A Probabilistic Latent Variable Model for Acoustic Modeling

Paris Smaragdis
Mitsubishi electric Research Laboratories
Cambridge MA, USA
paris@merl.com

Bhiksha Raj
Mitsubishi electric Research Laboratories
Cambridge MA, USA
bhiksha@merl.com

Madhusudana Shashanka[*]
Dept. of Cognitive and Neural Systems
Boston University
shashanka@cns.bu.edu

Abstract

In this paper we describe a model developed for the analysis of acoustic spectra. Unlike decompositions techniques that can result in difficult to interpret results this model explicitly models spectra as distributions and extracts sets of additive and semantically useful components that facilitate a variety of applications ranging from source separation, denoising, music transcription and sound recognition. This model is probabilistic in nature and is easily extended to produce sparse codes, and discover transform invariant components which can be optimized for particular applications.

1 Introduction

Decompositions of acoustic spectra have been a subject of research for few years now. There have been multiple publications that decompose spectra into principal or independent components, non-negative factors or other types of sparse codes. However quite often these decompositions leave things to be desired either in terms of approximation quality, in terms of computational complexity, or in terms of lacking interpretability. A “good” decomposition would ideally decompose the data into a combination of semantically meaningful components, be computationally inexpensive to derive, and must enable effective signal processing of the audio. In this paper we present a statistical model for the decomposition of acoustic spectra that satisfies all these requirement. In the following sections we describe the model, and show how it enables simple solutions to several common problems in audio.

1.1 Probabilistic Latent Component Analysis

In this section we describe the statistical model we will use for acoustic modeling. Probabilistic Latent Component Analysis (PLCA) is a straightforward extension of Probabilistic Latent Semantic Indexing (PLSI) [1] which deals with an arbitrary number of dimensions and can exhibit various features such as sparsity or shift-invariance. The basic model is defined as:

$$P(\mathbf{x}) = \sum_z P(z) \textstyle\prod_{j=1}^{N} P(x_j|z) \tag{1}$$

[*] Work performed while at Mitsubishi Electric Research Laboratories

where $P(\mathbf{x})$ is an N-dimensional distribution of the random variable $\mathbf{x} = x_1, x_2, ..., x_N$. The z is a latent variable, and the $P(x_j|z)$ are one dimensional distributions. Effectively this model represents a mixture of marginal distribution products to approximate an N-dimensional distribution. Our objective is to discover the most appropriate marginal distributions.

The estimation of the marginals $P(x_j|z)$ is performed using a variant of the EM algorithm. In short this algorithm contains an expectation and a maximization step which we alternate between in an iterative manner. In the expectation step we estimate the 'contribution' of the latent variable z:

$$R(\mathbf{x}, z) = \frac{P(z)\prod_{j=1}^{N} P(x_j|z)}{\sum_{z'} P(z')\prod_{j=1}^{N} P(x_j|z')} \tag{2}$$

and in a maximization step we re-estimate the marginals using the above weighting to obtain a new and more accurate estimate:

$$P(z) = \int P(\mathbf{x})R(\mathbf{x}, z)d\mathbf{x} \tag{3}$$

$$P(x_j|z) = \frac{\int \cdots \int P(\mathbf{x})R(\mathbf{x}, z)dx_k, \forall k \neq j}{P(z)} \tag{4}$$

$P(x_j|z)$ will contain a latent marginal distribution across the dimension of variable x_j, relating to the latent variable z, and $P(z)$ will contain the prior of that latent variable. Repeating the above steps in an alternating manner multiple times produces a converging solution for the marginals and the latent variable priors. This above process can also be adapted to work for a discrete $\mathbf{x}$ and z (or all possible combinations). This process will also work if the provided input $P(\mathbf{x})$ is an unnormalized histogram as opposed to a density. The only added measure we need to take in this case is to normalize each $P(x_j|z)$ to integrate (or sum) to one in every iteration to ensure that it corresponds to a true marginal distribution.

2 Applications of PLCA in audio

Although not traditionally seen as such, energy and power spectra are distributions of acoustic energy over frequency. By extension a magnitude spectrogram is a distribution of acoustic energy across the time-frequency plane. This view is more apparent if one considers the spectrogram to be a sort of histogram which measures the amount of time-frequency "sound quanta" at each point. The only thing that separates a spectrogram from a true distribution is proper normalization.

Adopting this view of spectrograms allows us to use statistical techniques directly. In particular using a technique such as PLCA on spectrograms can yield a lot of desirable properties. The marginal distributions which we can extract will be distributions over time and frequency which can then be used directly with their statistical capacity for sound classification or speech recognition. An additional advantage of this approach is that unlike some other spectral decomposition techniques we maintain the attributes of the input, meaning that we extract actual spectra and time envelopes. Generic techniques such as PCA and ICA are more likely to extract components that contain negative elements which are hard to justify in this setting. Although non-negative matrix factorization can resolve this issue, it is however not statistical in nature which poses a problem when the intent is to use the results in a learning framework. The use of an EM algorithm for the estimation of the marginals and the statistical framework also opens up a lot of possibilities for optional features. Imposing sparsity and various transformation invariances are relatively straightforward operations which are easily incorporated in the estimation step.

Finally we should note that the 2-dimesional PLCA model is numerically identical to the non-negative matrix factorization model. Higher dimensional PLCA corresponds to a non-negative tensor factorization problem.

In the following few sections we briefly describe some applications in which we have used the PLCA model and its extensions.

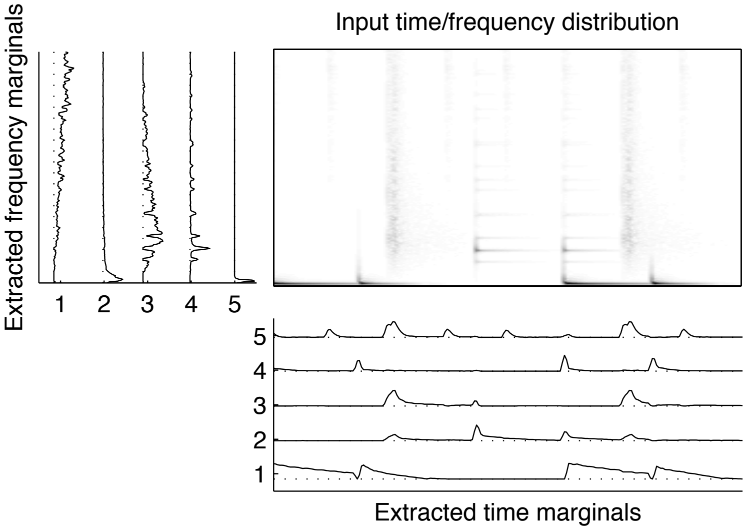

Figure 2.1: Application of PLCA on a spectrogram. The large middle plot displays the spectrogram of a drum loop. The left plot displays the extracted marginals in the frequency axis and the bottom plots the corresponding distributions at the time axis.

2.1 Feature extraction

One of the most straightforward application for this technique is feature extraction. In figure 2.1 we display a spectrogram and the corresponding marginal distributions we have obtained upon application of PLCA. The spectrogram itself contained a drum loop. The extracted marginals are either in the frequency or the time axis. The frequency marginals describe the spectral characteristics that composed the spectrogram. Upon closer examination we can see that the spectrally constituent elements of the spectrogram are described very succinctly in the frequency marginals. Likewise the time marginals describe the temporal elements in the spectrogram. Pairwise combination of corresponding frequency and time marginals gives us a description of the frequency profile and the temporal evolution of all the scene elements.

Note that the extracted marginals, being distributions themselves, are proper spectra and envelopes which do not assume negative values such as the results that a decomposition like PCA or ICA would produce. The number of distributions that we extract (the extend of the latent variable z), determines the detail of the analysis. A small number of components will result in averaged components that give a coarse description of the input. A large number of components will result in more detailed information that uses multiple distributions to describe one source. Depending on the desired use of the features one can decide what level of analysis is best.

This technique can be used to perform a decomposition of an audio scene into its constituent elements and can be further specialized for polyphonic music transcription [2].

2.2 Source recognition in mixtures

One of the fundamental problems in recognition of sounds is that of mixing. Sounds are inherently captured mixed, a property that is not usually addressed in conventional statistical models and machine learning techniques for classification where decisions are mutually exclusive and don't allow characterization for mixtures. The statistical formulation and the additive nature of this decomposition allows us to do so in a seamless manner. The procedure is as follows. For each candidate sound class i we can learn a set of frequency marginals $P_i(f|z)$ from training examples of each sound class. Once confronted with a mixture we have to make the assumption that the mixture is composed of the already known frequency marginals, albeit with different time marginals. The time marginals will reveal how much of each classes frequency marginals are present at any time. To estimate the time marginals we perform PLCA as described above, holding the frequency marginals to

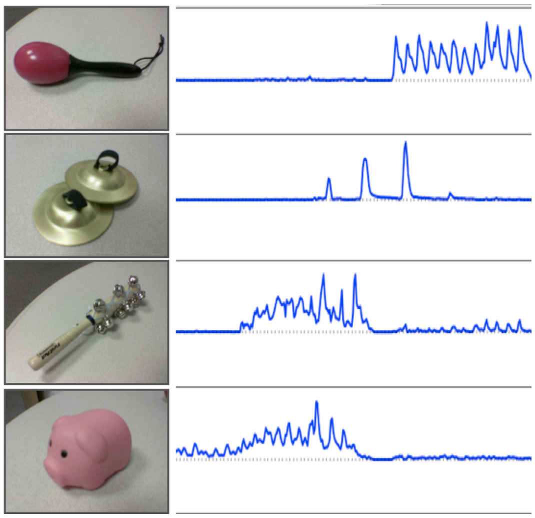

Figure 2.2: Using PLCA for sound recognition. The pictures on the left show the objects that generated each sound class. the Plots on the right show the averages time marginals for each class's frequency marginals as estimated from a mixture spectrogram. The resulting plots accurately describe the manner and order in which the sounds were recorded. With some further manipulation of the time marginals it is straightforward to actually obtain a probabilistic estimate of the dominant source.

those learned from the trained set. Upon conclusion of analysis we will have obtained the priors of each class's marginals $P_i(z)$ and their corresponding time marginals $P_i(t|z)$. The sum over $P_i(t|z)$ over z for each i will be a measure of the presence of class i at any point in time.

The results of an illustrative experiment are shown in figure 2.2. Four different sound classes were trained and then recorded in an overlapping manner (one after the other with a transition overlap). We performed PLCA using all trained frequency marginals from all classes and estimated only the time marginals (the frequency marginals were already estimated from training examples for all sounds). The normalized and averaged time marginals from each class after analyzing the mixture spectrogram are shown in figure 2.2. It is easy to see that the marginals that correspond to each class appropriately reveal the presence of the proper class at the right time, and also reveal the manner in which the sounds were performed.

2.3 Source separation

Extending the idea above we show how we can employ PLCA to perform separation of sources from a busy scene. If we know the frequency marginals that correspond to each source in a mixture we can try to reconstruct the mixture using them. This can be achieved using the process described in the above example. We once again assume that we know the types of sounds in the mixture and that we have pre-trained frequency marginals describing them. Once we estimate the corresponding time marginals that best describe an input we can perform reconstruction by appropriately multiplying and summing all marginals (as shown in equation 2). A selective reconstruction using the marginals from a single source will result in a reconstruction that contains only one source. An example of this is shown in figure 2.3. One of the sounds is a speaker uttering the word "noise" and the other source is a set of chimes. From other instances of these two sounds we learned 100 frequency marginals for each sound. We approximated the input spectrogram with the known frequency marginals and learned the appropriate time marginals. We then selectively reconstructed the input using the marginals corresponding to each source. The results are also shown in figure 2.3.

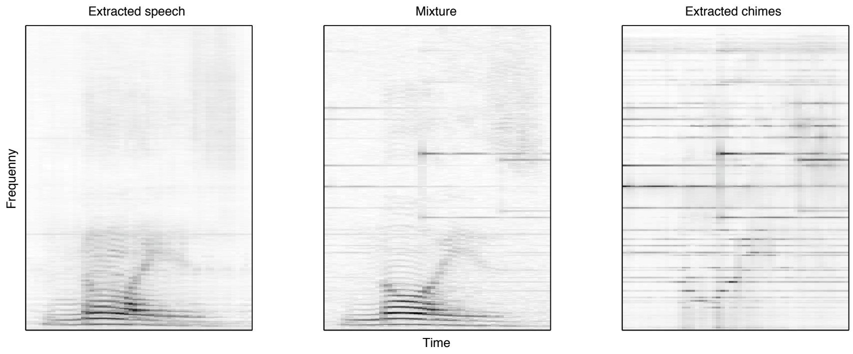

Figure 2.3: Using PLCA for source separation. The middle plot is a mixture spectrogram containing speech and chimes. The left plot is the reconstruction using only speech frequency marginals, and on the right there is a plot of the reconstruction using only the chime frequency marginals. These partial reconstructions effectively separate the sounds in the mixture.

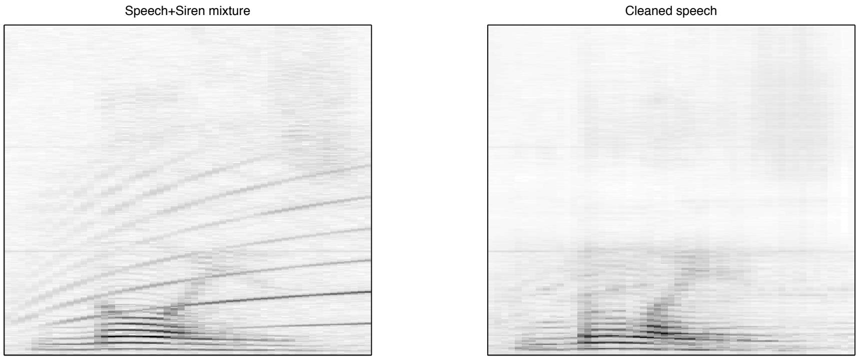

Figure 2.4: Using PLCA for denoising. The left spectrogram is a mixture of speech and a siren. The right spectrogram is the result of a selective reconstruction after using the speech frequency marginals in addition to some learned ones to fit the input, and reconstructing using only the speech marginals.

2.4 Denoising

Finally, a special case of source separation, is denoising. We can use PLCA in a similar manner as above when we know only the frequency marginals of a single source in a mixture and we wish to extract it. The procedure is straightforward. Once again, we perform PLCA holding the known frequency marginals of the desired source fixed. However, we allocate a few additional frequency marginals that are learned; these will adapt to the frequency content of interfering sounds. The time marginals corresponding to all frequency marginals are computed. Upon conclusion of the iterations we will have a set of time marginals which will correspond to frequency marginals of the source we are interested and a set of marginals that describes everything else. We can then use the selective reconstruction method we mentioned in the previous section to extract only the source we are interested in, and all the other sources separately. An example using speech corrupted by a siren sound is shown in figure 2.4.

3 Conclusions

In this document we provided an overview of the PLCA model and how it can be applied for audio related operations. We show how the extracted components are semantically meaningful and maintain the desirable non-negativity property in addition to having a statistical interpretation. This is a model which can easily be extended and employed as a kernel in various other machine learning algorithms. The above and the additivity property which allows us to deal with mixtures have so far proven to be a very useful combination for performing a variety of processing tasks on sounds.

The model is also easily exensible to allow overcomplete sparse representations, invariance to transformations etc. These extensions and their applications will be presented at a future venue.

References

[1] Hofmann, T., 1999. Probabilistic Latent Semantic Indexing in *Proceedings of the Twenty-Second Annual International SIGIR Conference on Research and Development in Information Retrieval* (SIGIR'99).

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