MITSUBISHI ELECTRIC RESEARCH LABORATORIES
http://www.merl.com

Supervised and Semi-Supervised Separation of Sounds from Single-Channel Mixtures

Paris Smaragdis, Bhiksha Raj, madhusudana Shashanka

TR2007-062 July 2006

Abstract

In this paper we describe a methodology for model-based single channel separation of sounds. We present a sparse latent variable model that can learn sounds based on their distribution of time/frequency energy. This model can then be used to extract known types of sounds from mixtures in two scenarios. One being the case where all sound types in the mixture are known, and the other being being the case where only the target or the interference models are known. The model we propose has close ties to non-negative decompositions and latent variable models commonly used for semantic analysis.

International Conference on Independent Component Analysis and Signal Separation

Supervised and Semi-Supervised Separation of Sounds from Single-Channel Mixtures

Paris Smaragdis[1], Bhiksha Raj[1], and Madhusudana Shashanka[2⋆]

[1] Mitsubishi Electric Research Laboratories
Cambridge MA, USA
[2] Department of Cognitive and Neural Systems
Boston University, Boston MA, USA

Abstract. In this paper we describe a methodology for model-based single channel separation of sounds. We present a sparse latent variable model that can learn sounds based on their distribution of time/frequency energy. This model can then be used to extract known types of sounds from mixtures in two scenarios. One being the case where all sound types in the mixture are known, and the other being being the case where only the target or the interference models are known. The model we propose has close ties to non-negative decompositions and latent variable models commonly used for semantic analysis.

1 Introduction

Separation of sounds from single-channel mixtures can be a daunting task. There is no exact solution nor a process that guarantees good separation behavior. Most approaches in this scenario are model-based and perform separation by splitting the spectrogram of the mixture in parts that correspond to a single source. This approach has been taken in [1–4] among many others and has been one of the easiest ways to obtain reasonable results. In this paper we employ a similar approach using a new decomposition algorithm which is best suited for spectrogram analysis. We show how this approach can be used both supervised and semi-supervised settings for separation from monophonic mixtures, and draw connections to various types of known analysis methods.

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) [5] which deals with an arbitrary number of dimensions and can easily extended to exhibit various features such as sparsity or shift-invariance. The basic model is defined as:

⋆ Work performed while at Mitsubishi Electric Research Laboratories. M. Shashanka was partially supported by an AFOSR grant to Prof. Barbara Shinn-Cunningham.

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

where $P(\mathbf{x})$ is a distribution over the N-dimensional random variable $\mathbf{x}$ and $x^{(j)}$ denotes j'th dimension. The random variable $z$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 the EM algorithm. In the expectation step we estimate the posterior probability of the latent variable z:

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

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

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

$$P^*(x^{(j)}|z) = \int \cdots \int P(\mathbf{x}) P(z|\mathbf{x}) dx^{(k)}, \forall k \neq j \tag{4}$$

$$P(x^{(j)}|z) = \frac{P^*(x^{(j)}|z)}{P(z)} \tag{5}$$

Repeating the above steps in an alternating manner multiple times produces a converging solution for the marginals $P(x^{(j)}|z)$ along each dimension j, and the latent variable priors $P(z)$. In the case where $P(\mathbf{x})$ is discrete we only have to substitute the integrations with summations. Likewise the latent variable z can be continuous valued in which case the summations over z become integrals. In practical applications $P(\mathbf{x})$ and z will both be discrete and we assume that to be the case in the remainder of this paper.

1.2 Sparsity Constraints

In this section we will introduce a modification to the PLCA algorithm which enables us to produce sparse (or maximally non-sparse) estimates of $P(x^{(j)}|z)$. Since the estimated quantities of PLCA are probability distributions, we can directly obtain sparsity by imposing an *entropic prior* instead of obtaining the effect by more traditional means such as L1-norm minimization. This prior can can impose a bias towards estimating a low (or high) entropy $P(x^{(j)}|z)$. We can thus obtain a sparse estimate by requesting low entropy results, a flatter estimate by requesting high entropy results, or any combination of the two cases for different values of the latent variable z.

Let us assume that we wish to manipulate the entropy of the distribution $P(x^{(j)}|z)$. The form of the entropic prior for this distribution is defined as $e^{-\beta\mathcal{H}(P(x^{(j)}|z)} = e^{\beta\sum_i P(x_i^{(j)}|z)logP(x_i^{(j)}|z)}$, where $P(x_i^{(j)}|z)$ denotes the i'th element of the distribution $P(x^{(j)}|z)$. Incorporating the entropic prior in the PLCA model and adding the constraint that $\sum_i P(x_i^{(j)}|z) = 1$ results into optimizing the following function:

$$\frac{P^*(x^{(j)}|z)}{P(x_i^{(j)}|z)} + \beta + \beta logP(x_i^{(j)}|z) + \lambda = 0, \tag{6}$$

where $P^*(x^{(j)}|z)$ is defined in equation 4 and λ is the Langrange multiplier enforcing the unity summation constraint. As shown in [6] this equation can be solved using Lambert's $\mathcal{W}$ function resulting in:

$$P(x^{(j)}|z) = \frac{P^*(x^{(j)}|z)/\beta}{\mathcal{W}(-P^*(x^{(j)}|z)e^{1+\lambda/\beta}/\beta)}. \tag{7}$$

Alternating between the last two equations for a couple of iterations we can obtain a refined estimate of $P(x^{(j)}|z)$ which accommodates the entropy constraint. This process is described in more detail in [7].

2 Applications of PLCA for Source Separation

The two separation scenarios we will introduce in the next sections are both making use of PLCA models of sounds. We will now briefly introduce how we can model a class of sounds using PLCA. One major feature that we can use to describe a sound is that of its frequency distribution. For example we know that speech tends to have a harmonic distribution with most energy towards the low end of the spectrum, whereas, say, a siren would have a more simple timbral profile mostly present at higher frequencies. We can use the PLCA model to obtain a dictionary of spectral profiles that best describe a class of sounds. To do so we consider the 2-d formulation of PLCA when applied on time-frequency distributions of sounds. The model will be:

$$P(f,t) = \sum_z P(z)P(f|z)P(t|z) \tag{8}$$

where $P(f,t)$ is a magnitude spectrogram. The decomposition will result into two sets of marginals, one for the frequency axis and one for the time axis. The time axis marginals are not particularly informative, the frequency axis marginals however will contain a dictionary of spectra which best describe the sound represented by the input spectrogram. To illustrate this operation consider the spectrograms in figure 2.1 and their corresponding frequency marginals. One can easily see that the extracted marginals are latching on to the specific spectral structure of each sound. These frequency marginals can be used as a model of a class of sounds such as human voice, speech of a specific speaker, a specific type of background noise, etc.

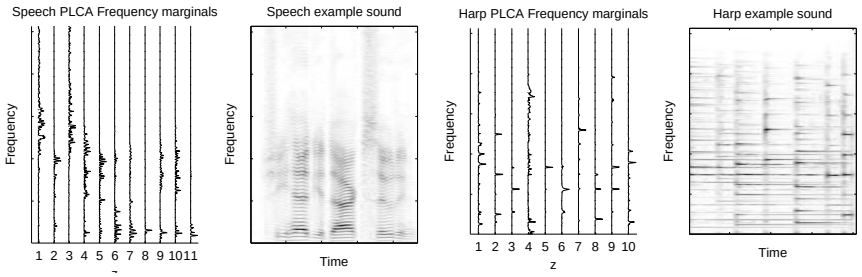

Fig. 2.1. Example of PLCA models of two different sounds. The two left plots display a spectrogram of speech and a set of speech-derived frequency marginals. Likewise the two right plots display the same information for a harp sound. Note how the derived marginals in both cases extract representative spectra for each sound.

The the next sections we describe how this model can be used for a supervised and semi-supervised source separation.

2.1 Supervised Separation

In the case of supervised separation we assume that the mixture we are operating on contains classes of sounds for which we have already trained PLCA models (in the form of frequency marginals, as described above). If the kind of time-frequency distribution that we use is (at least approximately) linearly additive in nature, we can assume that the marginal distributions of our trained models can be used to approximate the mixture's distribution. For the experiments presented in this paper we employ the magnitude short time Fourier transform. Although the linearity assumption does not exactly apply for this transform, it is sufficiently approximately correct in the context of sound mixtures.

In order to perform the separation let us consider a mixture composed out of samples from the two sound classes analyzed in figure 2.1. Let us denote the already known frequency marginals from these two sounds as $P_1(f|z)$ and $P_2(f|z)$. The spectrogram of the mixture, which we denote by $P(f,t)$, is shown in figure 2.2. One can easily see elements of both sounds present in it. Once we obtain the spectrogram of the mixture we need to find how to use the already known marginals from prior analysis to approximate it. Doing so it a very simple operation which involves partial use of the training procedure shown above. First we consolidate the marginals of the known sounds into one set $P(f|z) = \{P_1(f|z) \bigcup P_2(f|z)\}$. Since all the of the marginals in $P(f|z)$ should explain the mixture spectrogram $P(f,t)$ we only need to estimate a set of time marginals $P(t|z)$ which will facilitate the approximation. We therefore perform the training outlined in the previous sections, only this time we only estimate $P(t|z)$ and keep $P(f|z)$ fixed to the already known values. After we obtain a satisfactory estimate of $P(t|z)$ we appropriately split it to two sets which correspond to each $P_i(f|t)$. We can then reconstruct the elements of the input spectrogram that correspond to only one sound class by using only the time and frequency marginals that

correspond to that sound class. The results in this particular case are shown in figure 2.2. As is evident the contribution to the mixture from each of the two sources is cleanly separated into two spectrograms. Once the spectrograms of each known sound have been recovered we can easily transform them back to the time domain by using the corresponding phase values from the original mixture.

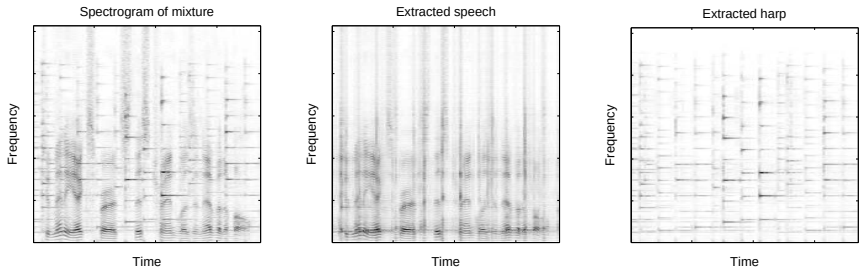

Fig. 2.2. Example of supervised separation using PLCA. The leftmost plot displays the inputs spectrogram. We can easily see features of the speech and harp sounds. The two remaining plots show the mixture spectrogram as approximated by the speech marginals (center plot), and the harp marginals (right plot).

As one might suspect this approach does not allow the separation of spectrally similar sounds since there will be significant similarity between the marginals of each sound class. The more dissimilar the sounds in the mixture are the better the quality of the separation will be. In this particular example separation was almost flawless since the two sounds had a very different spectral profile. For experiments using $0dB$ speech mixtures the target source improvement ranged from $3dB$ to $10dB$ depending on the similarity between the speakers. Using examples such as various types of ambient noise and speech we often achieved separation of more than $12dB$.

2.2 Semi-Supervised Separation

In the case of semi-supervised separation we assume that we only have a PLCA model for one of the sounds in a mixture. In this case we cannot directly use the aforementioned procedure to perform separation. In this section we introduce a methodology which deals with this problem.

Assume that we have a mixture of multiple sounds and we only have a PLCA model for one of them. We can perform PLCA on the mixture using the known frequency marginals for one of the sounds and in the process estimate additional marginals to explain the elements in the mixture we can't already. Doing so with the training procedure we have shown in the previous sections is very easy. We train as we usually do when learning both the frequency and the time marginals, but we make sure that a portion of the frequency marginals are kept fixed as

we update only the remaining ones using the same training procedure as before. The fixed marginals are the ones we already know as a model for one of the sounds. Conclusion of training will result into a set of new frequency marginals which are best suited to explain the sources in the mixture other than the one we already know. Since there will most likely be some spectral similarity between the known sound and the rest of the sources we also encourage sparsity on the time marginals to ensure that there is minimal co-occurence of frequency marginals at any time.

Once the marginals of the additional sources have been identified we can revert back to the supervised separation methodology to obtain the results we seek. The additional complication in this scenario is that by having a model of only one of the sources results into the ability to extract either that source by itself or all the other sources as one. This means that we can use this method for applications akin to denoising where we either know the target characteristics, of the background noise characteristics. In our experiments we have used this approach to separate speech from music, where the results often are very impressive [3]. A separation example is shown in figure 2.3, where a soprano is separated from a piano. We only had a model for the piano and learned the soprano model using the aforementioned methodology. The suppression of the piano was audibly flawless and the only artifact of this approach was a slight coloring of the extracted soprano voice (attributed mostly to the usage of phase of the original mixture).

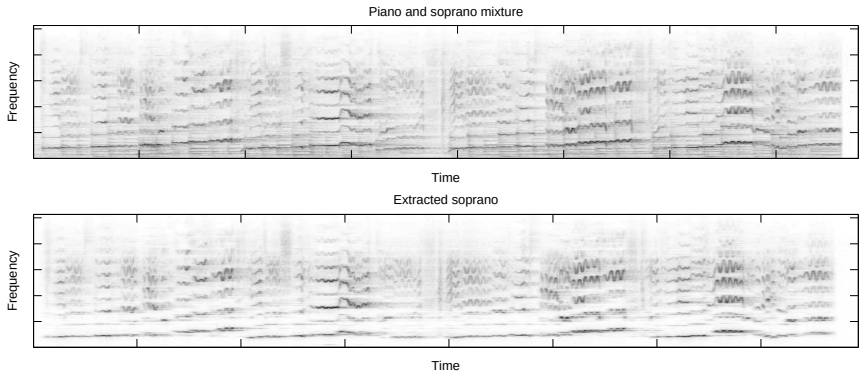

Fig. 2.3. Example of semi-supervised separation using PLCA. The left plot displays the mixture of a piano and a soprano, the right plot displays the extracted soprano voice. One can easily see that the harmonic series corresponding to the piano notes are strongly suppressed.

[3] Demonstration sound samples of this approach can be found in http://www.merl.com/people/paris/sep.html under the section "PLCA for spectral factoring"

3 Discussion

In this section we discuss the selection of parameters and their effect in separation performance and point to some of the relationships of the PLCA model to other known decompositions.

3.1 Parameter selection

In order to obtain reasonable results we have to make sure that the right parameters are used in the process. First we need to ensure that the time/frequency decomposition we employ is adequate to perform separation. In our experience a $1/10sec$ analysis window is usually a good choice for separation. As this window becomes smaller it results in inadequate frequency resolution, and as it grows larger it results in time smearing. The hop size of the transform also needs to be small enough to ensure a clean reconstruction during the transformation from time/frequency to time (a fourth of the transform size is a good choice). Applying a Hanning window for the frequency analysis is also advised since it minimizes high frequency artifacts which are not part of the sound we model and can result into a skewed representation.

The selection of the PLCA parameters is very important in order to achieve good results. In most of our simulations, sounds were modeled using around 100 marginals (i.e. $z = \{1, 2, 3, ..., 100\}$). Using a small number of marginals results into a poor representation which attains spectrally quantized results, whereas a large number of marginals results into large sets of simple marginals which can also describe elements of the interfering sounds. The tradeoff in this case is between accuracy of model versus separability of models. The sparsity parameter is something we only use in the case of the semi-supervised learning on the time marginals. It ensures that the new marginals that we learn will not overlap as much with the already known ones. Common usage values in our experiments were $\beta = \{0, 0.01, 0.05, 0.1\}$, where larger values were used in harder to separate problems where more spectral overlap between sources was present. The audible effect of using sparsity is a degradation of reconstruction of the sound quality of the source to be learned. Therefore using the sparsity parameter is best when we have a model of the target source and we wish to remove the remaining sources.

3.2 Relation to similar decompositions

The PLCA model which we introduced is closely related to a variety of known decompositions. The non-sparse 2-d manifestation is identical to the Probabilistic Latent Semantic Indexing (PLSI) algorithm [5], which itself is a probabilistic generalization of the Singular Value Decomposition. The functional difference is that PLSI/PLCA operate on distributions instead of raw data which means that they can effectively only analyze non-negative inputs. If we rewrite the 2-d PLCA model in terms of matrix operations, this relationship is more evident:

$$P(f,t) = \sum_z P(z)P(f|z)P(t|z) \equiv \mathbf{V} = \mathbf{W} \cdot \mathbf{S} \cdot \mathbf{H} \tag{9}$$

where $\mathbf{V}$ is a matrix containing the distribution $P(f,t)$, $\mathbf{W}$ is a matrix containing in its columns $P(f|z)$ for every z, $\mathbf{S}$ is a diagonal matrix containing in its diagonal the values of $P(z)$, and $\mathbf{H}$ is a matrix containing in its rows $P(t|z)$ for every z.

Additionally if we absorb the values of $\mathbf{S}$ into the two matrices $\mathbf{W}$ and $\mathbf{H}$ so that: $\mathbf{V} = \mathbf{W} \cdot \mathbf{S} \cdot \mathbf{H} = \bar{\mathbf{W}} \cdot \bar{\mathbf{H}}$, we can make a connection to the Non-negative Matrix Factorization (NMF) decomposition [8]. NMF can employ the Kullback-Leibler divergence to measure how well the factorization $\bar{\mathbf{W}} \cdot \bar{\mathbf{H}}$ approximates the input $\mathbf{V}$. The EM training which we perform also indirectly optimizes the same cost function as it improves the model's log-likelihood. In fact the two training procedures for 2-d PLCA and NMF can be shown to be numerically identical.

Finally we can make a loose connection to non-negative ICA by noting that by using the entropic prior to manipulate the joint entropy of $\mathbf{H}$ we can obtain the equivalent of an ICA mixing matrix in $\mathbf{W}$. Although this is only a conjecture on our part, preliminary results from simulations are encouraging.

4 Conclusions

In this document we introduced a sparse latent variable model which can be employed for the decomposition of time/frequency distributions to perform separation of sources from monophonic recordings. We demonstrated the use of this model for both supervised and semi-supervised source separation, and discussed its relationship with other known decompositions. Our results are very encouraging and amenable to various modifications, such as the use of convolutive bases and transformation invariance, which can help to successfully apply this work to even more challenging source separation problems.

References

1. Casey, M. and Westner, A. "Separation of Mixed Audio Sources by Independent Subspace Analysis" in proceedings *ICMC* 2000.
2. Roweis, S.T. "One Microphone Source Separation" in *NIPS* 2000.
3. Benaroya, L., McDonagh, L, Bimbot F. and Gribonval, R. "Non negative sparse representation for Wiener based source separation with a single sensor" in proceedings of the *ICASSP* 2003.
4. Vincent, E. and Rodet, X. "Music transcription with ISA and HMM" in proceedings of *ICA* 2004.
5. Hofmann, T. "Probabilistic Latent Semantic Indexing" in proceedings *SIGIR'99*
6. Brand, M.E., "Structure Learning in Conditional Probability Models via an Entropic Prior and Parameter Extinction" in *Neural Computation Journal*, Vol. 11, No. 5, pp. 1155-1182, July 1999.
7. Shashanka, M.V.S., "A Unified Probabilistic Approach to Modeling and Separating Single-Channel Acoustic Sources", Ph.D. Thesis, Department of Cognitive and Neural Systems. Boston University, Boston 2007.
8. Lee, D.D and Seung, H.S. "Algorithms for Non-negative Matrix Factorization", in *NIPS* 2001.