Topic Models for Audio Mixture Analysis

Paris Smaragdis
Adobe Systems Inc.
paris@adobe.com

Madhusudana Shashanka
Mars Corporation
shashanka@alum.bu.edu

Bhiksha Raj
Carnegie Mellon University
bhiksha@cs.cmu.edu

1 Introduction

Sound consists of pressure variations propagating through a medium such as air. The common digital representation of an acoustic signal is the sampled waveform, where each sample represents the sound pressure level at a particular time instant. However, sounds that as human listeners we find meaningful are best represented in the *time-frequency domain*.

In contrast to the time-domain waveform, time-frequency representations explicitly represent the time-varying frequency content of a sound and effectively visualize the signal's activity at any time-frequency bin. These transforms are often complex-valued and include well known tools such as the short-time Fourier transform, constant-Q transforms, wavelets, etc. However, because our hearing system is more sensitive to the relative energy between different frequencies, for most practical applications we study the modulus of these transforms and discard the phase which is useful only in special cases. These kinds of representations are essentially counting the number of time-frequency acoustic quanta that collectively make up complex sound scenes, similar to how we count words that make up documents. With this representation we can use the analogy of *bag of frequencies*, which we describe later in this document.

Historically, audio research (speech/sound recognition, pitch tracking, etc) has been focused on analyzing sounds that have been isolated. But as interest has shifted towards more ambitious goals such as scene analysis, and also more applied goals, such as speech recognition in noisy environments, the problem of performing audio analysis on mixtures of sounds has become increasingly prevalent. The traditional signal processing framework has not been adequate for these problems, but topic-like models that operate on the time-frequency domain have been proven to be exceptional performers for this job. This is for two predominant reasons. First audio researchers have generally come to accept that the modulus of a time-frequency representation of a mixture of two sounds is approximately equal to the modulus of the two sounds isolated. This allows us to talk about mixtures of sounds as being mixtures of topics. Second, from a perspective of human perception a mixture of topics model is more in tune with how we interpret scenes. Unlike related models like PCA, ICA, etc, which use cross-cancellations to describe a mixture, topic models decompose mixtures in additive terms which best map to our perceptual interpretation of an auditory scene. This is because we never think of what sounds are missing from a auditory scene, but only of what sounds are added in it.

In the following sections we highlight some of the applications of topic models in audio analysis. We describe models that have been successfully used for unsupervised music transcription, source separation, scene analysis, denoising, missing audio data imputation, and auditory user interfaces. These applications are all based on topic models as applied on time-frequency data and present an exciting new direction of research that has been recently very active, but still largely disjoint from the general topic-modeling literature. We hope that this presentation will introduce some of the

main topic model approaches in the audio world and foster more cross-pollination between these two communities.

2 Bag of Frequencies

In this section we consider some applications of a bag of frequencies approach. In contrast to the use of a word count matrix in which each input element represents the frequency of occurrence of a word in a document, we use the time-frequency representation which provides a sense of the occurrence of each frequency at each point in time. So instead of a $\{\text{word} \times \text{document}\}$ matrix we use a $\{\text{frequency} \times \text{time}\}$ matrix. The objective is to use this representation to carry out various tasks. The basic model we start with is equivalent to PLSA where the spectral frame at time t is modeled as the result of repeated draws of frequencies f from latent multinomial distributions $P(f|z)$, and can be written as:

$$P_t(f) = \sum_z P_t(z)P(f|z), \tag{1}$$

where $P_t(f)$ represents the probability of observing frequency f in the time frame t, $P(f|z)$ represents the probability of observing frequency f given the latent variable z, and $P_t(z)$ represents the probability of z in the t-th frame.

2.1 Modeling a mixture of sources

When modeling mixtures in terms of known "acoustic topics" we can reformulate the basic model as:

$$P_t(f) = \sum_s P_t(s) \sum_{z \in \{\mathbf{z}_s\}} P_t(z|s)P_s(f|z) \tag{2}$$

where latent variable s represents the constituent sources, $P_t(s)$ is the probability of observing source s in time frame t, and $\{\mathbf{z}_s\}$ represents the set of values that z can take for that source (or "topic"). In a typical source separation scenario, we would assume that we know the types of sources that compose a mixture (e.g. the speaker and the background noise), and use regular PLSA modeling to learn their frequency dictionaries $P_s(f|z)$. Once these are known we can use them to estimate the $P_t(z|s)$ and $P_t(s)$ of a given mixture and thus segment the time-frequency distribution to the two constituent sources. This approach has been used successfully to separate sounds in monophonic mixtures, and to obtain some of the state-of-the-art results. In the case where we know one source but not the rest we can modify this procedure so that we estimate $P_t(z|s)$, $P_t(s)$ and only the $P_s(f|z)$ of the sources that we do not know. This allows us to design source separators which can operate given only a known target, or only a known background model. These algorithms are described in [1].

2.2 Entropic priors

It was observed that the use of the above approach did result in modeling the data using a dense mixture of multinomials which often resulted in additional noise in the separations. In order to address this issue we developed a sparse variant of the above model [2] by employing the entropic prior proposed in [3]. Using this prior we can manipulate the entropy of any of the estimated distributions. The most practical case in the problem at hand is using the entropic prior to obtain minimal entropy frequency bag activations $P_t(z|s)$. This results in a sparser representation of a mixture and consequently a cleaner quality of separation. As shown in [4] this approach also helps learn overcomplete models based solely on the training data without requiring extracting a dictionary.

2.3 Use of source and relational priors

The use of priors in the topic modeling community is well known and has resulted in a lot of interesting work. Likewise in the acoustic domain we can an LDA-type analysis [5] in order to enforce a particular belief on the bag of frequency shape, activations, and the source priors. Use of this was done in [6] and was shown to improve the separation results. Another kind of prior that was used was a relational prior which biased the estimated distributions to be either more related to each other or not, by optimizing their cross-entropy [7]. This allowed us to find groups of components that are related to each other, but unrelated to other groups.

2.4 Missing data

Once we have a PLSA model of a sound's time-frequency structure, we can also use it to perform data imputation. In [8] we present an iterative algorithm that recovers missing values in time-frequency representations of sounds. Using either training data, or even the missing data input itself, we can derive a model which is then used to find the most likely values in the missing time-frequency cells. Using this approach we can patch holes that can be caused by user editing of sound subtraction, but also upsample signals and recover frequencies that have been lost.

2.5 Scene analysis and object extraction

PLSA modeling has also been very useful for discovering objects in auditory scenes. A well-known example is the case of discovering that music is made out of notes by examining piano recordings. This was first reported using NMF in [9], but as shown in [10, 11] the process is equivalent to a PLSA model. In this example a topic analysis of a time-frequency decomposition of piano music resulted in bag of frequencies which each individually corresponded to a single piano note. Likewise their activation weights showed when in time these notes were active. This has by now become a dominant model in music transcription because of its flexibility and lack of reliance on complex (yet usually narrow) models of musical sound.

3 Bag of Spectrograms

The base model used in the previous section is a straightforward PLSA model. However in audio we often observe shift-invariance in both time and frequency and thus we extended the basic formulation to incorporate this feature. In the process we created a *bag of spectrograms* model which we describe below. This model can discover topics which have not only a shiftable-frequency structure, but also a shiftable temporal structure. This allows us to make models that unlike the ones in the previous section use more of the temporal statistics and respect the input's time order.

3.1 Convolutive Models

The shift-invariant version of PLSA was shown in [12]. In this case instead of modeling each column of the input as a sum of multinomial distributions, we model each patch of the input as a sum of shifted two-dimensional time-frequency distributions. This enforces structure in both time and frequency as opposed to only the frequency dimension that we see in PLSA. The full model in that case is defined as:

$$P(\mathbf{x}) = \sum_{z} \left(P(z) \int P(\mathbf{w}, \boldsymbol{\tau}|z) P(\mathbf{h} - \boldsymbol{\tau}|z) d\boldsymbol{\tau} \right) \tag{3}$$

where $\mathbf{w}$ and $\mathbf{h}$ are mutually exclusive subsets of components, $\mathbf{w} = \{x_i\}$, $\mathbf{h} = \{x_i\}$, such that $\mathbf{x} = \{\mathbf{w}, \mathbf{h}\}$. $\boldsymbol{\tau}$ is a random variable that is defined over the same domain as $\mathbf{h}$. By using convolution, the now multi-dimensional topics $P(\mathbf{w}, \boldsymbol{\tau}|z)$ can shift around the space of the input in both frequency and time. This allows us to model shift-invariance which in the audio domain is often seen in both the frequency and the time axes. Interestingly enough if we remove the latent variable and consider the rank-1 decomposition this reduces to the Lucy-Richardson [13] deconvolution operation.

Because the convolution operator in the above model is commutative there exists an ambiguity on the structure of $\mathbf{w}$ and $\mathbf{h}$, and the way to resolve this is to use the entropic prior shown above. Most usually we require that the activations $P(\mathbf{h} - \boldsymbol{\tau}|z)$ of the estimated patches $P(\mathbf{w}, \boldsymbol{\tau}|z)$ are maximally sparse, so that we discover a shift-invariant sparse code. This model can be used to repeat many of the operations shown using regular PLSA, but it predominantly most useful at extracting information.

3.2 Object discovery

Performing the above decomposition on long recordings of sounds we can find repeating patches of sounds that seem to be statistically significant. A particularly interesting case of this is performing this analysis on speech recordings trying to discover the building elements of speech. If we request

sparse activation codes the resulting patches $P(\mathbf{w}, \boldsymbol{\tau}|z)$ end up resembling phonemes or phones, which we accept to be the building blocks of speech. Likewise from music recordings we can recover entire notes or instrument phrases.

3.3 Musical priors

The frequency invariance is also very useful for musical applications. Using constant-Q tie-frequency representations we ensure that a change in pitch of an instrument will result in only a vertical shift of an otherwise constant bag of frequencies. As shown in [14] this can be used to extract an instrument's spectral character from a recording while at the same time estimating its pitch. In addition to that we can impose spectral shape and temporal continuity priors [12] to ensure that the pitch tracking is appropriately constrained to fit the instrument we wish to track.

3.4 Dereverb and echo canceling

FInally, as will be shown in upcoming publications the convolutive model allows us to perform operations such as dereverberation and echo cancelation by attempting to remove the repetitions of a "series of topics" using this convolutive formulation. This probabilistic reasoning for these operations is a new way of thinking about signals and results in the opportunity to perform exciting extensions and find new ways of treating some of the oldest problems in signal processing.

4 Conclusions

In this abstract we have tried to communicate some of the applications that topic models have found in audio and music processing. Unfortunately a presentation on paper is not as vivid as playing example sounds which show how powerful these techniques can be, but we hope it has given the flavor of some of the exciting new directions audio processing is moving towards.

References

[1] P Smaragdis, B Raj, and MV Shashanka. Supervised and Semi-Supervised Separation of Sounds from Single-Channel Mixtures. In ICA 2007.

[2] MVS Shashanka, B Raj, and P Smaragdis. Sparse overcomplete latent variable decomposition of counts data. In NIPS 2007.

[3] ME Brand. Pattern discovery via entropy minimization. In Uncertainty 99: AISTATS 99, 1999.

[4] P Smaragdis, M Shashanka, and B Raj. A sparse non-parametric approach for single channel separation of known sounds. In NIPS 2009.

[5] DM Blei, AY Ng, and MI Jordan. Latent Dirichlet allocation. JMLR, 3:9931022, 2003.

[6] B Raj, MVS Shashanka, and P Smaragdis. Latent Dirichlet decomposition for single channel speaker separation. In ICASSP 2006.

[7] P Smaragdis, M Shashanka, B Raj, and G Mysore. Probabilistic factorization of non-negative data with entropic co-occurrence constraints. In ICA 2009.

[8] P Smaragdis, B Raj, MV Shashanka, Missing data imputation for spectral audio signals. In MLSP 2009.

[9] P Smaragdis and JC Brown. Non-negative matrix factorization for polyphonic music transcription. In WASPAA 2003.

[10] E Gaussier and C Goutte. Relation between PLSA and NMF and implications. In Proc. ACM SIGIR Conf. on Research and Dev. in Information Retrieval, pages 601602, 2005.

[11] M Shashanka, P Smaragdis, and B Raj. Probabilistic latent variable models as nonnegative factorizations. Computational Intelligence and Neuroscience, 2008.

[12] P Smaragdis, B Raj, and MVS Shashanka. Sparse and shift-invariant feature extraction from non-negative data. In ICASSP 2008.

[13] WH Richardson. "Bayesian-Based Iterative Method of Image Restoration". JOSA 62 (1): 5559. 1972.

[14] P Smaragdis. Relative Pitch Tracking of Multiple Arbitrary Sounds. In JASA, Volume 125, Issue 5, pp. 3406-3413, May 2009.

[15] G Mysore and P Smaragdis. Relative Pitch Estimation of Multiple Instruments, In ICASSP 2009.