-window spectrograms to compute basis spectra.
- We then identify spectra stemming from harmonic instruments and percussive instruments, respectively.

The following calculation is used to measure the amount of spectral discontinuity in the j^{th} column of $\mathbf{B}$ (a basis vector) in order to identify spectra stemming from harmonic instruments

$$d_s(\mathbf{B}^j) = \frac{\sum\limits_{k=2}^{K} (\mathbf{B}_{k,j} - \mathbf{B}_{k-1,j})^2}{\sum\limits_{k=1}^{K} (\mathbf{B}_{k,j}^2)}$$

where K is the number of bins in the spectrogram.

Similarly, the following calculation is used to measure the amount of discontinuity in the activation gains in the j^{th} column of $\mathbf{G}$ in order to identify spectra stemming from percussive instruments

$$d_s(\mathbf{G}^j) = \frac{\sum_{t=2}^{T}(\mathbf{G}_{j,t} - \mathbf{G}_{j,t-1})^2}{\sum_{t=1}^{T}(\mathbf{G}_{j,t}^2)}$$

where T is the number of frames in the spectrogram.

Given a spectrogram $\mathbf{X}$ of K bins and T frames, NMF will find an approximate factorization

$$\mathbf{X} \approx \mathbf{BG}$$

where $\mathbf{B}$ is a matrix of K rows and J columns and $\mathbf{G}$ is a matrix of J rows and T columns.

When using the Kullback-Leibler divergence and interpreting it as a β-divergence where $\beta = 1$, then according to Hennequin et al. (2011b), the divergence is non-increasing when using the following update rules:

$$\mathbf{B} \leftarrow \mathbf{B}.\frac{\frac{\mathbf{X}}{\mathbf{BG}}\mathbf{G}^T}{\mathbf{1}\mathbf{G}^T}$$

$$\mathbf{G} \leftarrow \mathbf{G}.\frac{\mathbf{B}^T\frac{\mathbf{X}}{\mathbf{BG}}}{\mathbf{B}^T\mathbf{1}}$$

Where . indicates element-wise multiply and indicates element-wise division.