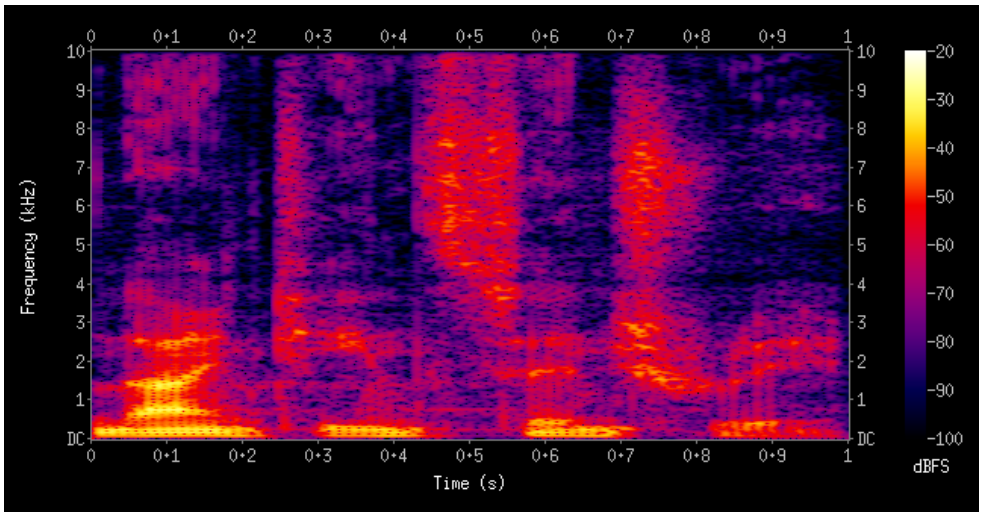
TSIA206

Reading Note

2022 - 2023

Multichannel Nonnegative Matrix Factorization in Convolutive Mixtures for Audio Source Separation

Ozerov, *Member, IEEE*, and Cédric Févotte, *Member, IEEE*



Jeanne Malécot

Abstract

The aim of this reading note is to offer a critical view of an article, briefly outlining the major issues, and then to suggest improvements or share comments on the proposed methods, for example by trying to implement them.

Summary

Non-negative matrix factorization (NMF) is an unsupervised data decomposition technique commonly used in machine learning for processing images and signals, including audio. In the case of audio, spectrograms (magnitude or power) are chosen as data. NMF is particularly well suited to single-channel data. In practice, however, we are often confronted with stereo recordings. One solution is to stack the spectrograms of each channel into a single matrix. This approach can be extended to non-negative tensor factorization (NTF) as part of a parallel factor analysis structure (PARAFAC), where the channel spectrograms form the slices of a 3-valoence tensor. Both NMF and NTF methods rely on certain assumptions about the sources and the mixing process, which are not always realistic. The original sources are often mixed instantaneously, and a subsequent linking step is required to group the elementary components into instrumental sources. In addition, redundancy between channels is not optimally exploited in these methods. The aim of this article is to propose alternatives to the model that remedy these problems.

Two NMF-based audio source separation methods are proposed. The first method uses an expectation-maximization (EM) algorithm to maximize the joint log-likelihood of multichannel data. This method takes advantage of the redundancy between channels and is statistically optimal. It is similar to other model-based multichannel source separation techniques. The second method maximizes the sum of the individual log-likelihoods of all channels using a multiplicative updating (MU) algorithm inspired by NMF. This method takes into account convolutional mixing and does not require the linking of elementary components into sources. Several blind source separation (BSS) techniques have been designed on the basis of similar models. For example, the Independent Component Analysis (ICA) algorithm can be applied to data in each frequency sub-band. However, permutation indeterminacy in ICA leads to the well-known problem of FD-ICA permutation alignment, which cannot be solved without additional a priori knowledge of the sources and/or mixing filters. The EM-based method presented in this paper models sources in the short-time Fourier transform (STFT) domain using multivariate Gaussian mixture models (GMMs). This approach takes into account temporal correlations in the audio signal, assuming stationarity in each window. In the supervised framework, sources are trained in advance.

The source model used in this article models each source frame as a sum of elementary components, rather than as a multi-state process characterized by a covariance matrix.

Both methods have been evaluated on four different data sets and have led to a number of observations. The algorithms are evaluated using the signal-to-distortion ratio (SDR), and the quality of the mixing system estimates is assessed using the mixing error ratio (MER). However, these evaluation criteria can only be calculated when the original spatial images and mixing systems are available. Various algorithm parameters are tested in the experiments. The choice of window size is important, as it enables a balance to be struck between good frequency resolution, the validity of the convolutional mixing approximation and the assumption of local source stationarity. The model's order parameter, which represents the total number of components, is generally set by hand as a function of the number of instrumental sources. However, for non-point sources or similarly mixed sources, this choice becomes more difficult. Artificial noise injection can therefore improve the results.

The results show that the EM algorithm, which maximizes joint likelihood, systematically

improves source separation performance in terms of SDR compared with initial values. On the other hand, the MU algorithm tends to deteriorate the SDR values obtained from oracle initialization. This discrepancy can be attributed to the elimination of mutual information between channels. The EM algorithm, by effectively exploiting the redundancy between channels, achieves better separation results. However, it should be noted that the performance of the EM algorithm can be affected when the assumptions of the mixing process are not verified. In terms of future research directions, they suggest exploring faster algorithms than the EM algorithm for optimizing joint likelihood. They suggest considering strategies that combine EM with faster gradient search methods when a solution is close. In addition, they mention the possibility of incorporating Bayesian extensions to the algorithm, such as primers favoring sparse activation coefficients or more complex primers like Markov chains that favor smoothing of activation coefficients. Automatic order selection is identified as an important area for further research. Determining the total number of components, the number of sources and the partition remains a challenge.

Critical Analysis

The proposed models are based on certain assumptions about the sources and the mixing process, such as instantaneous mixing and spatially uncorrelated noise. These tend not to be verified in basic applications, which limits the practical usefulness of the model. Parameter initialization is therefore of fundamental importance, since performance can be greatly affected. This article addresses this probabilistic issue by introducing a probabilistic framework to represent the data. This approach makes it possible to capture the statistical properties of audio signals and to model the sources and mixing process in a way that is based on reality-based principles, thus adding a realistic aspect. However, the conclusion confirms that further research is still more than necessary to provide better, more automated initialization methods.

The two different models proposed in the article have their respective advantages. The first is an expectation-maximization (EM) algorithm for maximizing the joint log-likelihood of multichannel data, which takes advantage of the redundancy between channels and is statistically optimal. It has been shown to outperform the latest methods in terms of source separation when appropriate initializations are used. We also have a multiplicative updating (MU) algorithm inspired by NMF, which maximizes the sum of the individual log-likelihoods of all channels and takes into account convolutional mixing. It provides an alternative approach that does not require the linking of elementary components in sources.

Depending on the case and the information and methods available for initialization, it is possible to vary and adapt the method used.

Both models have a linear complexity with the number of components, which makes them usable, but implies at least one hour of computation for just one song, which restricts the field of application of these methods in real time.

Another avenue explored by the article is the data-driven approach, which offers a more robust and adaptive solution: only available data is exploited, without the need for pre-trained parameters.

However, some of the complexities of music recordings, such as non-point sources or excessively long reverberations, are still overlooked here.

Ideas and Improvements

My Ideas

To overcome the problems of parameter initialization, deep learning comes to mind as the best solution for almost any problem. Here, the use of Recurrent Neural Networks might seem relevant: we wouldn't need to model anything explicit, as the network would learn complex, non-linear relationships as it analyzed the audio signal.

We might also be interested in methods such as variational autoencoders, again derived from deep learning, but with a generative approach. From different tracks and audios, we could then create a relevant and compact representation of the signal, from which the VAE could reconstruct an audio, enabling us to control the general "shape" of the results we obtain. Another use for VAE would be to perform data augmentation, to improve the robustness of our model, by also proposing data with synthetically added noise, for example.

Interests for this thematic

I was interested in this theme because of its application to music. I've done a lot of instrumental practice myself, and have recently become interested in different recording and mixing methods, so I found it interesting to have a more "scientific" vision of this hobby, and conversely to be able to have concrete images in my head when reading the article. I would have liked to discover more in this article about what deep learning could bring to this field, since it's a subject I particularly enjoyed this end of year, but I still find it interesting that we can get such good results with such explicit models, and therefore different from the neural networks we manipulate, which all seem similar in appearance.

I'm also very fond of arranging and rewriting the scores I listen to, and I think that such methods would be extremely useful and could facilitate, even automate, the time-consuming task of distinguishing by ear a single voice from all those heard on a track. Even if the number of voices is not very high for a program, it doesn't take much for the human ear to get into trouble, influenced by the voice it already knows, for example.