Melody and accompaniment separation using enhanced binary masks

Shayenne da Luz Moura[1*], **Marcelo Queiroz**[1]

[1]Computer Science Department, University of São Paulo

shayenne@ime.usp.br, mqz@ime.usp.br

Abstract

Recovering melodic information from sound signals has several applications, being an important task in Computational Auditory Scene Analysis. This work presents enhancement methods for filtering the spectrogram of an audio signal based on melodic annotations for the separation of melody and accompaniment in different audio tracks. Preliminary quantitative results correlate well with subjective evaluations, showing that enhanced binary masks provide a reasonable starting point for the refinement of automatic melodic separation strategies based on spectrogram resynthesis.

Introduction

This work deals with the processing of musical signals aiming at the extraction of melodies based on annotated melodic transcriptions. The main contribution is the proposition of enhanced harmonic binary masks aimed at preserving important timbral and transient characteristics of the melodic instrument being extracted.

The input audio signal x is first transformed into a time-frequency representation in the form of a spectrogram $\mathcal{X}(m,k)$where m and k are frame and frequency indices.This spectrogram is invertible by taking the IFFT of each audio frame and overlap-adding the results. The goal is to define a binary mask $\mathcal{B}(m,k)$ that acts as a spectrogram filter allowing the decomposition of the signal in two parts, $\mathcal{X}_{\mathcal{M}} = \mathcal{B}\mathcal{X}$ and $\mathcal{X}_{\mathcal{A}} = (1-\mathcal{B})\,\mathcal{X}$, in such a way that the resynthesis of $\mathcal{X}_{\mathcal{M}}$ contains the melody and the resynthesis of $\mathcal{X}_{\mathcal{A}}$ contains the accompaniment.

By knowing the spectrogram and an annotated melodic line, given by either a human expert of as the result of an automatic pitch tracker such as Melodia [1], it is possible to create a crude binary mask containing only the spectrogram bins corresponding to this F0-profile and use it for resynthesis. Of course most melodic instruments will produce harmonic spectra, so extending this binary mask to include all the harmonic components of this F0-profile is a natural idea. Once a binary melodic mask is defined, the remaining spectrogram bins would produce the accompaniment through resynthesis, using the binary complement of the melodic mask.

It is known that distinct sound sources have different spectral behaviors: some sources have smooth onsets while others produce many transient noises at the attack section which are important for the characterization of their timbre; These individual timbral traits of music sound sources demand the refinement of melodic masks so that the audio signal obtained through resynthesis preserves these characteristics.

Melodic mask enhancements

In order to achieve perceptually expressive results, a number of binary masks were proposed and compared starting with the annotated F0-profile provided in the input: (1) Melodic mask selecting spectrogram bins of the pure F0-profile; (2) Harmonic Mask containing the F0-profile and its higher harmonics; (3) Dilated harmonic mask including upper and lower neighboring bins; (4) Adaptive dilated harmonic mask with spectral width given by a spectral novelty function [2]; (5) Adaptive percussive dilated mask, as the previous mask but expanding only through percussive[1] bins (see example in Figure 1); (6) Adaptive dilated mask with backward spreaded onsets[2]; (7) Adaptive percussive dilated mask with backward spreaded onsets.

*Both authors acknowledge their support by CNPq.

[1]Based on percussive + harmonic binary spectrogram masks [2].

[2]Wherever an onset is detected in a harmonic partial this mask spreads and dilates through bins of previous frames to colect transient/noisy elements

After resynthesis from the filtered spectrogram it is possible to study the perceptual impact of each one of the strategies for melody extraction. Whenever a ground truth is available (e.g. in multichannel recordings where the audio of the melodic instrument is known) it is also possible to correlate perceptual data with objective measurements.

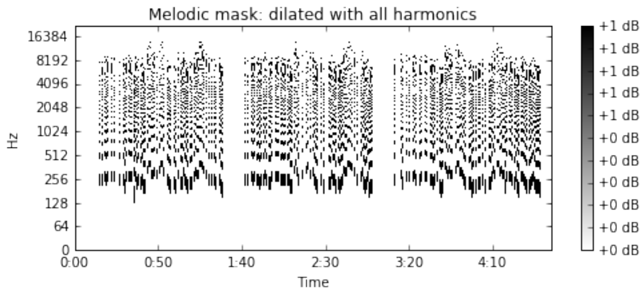

Figure 1: Example of mask of type (5)

Implementation and evaluation

The melody-accompaniment separation system was developed in Python, using specialized libraries such as *libROSA*[3] for audio processing and *SciPy*[4] for image processing. The Jupyter Notebook environment was used to present the results in an interactive way.

The MedleyDB dataset [3] was used to evaluate the system. Five audio signals were selected which had different musical styles and contained annotated F0-profiles and also separate audio tracks for individual instruments. These were used to evaluate the melody-accompaniment separation, both objectively and perceptually, using all the proposed enhanced masks.

For each piece, the spectrogram of the isolated melodic instrument is used as ground truth. The correlation between each mask and ground truth is calculated, along with *accuracy* and *signal level*. These are defined similarly to the well-known information retrieval measures precision and recall, but refer to energy values: accuracy is defined as how much of the spectral energy of the melodic instrument was retrieved from the melodic mask, and signal level is defined as the part of the recovered energy that was actually part of the melodic instrument. These values are summarized via their harmonic mean (akin to the F-measure in information retrieval, see Figure 2).

[3]http://librosa.github.io
[4]http://docs/scipy.org

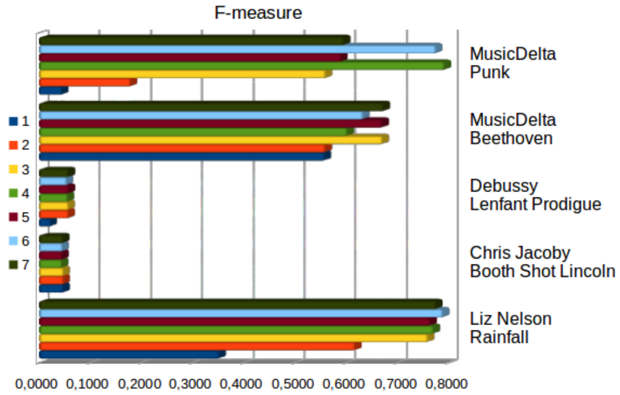

Figure 2: Results for enhanced masks

The subjective evaluation consisted of the perceptual evaluation of the extracted melodies. It was verified that subjectively preferred melodies were generated by masks with higher accuracy or harmonic mean values.

Conclusion

This work investigated melody-accompaniment separation based on enhanced binary masks. Comparison of objective and subjective results showed that these masks are able to filter out the melodic instrument from polyphonic musical contexts, but each enhancement might be better suited to different inputs according to timbral characteristics of the corresponding melodic instrument.

References

[1] J. Salamon and E. Gómez. Melody extraction from polyphonic music signals using pitch contour characteristics. *IEEE Transactions on Audio, Speech, and Language Processing*, 20(6):1759–1770, Aug. 2012.

[2] M. Müller. *Fundamentals of Music Processing: Audio, Analysis, Algorithms, Applications*. Springer International Publishing, 2015.

[3] R. M. Bittner, J. Salamon, M. Tierney, M. Mauch, C. Cannam, and J. P. Bello. MedleyDB: A multitrack dataset for annotation-intensive MIR research. In *15th Int. Soc. for Music Info. Retrieval Conf.*, pages 155–160, Taipei, Taiwan, Oct. 2014.