Non-negative Hidden Markov Modeling of Audio with Application to Source Separation

Gautham J. Mysore[1⋆], Paris Smaragdis[2], and Bhiksha Raj[3]

[1] Center for Computer Research in Music and Acoustics, Stanford University,
[2] Advanced Technology Labs, Adobe Systems Inc.,
[3] School of Computer Science, Carnegie Mellon University

Abstract. In recent years, there has been a great deal of work in modeling audio using non-negative matrix factorization and its probabilistic counterparts as they yield rich models that are very useful for source separation and automatic music transcription. Given a sound source, these algorithms learn a dictionary of spectral vectors to best explain it. This dictionary is however learned in a manner that disregards a very important aspect of sound, its temporal structure. We propose a novel algorithm, the non-negative hidden Markov model (N-HMM), that extends the aforementioned models by jointly learning several small spectral dictionaries as well as a Markov chain that describes the structure of changes between these dictionaries. We also extend this algorithm to the non-negative factorial hidden Markov model (N-FHMM) to model sound mixtures, and demonstrate that it yields superior performance in single channel source separation tasks.

1 Introduction

A common theme in most good strategies to modeling audio is the ability to make use of structure. Non-negative factorizations such as non-negative matrix factorization (NMF) and probabilistic latent component analysis (PLCA) have been shown to be powerful in representing spectra as a linear combination of vectors from a dictionary [1]. Such models take advantage of the inherent low-rank nature of magnitude spectrograms to provide compact and informative descriptions. Hidden Markov models (HMMs) have instead made use of the inherent temporal structure of audio and have shown to be particularly powerful in modeling sounds in which temporal structure is important, such as speech [2]. In this work, we propose a new model that combines the rich spectral representative power of non-negative factorizations and the temporal structure modeling of HMMs.

In [3], ideas from non-negative factorizations and HMMs have been used by representing sound mixtures as a linear combination of spectral vectors and also modeling the temporal structure of each source. However, at a given time frame, each source is represented by a single spectral vector rather than a linear combination of multiple spectral vectors. As pointed out, this has some virtue in speech as it is monophonic but it can break down when representing rich polyphonic sources such as music, for which one would resort to using standard NMF. In our

⋆ This work was performed while interning at Adobe Systems Inc.

proposed method, at a given time frame, a given source is represented as a linear combination of multiple spectral vectors from one (of the many) dictionaries of the source. This allows us to model finer details in an input such as variations in a phoneme or a note. As shown in the results section, the performance improves even for speech.

2 Models of Single Sources

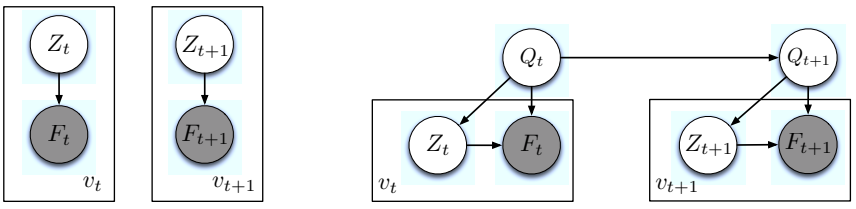
In this section, we first briefly describe probabilistic spectrogram factorization for modeling a single source. We then describe the non-negative hidden Markov model (N-HMM) and parameter estimation for the model.

2.1 Probabilistic Spectrogram Factorization

The magnitude spectrogram of a sound source can be viewed as a histogram of "sound quanta" across time and frequency. With this view, probabilistic factorization [4], which is a type of non-negative factorization, has been used to model a magnitude spectrogram as a linear combination of spectral vectors from a dictionary. The model is defined by two sets of parameters:

1. $P(f|z)$ is a multinomial distribution of frequencies for latent component z. It can be viewed as a spectral vector from a dictionary.
2. $P(z_t)$ is a multinomial distribution of weights for the aforementioned dictionary elements at time t.

Given a magnitude spectrogram, these parameters can be jointly estimated using the Expectation–Maximization (EM) algorithm. As can be seen in the graphical model representation in Fig. 1a, the weights of each time frame are estimated independently of the other time frames, therefore failing to capture the temporal structure of the sound source.



(a) Factorization (b) N-HMM

Fig. 1: Probabilistic factorization models each time frame independently, whereas the N-HMM models the transitions between successive time frames.

2.2 Non-negative Hidden Markov Model

There is a temporal aspect to the proposed model, the N-HMM, shown in Fig. 1b. The model has a number of states, q, which can be interpreted as individual dictionaries. Each dictionary has a number of latent components, z, which can be interpreted as spectral vectors from the given dictionary. The spectral vector z of state q is defined by the multinomial distribution, $P(f|z,q)$.

As in traditional HMMs, in a given time frame, only one of the states is active. The given magnitude spectrogram at that time frame is modeled as a linear combination of the spectral vectors of the corresponding dictionary (state), q. At time t, the weights are defined by the multinomial distribution $P(z_t|q_t)$.

This notion of modeling a given time frame with one (of many) dictionaries rather than using a single large dictionary globally caters well to the non-stationarity of audio signals. The idea is that as an audio signal dynamically changes towards a new state, a new and appropriate dictionary should be used. We capture the temporal structure of these changes with a transition matrix, defined by, $P(q_{t+1}|q_t)$. The initial state probabilities (priors) are defined by $P(q_1)$. We also define a distribution $P(v|q)$ which is a distribution of the energy of a given state. It is modeled as a Gaussian distribution. It has been left out of the graphical model for clarity. The overall generative process is as follows:

1. Set $t = 1$ and choose a state according to the initial state distribution $P(q_1)$.
2. Choose the number of draws (energy) for the given time frame according to $P(v_t|q_t)$.
3. Repeat the following steps v_t times:
   (a) Choose a latent component according to $P(z_t|q_t)$.
   (b) Choose a frequency according to $P(f_t|z_t,q_t)$.
4. Transit to a new state q_{t+1} according to $P(q_{t+1}|q_t)$
5. Set $t = t+1$ and go to step 2 if $t < T$.

2.3 Parameter Estimation for the N-HMM

Given the scaled magnitude spectrogram, V_{ft}, of a sound source[4], we use the EM algorithm to estimate the model parameters of the N-HMM. The E-step is computed as follows:

$$P(z_t, q_t|f_t, \overline{\mathbf{f}}, \overline{\mathbf{v}}) = \frac{\alpha(q_t)\beta(q_t)}{\sum_{q_t}\alpha(q_t)\beta(q_t)} P(z_t|f_t,q_t)\,, \tag{1}$$

where

$$P(z_t|f_t,q_t) = \frac{P(z_t|q_t)P(f_t|z_t,q_t)}{\sum_{z_t}P(z_t|q_t)P(f_t|z_t,q_t)}\,. \tag{2}$$

$P(q_t, z_t|f_t, \overline{\mathbf{f}}, \overline{\mathbf{v}})$ is the posterior distribution that is used to estimate the dictionary elements and the weights vectors. $\overline{\mathbf{f}}$ denotes the observations across all time frames[5], which is the entire spectrogram. $\overline{\mathbf{v}}$ denotes the number of draws over all

[4] Since the magnitude spectrogram is modeled as a histogram, the entries should be integers. To account for this, we weight it by an appropriate scaling factor.

[5] It should be noted that f_t is part of $\overline{\mathbf{f}}$. It is however mentioned separately to indicate that the posterior over z_t and q_t is computed separately for each f_t.

time frames. The forward/backward variables $\alpha(q_t)$ and $\beta(q_t)$ are computed using the likelihoods of the data, $P(\mathbf{f_t}, v_t|q_t)$, for each state (as in standard HMMs [2]). The likelihoods are computed as follows:

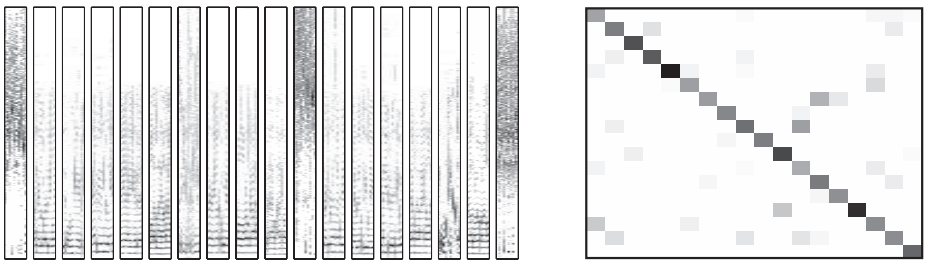
$$P(\mathbf{f_t}, v_t|q_t) = P(v_t|q_t) \prod_{f_t} \left(\sum_{z_t} P(f_t|z_t, q_t) P(z_t|q_t) \right)^{V_{ft}}, \tag{3}$$

where $\mathbf{f_t}$ represents the observations at at time t, which is the magnitude spectrum at that time frame. The dictionary elements and their weights are estimated in the M-step as follows:

$$P(f|z, q) = \frac{\sum_t V_{ft} P(z_t, q_t|f_t, \overline{\mathbf{f}}, \overline{\mathbf{v}})}{\sum_{f_t} \sum_t V_{ft} P(z_t, q_t|f_t, \overline{\mathbf{f}}, \overline{\mathbf{v}}))}, \tag{4}$$

$$P(z_t|q_t) = \frac{\sum_{f_t} V_{ft} P(z_t, q_t|f_t, \overline{\mathbf{f}}, \overline{\mathbf{v}}))}{\sum_{z_t} \sum_{f_t} V_{ft} P(z_t, q_t|f_t, \overline{\mathbf{f}}, \overline{\mathbf{v}}))}. \tag{5}$$

The transition matrix, $P(q_{t+1}|q_t)$, and priors, $P(q_1)$, are computed exactly as in standard HMMs [2]. The mean and variance of $P(v|q)$ are also estimated from the data. The learned dictionaries and transition matrix for an instance of speech data can be seen in Fig. 2. This model can be interpreted as an HMM in which



(a) Learned dictionaries (b) Transition matrix

Fig. 2: (a) Dictionaries were learned from speech data of a single speaker. Shown are the dictionaries for 18 states, each state dictionary comprised of 10 elements. Each of these dictionaries roughly corresponds to a subunit of speech, either a voiced or unvoiced phoneme. (b) The learned transition matrix describes the transitions between the learned dictionaries in Fig. 2a. As can be seen by the strong diagonal, the algorithm correctly learns a model with state persistence.

the observation model $P(f_t|q_t)$ is a multinomial mixture model:

$$P(f_t|q_t) = \sum_{z_t} P(f_t|z_t, q_t) P(z_t|q_t). \tag{6}$$

However, this implies that for a given state, q, there is a single set of spectral vectors $P(f|z, q)$ and a single set of weights $P(z|q)$. If the weights did not change

across time, the observation model would collapse to a single spectral vector per state. In the proposed model however, the weights $P(z_t|q_t)$ change with time. This flexible observation model allows us to model variations in the occurrences of a given state. This idea has previously been explored for Gaussian mixture models [5]. It should be noted that the proposed model collapses to a regular non-negative factorization if we use only a single state, therefore only one dictionary.

3 Model for Sound Mixtures

In this section, we describe the non-negative factorial hidden Markov model (N-FHMM) for modeling sound mixtures. We then describe how to perform source separation using the model.

As shown in the two-source graphical model in Fig. 3, the N-FHMM combines multiple N-HMMs of single sources. The interaction model introduces a new variable s_t that indicates the source. In the generative process, for each draw of each time frame, we first choose the source and then choose the latent component as before. In order to perform separation, we use trained models of

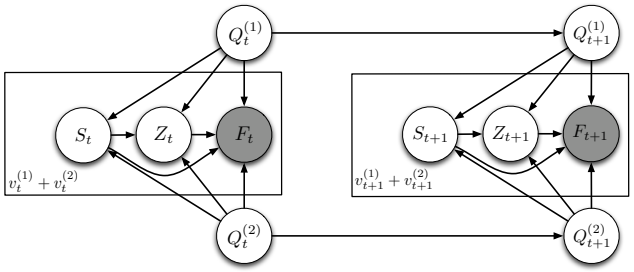


Fig. 3: N-FHMM. The structure of two individual N-HMMs (one in the upper half and one in the lower half) can be seen in this model.

individual sources. We train an N-HMM and learn the dictionaries and the transition matrix for each class of sound we expect to encounter in a mixture. We then use the a priori source information in these trained models to resolve mixtures that involve such sources. The dictionaries and transition matrices of the N-FHMM will therefore already be defined, and one will only need to estimate the appropriate weights from the mixture.

In a given time frame t, each source is explained by one of its dictionaries. Therefore, a given mixture is modeled by a pair of dictionaries, $\{q_t^{(1)}, q_t^{(2)}\}$, one for each source (superscripts indicate the source). For a given pair of dictionaries, the mixture spectrum is defined by the following interaction model:

$$P(f_t|q_t^{(1)}, q_t^{(2)}) = \sum_{s_t} \sum_{z_t} P(f_t|z_t, s_t, q_t^{(s_t)}) P(z_t, s_t|q_t^{(1)}, q_t^{(2)}) \,. \tag{7}$$

As can be seen, the mixture spectrum is modeled as a linear combination of the individual sources which are in turn modeled as a linear combination of spectral

vectors from the given dictionaries. This allows us to model the mixture as a linear combination of the spectral vectors from the given pair of dictionaries [6].

3.1 Source Separation

In order to perform separation, we need to first estimate the mixture weights, $P(z_t, s_t|q_t^{(1)}, q_t^{(2)})$ for each pair of states. That can be done using the EM algorithm. The E-step is computed as follows:

$$P(z_t, s_t, q_t^{(1)}, q_t^{(2)}|f_t, \overline{\mathbf{f}}, \overline{\mathbf{v}}) = \frac{\alpha(q_t^{(1)}, q_t^{(2)})\beta(q_t^{(1)}, q_t^{(2)})}{\sum_{q_t^{(1)}} \sum_{q_t^{(2)}} \alpha(q_t^{(1)}, q_t^{(2)})\beta(q_t^{(1)}, q_t^{(2)})} P(z_t, s_t|f_t, q_t^{(1)}, q_t^{(2)}) \,, \tag{8}$$

where

$$P(z_t, s_t|f_t, q_t^{(1)}, q_t^{(2)}) = \frac{P(f|z_t, s_t, q_t^{(s_t)})P(z_t, s_t|q_t^{(1)}, q_t^{(2)})}{\sum_{s_t} \sum_{z_t} P(f|z_t, s_t, q_t^{(s_t)})P(z_t, s_t|q_t^{(1)}, q_t^{(2)})} \,. \tag{9}$$

$\alpha(q_t^{(1)}, q_t^{(2)})$ and $\beta(q_t^{(1)}, q_t^{(2)})$ are computed with a two dimensional forward–backward algorithm [6] using the likelihoods of the data, $P(\mathbf{f_t}, v_t|q_t^{(1)}, q_t^{(2)})$, for each pair of states. The likelihoods are computed as follows:

$$P(\mathbf{f_t}, v_t|q_t^{(1)}, q_t^{(2)}) = P(v_t|q_t^{(1)}, q_t^{(2)}) \prod_{f_t} \left(\sum_{s_t} \sum_{z_t} P(f_t|z_t, s_t, q_t^{(s_t)})P(z_t, s_t|q_t^{(1)}, q_t^{(2)}) \right)^{V_{ft}} \,. \tag{10}$$

The weights are computed in the M-step as follows:

$$P(z_t, s_t|q_t^{(1)}, q_t^{(2)}) = \frac{\sum_{f_t} V_{ft} P(z_t, s_t, q_t^{(1)}, q_t^{(2)}|f_t, \overline{\mathbf{f}}, \overline{\mathbf{v}})}{\sum_{s_t} \sum_{z_t} \sum_{f_t} V_{ft} P(z_t, s_t, q_t^{(1)}, q_t^{(2)}|f_t, \overline{\mathbf{f}}, \overline{\mathbf{v}})} \,. \tag{11}$$

Once we estimate the weights using the EM algorithm, we compute the proportion of the contribution of each source at each time–frequency bin as follows:

$$P(s_t|f_t, \overline{\mathbf{f}}, \overline{\mathbf{v}}) = \frac{\sum_{q_t^{(1)}} \sum_{q_t^{(2)}} P(q_t^{(1)}, q_t^{(2)}|\overline{\mathbf{f}}, \overline{\mathbf{v}}) \sum_{z_t} P(f|z_t, s, q_t^{(s)})P(z_t, s|q_t^{(1)}, q_t^{(2)})}{\sum_{s_t} \sum_{q_t^{(1)}} \sum_{q_t^{(2)}} P(q_t^{(1)}, q_t^{(2)}|\overline{\mathbf{f}}, \overline{\mathbf{v}}) \sum_{z_t} P(f|z_t, s, q_t^{(s)})P(z_t, s|q_t^{(1)}, q_t^{(2)})} \,, \tag{12}$$

where

$$P(q_t^{(1)}, q_t^{(2)}|\overline{\mathbf{f}}, \overline{\mathbf{v}}) = \frac{\alpha(q_t^{(1)}, q_t^{(2)})\beta(q_t^{(1)}, q_t^{(2)})}{\sum_{q_t^{(1)}} \sum_{q_t^{(2)}} \alpha(q_t^{(1)}, q_t^{(2)})\beta(q_t^{(1)}, q_t^{(2)})} \,. \tag{13}$$

This effectively gives us a soft mask with which we modulate the mixture spectrogram to obtain the separated spectrograms of the individual sources. In Eq.

[6] We deal with $P(z_t, s_t|q_t^{(1)}, q_t^{(2)})$ rather than dealing with $P(z_t|s_t, q_t^{(1)}, q_t^{(2)})$ and $P(s_t|q_t^{(1)}, q_t^{(2)})$ individually (as shown in the graphical model) so that we will have a single set of mixture weights over both sources.

12, we sum the contribution of every pair of states. This implies that the reconstruction of each source has contributions from each of its dictionaries. However, in practice, $P(q_t^{(1)}, q_t^{(2)}|\overline{\mathbf{f}}, \overline{\mathbf{v}})$ tends to zero for all but one $\{q_t^{(1)}, q_t^{(2)}\}$ pair, effectively using only one dictionary per time frame per source. This happens because the dictionaries of individual source models are learned in such a way that each time frame is explained almost exclusively by one dictionary. The provision of having a small contribution from more than one dictionary is sometimes helpful in modeling the decay of the active dictionary in the previous time frame.

4 Experimental Results

We performed speech separation experiments on data from the TIMIT database. Specifically, we performed separation on eight pairs of speakers. Each speaker pair consists of one male and one female speaker. We first used nine sentences of each speaker as training data and learned individual N-HMM model parameters as described in Sec. 2.3. Specifically, for each speaker, we first obtained a spectrogram with a window size of 1024 and a hop size of 256 (at F_s=16,000). We then learned a model of the spectrogram with 40 dictionaries with 10 latent components each (K=10). We then repeated the experiment with 1 latent component per dictionary (K=1). After training, we combined the models into a joint model as described in Sec. 3. We constructed test data by artificially mixing one unseen sentence from each speaker at 0dB and performed separation[7]. The separation yields estimated magnitude spectrograms for each source. We used the phase of the mixture and resynthesized each source.

As a comparison, we then performed the same experiments using a traditional non-negative factorization approach. The experimental procedure as well as the training and test data are the same as above. After thorough testing, we found that the optimal results were obtained in the non-negative factorization approach by using 30 components per speaker and we therefore used this for the comparison to the proposed model. The separation performance increases up to using 30 components. When more components are used, the dictionary of one source starts to explain the other source and the separation performance goes down. It should be noted that this is equivalent to using the proposed models with 1 dictionary of 30 components per speaker.

We evaluated the separation performance in terms of the metrics defined in [7]. The averaged results over the eight pairs of speakers are as follows:

	SDR (dB)	SIR (dB)	SAR (dB)
N-FHMM (K=10)	6.49	14.07	7.74
N-FHMM (K=1)	5.58	12.07	7.26
Factorization	4.82	8.65	7.95

As shown in the table, the performance of the N-FHMM (by all metrics) is better when we use 10 components rather than 1 component. This shows the need to use a dictionary to model each state rather than a single component. We found no appreciable improvement in performance by using more than 10 components per dictionary. We see an improvement over factorizations in the overall performance

[7] Examples at `https://ccrma.stanford.edu/~gautham/Site/lva_ica_2010.html`

of the N-FHMM (SDR). Specifically, we see a large improvement in the actual suppression of the unwanted source (SIR). We however see a small increase in the introduced artifacts (SAR). The results intuitively make sense. The N-FHMM performs better in suppressing the competing source by enforcing a reasonable temporal arrangement of each speaker's dictionary elements, therefore not simultaneously using dictionary elements that can describe both speakers. On the other hand, this exclusive usage of smaller dictionaries doesn't allow us to model the source as well as we would otherwise (with 1 component per dictionary being the extreme case). There is therefore an inherent trade-off in the suppression of the unwanted source and the reduction of artifacts.

Traditional factorial HMMs that use a Gaussian for the observation model have also been used for source separation [8, 9]. As in [3], these methods model each time frame of each source with a single spectral vector. The proposed model, on the other hand, extends non-negative factorizations by modeling each time frame of each source as a linear combination of spectral vectors. As shown above, this type of modeling can be advantageous for source separation.

5 Conclusions

We have presented new models and associated estimation algorithms that model the non-stationarity and temporal structure of audio. We presented a model for single sources and a model for sound mixtures. The performance of the proposed model was demonstrated on single channel source separation and was shown to have a much higher suppression capability than similar approaches that do not incorporate temporal information. The computational complexity is exponential in the number of sources (as with traditional factorial HMMs). Therefore, approximate inference algorithms [6] such as variational inference is an area for future work. Although the model was only demonstrated on source separation in this paper, it can be useful for various applications that deal with sound mixtures such as concurrent speech recognition and automatic music transcription.

References

1. Smaragdis, P., Brown, J.C.: Non-negative matrix factorization for polyphonic music transcription. In: WASPAA. (2003)
2. Rabiner, L.R.: A tutorial on hidden Markov models and selected applications in speech recognition. Proceedings of the IEEE **77**(2) (1989)
3. Ozerov, A., Fevotte, C., Charbit, M.: Factorial scaled hidden Markov model for polyphonic audio representation and source separation. In: WASPAA. (2009)
4. Smaragdis, P., Raj, B., Shashanka, M.: A probabilistic latent variable model for acoustic modeling. In: Advances in models for acoustic processing, NIPS. (2006)
5. Benaroya, L., Bimbot, F., Gribonval, R.: Audio source separation with a single sensor. IEEE TASLP **14**(1) (2006)
6. Ghahramani, Z., Jordan, M.: Factorial hidden Markov models. Machine Learning **29** (1997)
7. Vincent, E., Gribonval, R., Févotte, C.: Performance measurement in blind audio source separation. IEEE TASLP **14**(4) (2006)
8. Hershey, J.R., Kristjansson, T., Rennie, S., Olsen, P.A.: Single channel speech separation using factorial dynamics. In: NIPS. (2007)
9. Virtanen, T.: Speech recognition using factorial hidden Markov models for separation in the feature space. In: Proceedings of Interspeech. (2006)