CONVOLUTIVE SPARSE CODING OF AUDIO SPECTROGRAMS

Tuomas Virtanen

Tampere University of Technology
P.O. Box 553, 33101 TAMPERE, FINLAND
tuomas.virtanen@tut.fi

1. INTRODUCTION

Representations which reduce redundancy and estimate latent variables behind observed data have turned out to be efficient in machine learning. Most of the representations model each observation vector as as weighted sum of N basis functions $\mathbf{a}_n$, so that

$$\mathbf{x}_t = \sum_{n=1}^{N} \mathbf{a}_n s_{n,t}, \tag{1}$$

where $s_{n,t}$ is the amount of contribution of the n^{th} basis function in the t^{th} observation.

There are methods which use fixed basis functions, but recently, many algorithms for the estimation of adaptive representations have been proposed, and they have been successfully used in several applications. For example, independent component analysis (ICA) estimates the basis functions by finding a decomposition in which the gains of each basis function are statistically independent from each other. Other criteria are, for example, sparseness and non-negativity of $s_{n,t}$.

1.1. One-channel audio signals

Because phases are perceptually less meaningful, one-channel audio signals are often analyzed using a phaseless mid-level representation, for example power spectrogram, or the magnitude of the short-time Fourier transform, or, magnitude spectrogram. Furthermore, the phases of natural sound sources often behave very irregularly so that they cannot be modeled with a simple linear model.

When the model (1) is used, the spectrogram is representated as a sum of components, each of which has a fixed spectrum and a time-varying gain.

2. CONVOLUTIVE SIGNAL MODEL

Linear model (1) models each observation vector independently. For example, the order of the observation vectors does not usually affect the resulting adaptive representation. However, in many situations there are dependencies between observations. For example, the observations can be samples of a process which evolves slowly over time. By utilizing the dependencies between the observations it is possible to estimate higher-level latent variables, and gain robustness because of the utilization of the relations.

An extension of the linear model is a convolutive signal model. Instead of a static basis functions $\mathbf{a}_n$, we add a shift dimension, to result in D basis functions $\mathbf{a}_n^\tau$, $\tau = 0, \ldots, D-1$. The model is written as a convolution between the basis functions and gains:

$$\mathbf{x}_t = \sum_{n=1}^{N} \sum_{\tau=0}^{D-1} \mathbf{a}_n^\tau s_{n,t-\tau} \tag{2}$$

In the case of audio spectrograms, the model has an intuitive interpretation: the spectrogram is modeled as a sum of repetitions of audio objects n, each of which has a spectrogram $\mathbf{a}_n^\tau$ of length D frames (D should be significantly smaller than the number of observations). The non-zero entries of the gain $s_{n,t}$ describe the locations where the object sets on, and the gains of each repetition.

In the analysis of one-channel audio signals the model has been used by Virtanen in [1] and by Smaragdis in [2]. If each frequency line is considered as an observation instead of each frame, the same model allows time-varying fundamental frequencies, as proposed by FitzGerald [3]. The model has also been used to model time-varying images [4], [5], [6].

3. ESTIMATION CRITERIA AND ALGORITHMS

Magnitude and power spectra are non-negative by their definition. Therefore, it is natural to restrict the spectra of components to be non-negative. The components can also be limited to be purely additive, so that the gains are restricted to non-negative values. The non-negativity restrictions have turned out to be sufficient for the estimation of meaningful basis is several cases. The basis functions and gain can be estimated by minimizing the reconstruction error between the observations and the model, which can be measured, for example, using the Euclidean distance or divergence as proposed by Lee and Seung [7].

The repetitions of natural sound objects are usually sparse, so that the gain $s_{n,t}$ being zero can also be assumed to have a

high probability.

4. APPLICATION TO THE SOURCE SEPARATION OF MUSIC SIGNALS

The discussed methods suit particularly well for the analysis of music signals, since musical signals contain lots of redundancy. Furthermore, many musical signals can be rather well be represented with a fixed spectrum and a time-varying gain. An individual component may represent, for example, all the equal-pitched notes of an instrument.

We tested several algorithms based on the linear model (1) and the convolutive model (2) in source separation of music signals. The algorithms included ICA, non-negative matrix factorization, and non-negative sparse coding.

Test signals were generated by mixing samples of individual notes of pitched musical instruments and drums. The samples were taken randomly from a database which is a combination of samples from the McGill University Master Samples Collection [8], the University of Iowa website [9], IRCAM Studio Online [10], and the DFH Superior commercial sample database [11]. The number of samples within each mixture signal was randomly selected. The number of mixture signals was 300, the length of each being 7 seconds.

The quality of the separation was evaluated by calculating the signal-to-noise ratio (SNR) between each separated and original sample. The best algorithm based on the linear model (1) resulted in average SNR of 6.5 dB, while the best algorithm based on the convolutive model (2) resulted in average SNR of 7.2 dB, which suggests that the convolutive model enables a better separation quality than the linear model.

5. REFERENCES

[1] Tuomas Virtanen, "Separation of sound sources by convolutive sparse coding," in *ISCA Tutorial and Research Workshop on Statistical and Perceptual Audio Processing*, Jeju, Korea, 2004.

[2] Paris Smaragdis, "Discovering auditory objects through non-negativity constraints," in *ISCA Tutorial and Research Workshop on Statistical and Perceptual Audio Processing*, Jeju, Korea, 2004.

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

[4] Rafal Bogacz, Malcolm W. Brown, and Christophe G. Giraud-Carrier, "Emergence of movement sensitive neurons' properties by learning a sparse code for natural moving images.," in *Neural Information Processing Systems*, Denver, USA, 2000, pp. 838–844.

[5] B. A. Olshausen, "Sparse codes and spikes," in *Probabilistic Models of the Brain: Perception and Neural Function*, R. P. N. Rao, B. A. Olshausen, and M. S. Lewicki, Eds., pp. 257–272. MIT Press, 2002.

[6] B. A. Olshausen, "Learning sparse, overcomplete representations of time-varying natural images," in *IEEE International Conference on Image Processing*, Barcelona, Spain, 2003.

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

[8] F. Opolko and J. Wapnick, "McGill University Master Samples," Tech. Rep., McGill University, Montreal, Canada, 1987.

[9] "The University of Iowa Musical Instrument Samples Database," http://theremin.music.uiowa.edu.

[10] "IRCAM Studio Online," http://soleil.ircam.fr/.

[11] "DFH Superior," Toontrack Music, 2003, http://www.toontrack.com/superior.shtml.