Analysis of polyphonic audio using source-filter model and non-negative matrix factorization

Tuomas Virtanen* **and Anssi Klapuri**
Tampere University of Technology, Institute of Signal Processing
Korkeakoulunkatu 1, FI-33720 Tampere, Finland
tuomas.virtanen@tut.fi, anssi.klapuri@tut.fi

Abstract

This paper proposes a method for analysing polyphonic audio signals based on a signal model where an input spectrogram is modelled as a linear sum of basis functions with time-varying gains. Each basis is further represented as a product of a "source" spectrum and the magnitude response of a "filter". This formulation reduces the number of free parameters needed to represent realistic audio signals and leads to a more reliable parameter estimation. Two novel estimation algorithms are proposed, one extended from non-negative matrix factorization and the other from non-negative matrix deconvolution. In preliminary experiments with singing signals, both algorithms have been found to converge towards meaningful analysis results.

1 Introduction

Model-based analysis of audio signals has received increasing attention in recent years and even quite complex generative models have been proposed [2, 3]. Applications of this include for example sound separation, audio coding, music transcription, and sound source recognition.

This paper proposes a method for analysing the component sounds in a polyphonic audio signal based on the source-filter model of sound production. Here "source" refers to a vibrating object such as a guitar string, and "filter" represents the resonance structure of the rest of the instrument which colors the produced sound. The source varies with pitch and the degree of periodicity, whereas the filter varies with timbre. This framework has been used for decades in speech coding [7] and sound synthesis [9], but has not been properly adopted in signal analysis and classification problems. Usually a less structured approach is employed, modeling sound spectra directly with the Fourier transform or with Mel-frequency cepstral coefficient (MFCC). These have certain limitations that are addressed in this paper.

We have recently proposed an algorithm for modeling the time-varying spectral energy distribution of musical sounds when presented in isolation [5]. In that work, the source-filter model was found to lead to a clear improvement over MFCC-like models. Here the

*This work was supported by the Academy of Finland, project No. 5213462 (Finnish centre of Excellence program (2006 - 2011).

analysis is performed using a completely different estimation algorithm which is suitable for the analysis of polyphonic signals. Also, several different filters are allowed per one sound source, which makes the model more suitable for singing, for example, were the vocal tract filter varies over time.

As a framework for polyphonic signal analysis we use a linear signal model for the magnitude spectrum $x_t(k)$ of the mixture signal, where k is frequency index and t is frame index. The signal $x_t(k)$ is modelled as a sum of basis functions:

$$\hat{x}_t(k) = \sum_{n=1}^{N} g_{n,t} b_n(k) \tag{1}$$

where $g_{n,t}$ is the gain of basis function n in frame t, and $b_n(k)$, $n = 1, \ldots, N$ are the bases. The bases represent the magnitude spectra of different musical tones. Several active bases are allowed at each time, making the model suitable for polyphonic signals. Existing unsupervised learning methods for estimating the basis functions and gains include independent component analysis (ICA) [1], sparse coding [10], and non-negative matrix factorization (NMF) [8]. When used for sound separation, currently the best results have been obtained using NMF [11].

2 Proposed signal model

A problem with the model (1) is that the mixture signal is represented as a sum of fixed spectra: each pitch value of each instrument requires a distinct basis function. The large number of free parameters makes the estimation less reliable, and also, clustering the estimated bases to sound sources is difficult.

In the proposed signal model, each basis $b_n(k)$ is described as a product of the magnitude spectra of an excitation (source) $e_i(k)$ and a filter $h_j(k)$. This leads to the model

$$\hat{x}_t(k) = \sum_{i,j} g_{i,j,t} e_i(k) h_j(k) \tag{2}$$

which assigns one excitation per pitch value and one filter per instrument (or phoneme in singing). A polyphonic signal consists of several excitation and filter combinations occurring simultaneously or in sequence.

We denote the number of excitations by I and the number of filters by J. Note that the number of bases, $(I \times J)$, is significantly larger than that of excitations or filters. As a consequence, the number of parameters to estimate is smaller (bases are restricted to $b_n(k) = e_i(k)h_j(k)$) which makes the estimation more reliable. Also, the proposed signal model associates components with the same timbre (resp. pitch), leading to an automatic clustering of bases to sound sources (resp. musical notes).

In Section 4, we also introduce a signal model which allows the pitch of an individual excitation signal to vary. This makes it more suitable for singing, where pitch values are not quantized and therefore cannot be well represented with a countable set of excitations. In that latter model, a singing signal can be represented with a single pitch-varying excitation and multiple filters (one per phoneme).

3 Non-negative matrix factorization algorithm for parameter estimation

When the bases are magnitude or power spectra, it is natural to restrict them to be entry-wise non-negative. Furthermore, the model can be restricted to be purely additive by limiting the gains to be non-negative. NMF estimates the bases and their gains by minimizing

the reconstruction error between the observed spectrogram and the model while restricting the parameters to non-negative values. It has turned out that the non-negativity restrictions alone are sufficient for sound source separation [8].

Commonly used measures for the reconstruction error are the Euclidean distance, and divergence d, defined as

$$d(x, \hat{x}) = \sum_{k,t} x_t(k) \log \frac{x_t(k)}{\hat{x}_t(k)} - x_t(k) + \hat{x}_t(k) \tag{3}$$

The divergence is always non-negative, and zero only when $x_t(k) = \hat{x}_t(k)$ for all k and t. It can be minimized for example using the multiplicate updates proposed by Lee and Seung [6]: the parameters are initialized to random non-negative values, and updated by applying multiplicative update rules iteratively. Each update decreases the value of the divergence, until the algorithm converges.

We propose an augmented NMF algorithm for estimating the parameters of the model (2). Multiplicative updates which minimize the divergence (3) are given by

$$g_{i,j,t} \leftarrow g_{i,j,t} \frac{\sum_k r_t(k) e_i(k) h_j(k)}{\sum_k e_i(k) h_j(k)}, \tag{4}$$

$$e_i(k) \leftarrow e_i(k) \frac{\sum_{j,t} r_t(k) g_{i,j,t} h_j(k)}{\sum_{j,t} g_{i,j,t} h_j(k)}, \tag{5}$$

and

$$h_j(k) \leftarrow h_j(k) \frac{\sum_{i,t} r_t(k) g_{i,j,t} e_i(k)}{\sum_{i,t} g_{i,j,t} e_i(k)}, \tag{6}$$

where $r_t(k) = \frac{x_t(k)}{\hat{x}_t(k)}$ is evaluated using (2) before each update.

The overall estimation algorithm is given as follows:

1. Choose the number of excitations and filters. Initialize each parameter $g_{i,j,t}$, $e_i(k)$, and $h_j(k)$ with a random positive values.
2. Update the gains using (4).
3. Update the excitations using (5).
4. Update the filters using (6).
5. Repeat steps 2-4 until the algorithm converges.

It can be shown that the divergence (3) is non-increasing under each update. When prior knowledge about the sources is available, it can be used to initialize the excitations and filters.

4 Representing several pitch values with a single excitation

The linear model (1) requires multiple basis functions to represent tones with different pitch values. This limitation has been addressed, for example, by FitzGerald [4] and Virtanen [12, pp. 57-65], who translated basis functions on logarithmic frequency axis to produce different fundamental values. In this model, each gain $g_{n,t}$ in (1) is replaced by gain $g_{n,t,\tau}$, which denotes the amount of contribution of the n^{th} basis function, which is translated by τ frequency bins. The model can be written as

$$\hat{x}_t(k) = \sum_{n,\tau} g_{n,t,\tau} b_n(k - \tau). \tag{7}$$

In this model, the translation affects the entire basis function, i.e. the product of the excitation and the filter, and therefore the filter becomes translated, too. A more realistic model is obtained by producing different fundamental frequency values by translating a single harmonic excitation, and keeping the filter fixed. When the spectrum is modeled as a product of excitation $e_i(k)$ and filter $h_j(k)$, the model can be written as

$$\hat{x}_t(k) = \sum_{i,j,\tau} g_{i,j,t,\tau} e_i(k-\tau) h_j(k). \tag{8}$$

Parameters of the model (7) have been estimated by algorithms extended from NMF. Similar approach can be used to estimate the parameters of the proposed model (8). Multiplicative updates which minimize the divergence (3) for model (8) are given by

$$g_{i,j,t,\tau} \leftarrow g_{i,j,t,\tau} \frac{\sum_{k,z} r_t(k+z) e_i(k) h_j(k)}{\sum_{k,z} e_i(k+z) h_j(k+z)}, \tag{9}$$

$$e_i(k) \leftarrow e_i(k) \frac{\sum_{j,t,z} r_t(k+z) g_{i,j,t,z} h_j(k+z)}{\sum_{j,t,z} g_{i,j,t,z} h_j(k+z)}, \tag{10}$$

and

$$h_j(k) \leftarrow h_j(k) \frac{\sum_{i,t,z} r_t(k) g_{i,j,t,z} e_i(k-z)}{\sum_{i,t,z} g_{i,j,t,z} e_i(k-z)}. \tag{11}$$

The overall estimation is similar to the algorithm for NMF. The parameters are initialized with random positive values, and updated sequentially using Equations (9)-(11).

The model is especially suitable for singing voice, since only a single excitation is required to model all harmonic tones, and different phonemes can be modeled using different filters.

5 Conclusions

Estimation algorithms were proposed for two signal models (2) and (8), both of which aim at reducing the amount of free parameters needed to represent realistic audio signals. In preliminary experiments with singing signals, both algorithms were found to converge and to find more meaningful bases than the conventional NMF which uses the signal model (1).

References

[1] M. A. Casey and A. Westner. Separation of mixed audio sources by independent subspace analysis. In *International Computer Music Conference*, Berlin, Germany, 2000.

[2] A.T. Cemgil, B. Kappen, and D. Barber. A generative model for music transcription. *IEEE Trans. Speech and Audio Processing*, 13(6), 2005.

[3] M. Davy, S. Godsill, and J. Idier. Bayesian analysis of polyphonic Western tonal music. *Journal of the Acoustical Society of America*, 119(4):2498–2517, 2006.

[4] Derry FitzGerald, Matt Cranitch, and Eugene Coyle. Generalised prior subspace analysis for polyphonic pitch transcription. In *Proceedings of International Conference on Digital Audio Effects*, Madrid, Spain, 2005.

[5] A. Klapuri. Analysis of musical instrument sounds by source–filter–decay model. In *International conference of acoustics, speech and signal processing*, Honolulu, Hawaii, USA, 2007. **Submitted for review**.

[6] D. D. Lee and H. S. Seung. Algorithms for non-negative matrix factorization. In *Neural Information Processing Systems*, pages 556–562, Denver, USA, 2001.

[7] M.R. Schroeder and B.S. Atal. Code-excited linear prediction (CELP): High-quality speech at very low bit rates. In *IEEE ICASSP*, pages 937–940, Tampa, Florida, 1985.

[8] Paris Smaragdis and J. C. Brown. Non-negative matrix factorization for polyphonic music transcription. In *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, New Paltz, USA, 2003.

[9] V. Välimäki, J. Pakarinen, C. Erkut, and M. Karjalainen. Discrete-time modelling of musical instruments. *Reports on progress in physics*, 69:1–78, 2006.

[10] Tuomas Virtanen. Sound source separation using sparse coding with temporal continuity objective. In *International Computer Music Conference*, Singapore, 2003.

[11] Tuomas Virtanen. Monaural sound source separation by non-negative matrix factorization with temporal continuity and sparseness criteria. *IEEE Transactions on Audio, Speech, and Language Processing*, 2006. Accepted for publication.

[12] Tuomas Virtanen. *Sound Source Separation in Monaural Music Signals*. PhD thesis, Tampere University of Technology, 2006. available at `http://www.cs.tut.fi/~tuomasv`.