A Non-negative Approach to Language Informed Speech Separation

Gautham J. Mysore[1] and Paris Smaragdis[1,2]

[1] Advanced Technology Labs, Adobe Systems Inc.,
[2] University of Illinois at Urbana-Champaign

Abstract. The use of high level information in source separation algorithms can greatly constrain the problem and lead to improved results by limiting the solution space to semantically plausible results. The automatic speech recognition community has shown that the use of high level information in the form of language models is crucial to obtaining high quality recognition results. In this paper, we apply language models in the context of speech separation. Specifically, we use language models to constrain the recently proposed non-negative factorial hidden Markov model. We compare the proposed method to non-negative spectrogram factorization using standard source separation metrics and show improved results in all metrics.

1 Introduction

The cocktail party problem is a classical source separation problem in which the goal is to separate speech of multiple concurrent speakers. This is a challenging problem, particularly in the single channel case. It would therefore be beneficial to use any high level information that is available to us. Specifically, if it is known that the speakers follow a certain grammar (constrained sequences of words), this information could be useful. We refer to this as a language model. This is routinely used in automatic speech recognition [1] and is in fact crucial to obtaining recognition results with high accuracy.

Non-negative spectrogram factorization algorithms [2] are a major research area in the source separation community and have been quite successful. They provide rich models of the spectral structure of sound sources by representing each time frame of the spectrogram of a given source as a linear combination of non-negative spectral components (analogous to basis vectors) from a dictionary. However, they model each time frame of audio as independent and consequently ignore an important aspect of audio – temporal dynamics. In order to address this issue, we proposed the non-negative hidden Markov model (N-HMM) [3] in which we model a given source using multiple dictionaries of spectral components such that each time frame of audio is explained by a linear combination of spectral components from one of the dictionaries. This gives us the rich spectral modeling capability of non-negative spectrogram factorizations. Additionally, we learn a Markov chain that explains the temporal dynamics between the dictionaries.

The dictionaries therefore correspond to states of the Markov chain. We model mixtures by combining N-HMMs of individual sources into the non-negative factorial hidden Markov model (N-FHMM).

There has been some other work [4,5,6] that extends non-negative spectrogram factorizations to model temporal dynamics. Ozerov [4] and Nakano [5] modeled the temporal dynamics between individual spectral components rather than dictionaries. They therefore model each time frame of a given source with a single spectral component rather than a linear combination of spectral components and can thus be too restrictive. Smaragdis [6] introduced a model that does allow linear combinations of spectral components with transitions between dictionaries. However, it also allows all spectral components of all dictionaries to be active at the same time, which is often not restrictive enough.

Since we use a hidden Markov model structure, we can readily use the ideas of language modeling from automatic speech recognition in the context of source separation. That is the context of this paper. Specifically, we constrain the Markov chain of each individual source to explain a valid grammar.

There has been some previous work [6,7,8] on modeling concurrent speakers using hidden Markov models and factorial hidden Markov models with language models. However, the goal has been concurrent speech recognition of multiple speakers. These papers report speech recognition performance and are presumably optimized for this. On the other hand, our goal is high quality source separation and we make design decisions for this goal. Also, to the best of our knowledge, no previous work on using language models for multiple concurrent speakers has reported source separation metrics.

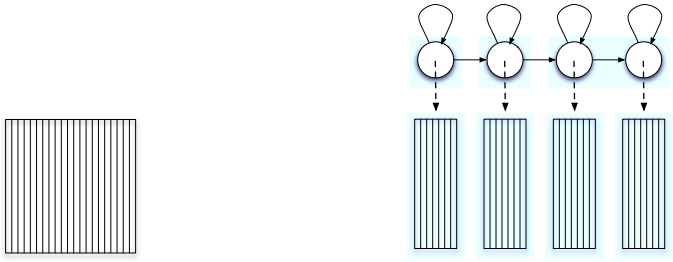
2 Models of Individual Speakers

In this section, we explain how we learn models of individual speakers. We first describe the Non-negative hidden Markov model (N-HMM). We then explain how to learn N-HMMs for individual words of a given speaker. Finally, we explain how to combine these individual word models into a single N-HMM according to the rules of the grammar, as dictated by the language model.

2.1 Non-negative Hidden Markov Model

Non-negative spectrogram factorizations (Fig. 1a) include non-negative matrix factorization (NMF) and their probabilistic counterparts such as probabilistic latent component analysis (PLCA). These models use a single dictionary of non-negative spectral components to model a given sound source. Specifically, they explain each time frame of the spectrogram of a given source with a linear combination of spectral components from the dictionary. These models however ignore two important aspects of audio – non-stationarity and temporal dynamics. To overcome this issue, we proposed the N-HMM (Fig.1b) [3]. This model uses multiple dictionaries such that each time frame is explained by any one of the several dictionaries (accounting for non-stationarity). Additionally it uses

a Markov chain to explain the transitions between dictionaries (accounting for temporal dynamics).



(a) Non-negative Spectrogram Factorization

(b) Non-negative Hidden Markov Model with left to right transitions

Fig. 1: Comparison of non-negative models. Non-negative spectrogram factorization uses a single large dictionary to explain a sound source, whereas the N-HMM uses multiple small dictionaries and a Markov chain.

The graphical model of the N-HMM is shown in Fig.2. Each dictionary corresponds to a state q. At time t, the N-HMM is in state q_t. Each spectral component of a given dictionary q is represented by z. A given spectral component is a discrete distribution. Therefore, spectral component z of dictionary q is represented by $P(f|z,q)$. The *non-negativity* in the N-HMM comes from the fact that the parameters of a discrete distribution are non-negative by definition. Since each column of the spectrogram is modeled as a linear combination of spectral components, time frame t (modeled by state q) is given by the following observation model:

$$P(f_t|q_t) = \sum_{z_t} P(f_t|z_t, q_t)P(z_t|q_t), \tag{1}$$

where $P(z_t|q_t)$ is a discrete distribution of mixture weights for time t. The transitions between states are modeled with a Markov chain, given by $P(q_{t+1}|q_t)$.

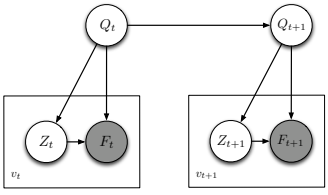


Fig. 2: Graphical model of the N-HMM

2.2 Word Models

Given an instance of a word, we can estimate the parameters of all of the distributions of the N-HMM using the expectation–maximization (EM) algorithm [3].

In this paper, we extend this idea to learn word models from multiple instances of a given word as routinely done in speech recognition [1]. We compute the E step of EM algorithm separately for each instance. The procedure is the same as in [3]. This gives us the marginalized posterior distributions $P_t^{(k)}(z, q|f, \bar{\mathbf{f}})$ and $P_t^{(k)}(q_t, q_{t+1}|\bar{\mathbf{f}})$ for each instance k. We use these in the M step of the EM algorithm. Specifically, we compute a separate weights distribution for each instance k as follows:

$$P_t^{(k)}(z_t|q_t) = \frac{\sum_{f_t} V_{ft}^{(k)} P_t^{(k)}(z_t, q_t|f_t, \bar{\mathbf{f}})}{\sum_{z_t}\sum_{f_t} V_{ft}^{(k)} P_t^{(k)}(z_t, q_t|f_t, \bar{\mathbf{f}})}, \tag{2}$$

where $V_{ft}^{(k)}$ is the spectrogram of instance k. However, we estimate a single set of dictionaries of spectral components and a single transition matrix using the marginalized posterior distributions of all instances as follows:

$$P(f|z, q) = \frac{\sum_k \sum_t V_{ft}^{(k)} P_t^{(k)}(z, q|f, \bar{\mathbf{f}})}{\sum_f \sum_k \sum_t V_{ft}^{(k)} P_t^{(k)}(z, q|f, \bar{\mathbf{f}})}, \tag{3}$$

$$P(q_{t+1}|q_t) = \frac{\sum_k \sum_{t=1}^{T-1} P_t^{(k)}(q_t, q_{t+1}|\bar{\mathbf{f}})}{\sum_{q_{t+1}} \sum_k \sum_{t=1}^{T-1} P_t^{(k)}(q_t, q_{t+1}|\bar{\mathbf{f}})}. \tag{4}$$

We restrict the transition matrix to use only left to right transitions.

2.3 Combining Word Models

Once we learn N-HMMs for each word of a given speaker, we combine them into a single speaker dependent N-HMM. We do this by constructing a large transition matrix that consists of each individual transition matrix. The transition matrix of each individual word stays the same. However, the transitions between words are dictated by a language model. Each state of the speaker dependent N-HMM corresponds to a specific dictionary of that speaker. Therefore, this N-HMM also contains all dictionaries of all words.

3 Model of Mixtures

We first describe how to combine models of individual speakers into a model of speech mixtures. We then explain how to use this model for speech separation. Finally, we describe the pruning that we use to reduce computational complexity.

3.1 Combining Speaker Dependent Models

We model a mixture of two speakers using the non-negative factorial hidden Markov model (N-FHMM) [3]. Given the N-HMM of two speakers, we can combine them into an N-FHMM. We use the dictionaries and the Markov chains of

the N-HMMs of the two speakers. A given time frame is then explained using any one dictionary of the first speaker and any one dictionary of the second speaker. Specifically, the given time frame is modeled using a linear combination of the spectral components of the two appropriate dictionaries. This is illustrated in Fig. 3.

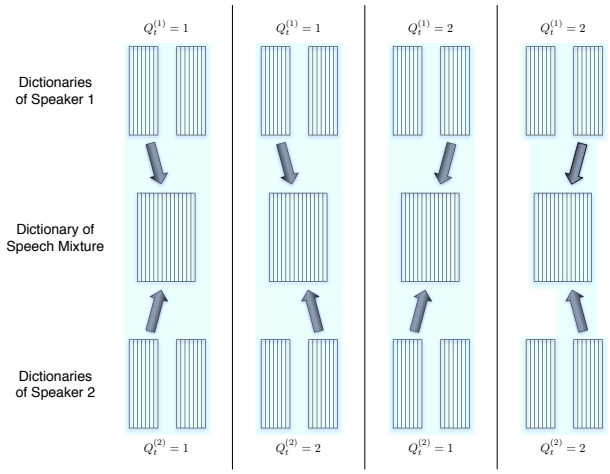


Fig. 3: Combining dictionaries of two sources to model a mixture. In this simple example, each source has two dictionaries so there are a total of four ways of combining them.

The graphical model of the N-FHMM is shown in Fig. 3. An N-HMM can be seen in the upper half of the graphical model and another one can be seen in the lower half. The interaction model (of the two sources) introduces a new variable s_t that indicates the ratio of the sources at a given time frame. $P(s_t|q_t^{(1)}, q^{(2)})$ is a Bernoulli distribution that depends on the states of the sources at the given time frame. The interaction model is given by:

$$P(f_t|q_t^{(1)}, q_t^{(2)}) = \sum_{s_t} \sum_{z_t} P(f_t|z_t, s_t, q_t^{(s_t)}) P(z_t, s_t|q_t^{(1)}, q_t^{(2)}), \tag{5}$$

where $P(f_t|z_t, s_t, q_t^{(s_t)})$ is spectral component z_t of state $q_t^{(s_t)}$ of source s_t.

3.2 Speech Separation

$P(z_t, s_t|q_t^{(1)}, q_t^{(2)})$ combines the new distribution $P(s_t|q_t^{(1)}, q^{(2)})$ and the weights distributions of each source into a single weights distribution of the mixture. Since the dictionaries and the Markov chain of each source are already specified, if we learn the weights distribution of the mixture, we can estimate soft masks to separate the two sources. This is done using the EM algorithm. Details on how to estimate the masks and then separate the sources can be found in [3].

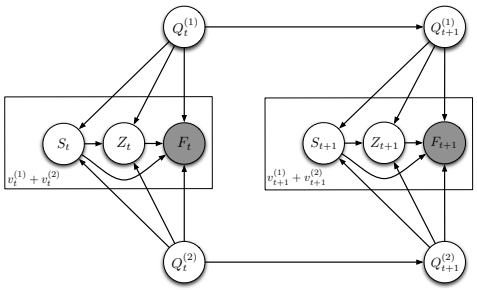


Fig. 4: Graphical model of the N-FHMM

3.3 Pruning

At every time frame, we need to compute the likelihood of every possible state pair (one state from each source). This causes the computational complexity of the N-FHMM to be exponential in the number sources. This can lead to intractable computation. However, we do not need to consider state pairs that have a very small probability. Specifically, we prune out all of the state pairs whose posterior probability $\gamma(q_t^{(1)}, q_t^{(2)})$, is below a pre-determined threshold. In our experiments, we set this threshold to -1000 in the log domain. Even though it is an extremely small number, this pruned out around 99% of the state pairs. This is due to the heavy constraining of the language model.

For each speaker, we used an N-HMM of 127 states (more details are in Sec. 4). Therefore, there are a total of 16129 possible state pairs. With our pruning, we need to consider less than 250 state pairs in most time frames. With the number of states that we used, this corresponds to computation complexity that is linear in the number of sources.

4 Experimental Results and Discussion

We performed experiments on a subset of the data from the speech separation challenge [9]. The test data from the challenge does not contain ground truth, without which we cannot compute source separation metrics. Therefore, we divided the training data into a training set and a test set. We trained N-HMMs for 10 speakers using 450 of the 500 sentences from the training set of each speaker. The remaining 50 sentences were used to construct the test set. We segmented the training sentences into words in order to learn individual word models as described in Section 2.2. We used one dictionary (state) per phoneme. This is less than what is typically used in speech recognition. However, we did not want to excessively constrain the model in order to obtain high quality reconstructions. We used 10 spectral components per dictionary as this number was previously found to give good results in N-HMMs [3].

We then combined the word models of a given speaker into a single N-HMM according to the language model, as described in Section 2.3.

We performed speech separation using the N-FHMM on speakers of different genders and the same gender[3]. For both categories, we constructed 10 test mixtures from our test set. The mixing was done at 0dB. We evaluated the source separation performance using the BSS-EVAL metrics [10]. As a comparison, we performed separation using a non-negative spectrogram factorization technique (PLCA) [2]. When using PLCA, we used the same training and test sets that we used with the proposed model. However, we simply concatenated all of the training data of a given speaker and learned a single dictionary for that speaker, which is customary when using non-negative spectrogram factorizations [2]. We used a dictionary size of 100 spectral components as this gave the best separation results. This is more than used in our previous paper [3] since the database used in this paper has much more training data for each speaker. The proposed method has the advantage (over PLCA) of using language information. However, the point that we are trying to make is that language information can lead to improved speech separation results.

Our results are shown in Table 1. The proposed model outperforms PLCA in all metrics of both categories. Specifically, we see a 7-8dB improvement in source to interference ratio (SIR) while still maintaining a higher source to artifacts ratio (SAR). This means that we are achieving much higher amounts of separation than PLCA and also introducing less artifacts. The source to distortion ratio (SDR), which reflects both of these things is therefore also higher.

Another observation is that when we compare the performance of the N-FHMM in the two categories, we see only a small deterioration in performance from the different gender to the same gender case (0.5-1 dB in each metric). With PLCA, however, we see a greater deterioration in SIR and SDR (2-3 dB). This is because the dictionaries of the two sources are much more similar in the same gender case than in the different gender case. With the N-FHMM, the language model helps disambiguate the sources. However, only the spectral information is used in the case of PLCA.

Table 1: Source separation performance of the N-FHMM and PLCA.

Different Gender	SIR	SAR	SDR
N-FHMM	14.91	10.29	8.78
PLCA	7.96	9.08	4.86

Same Gender	SIR	SAR	SDR
N-FHMM	13.88	9.89	8.24
PLCA	5.11	8.77	2.85

The introduction of constraints, priors, and additional structure in non-negative models often leads to improved separation quality (higher SIR), when compared to PLCA or NMF. However, this usually leads to more artifacts (lower SAR). Ozerov [4] noted this with the FS-HMM. We have improved results in both metrics. The reason is that the language model only attempts to determine the correct dictionary to explain each source but not the exact fitting of the spectral components of the given dictionary to the data. Once this dictionary of each source is determined for a given time frame, the algorithm fits the corresponding spectral components to the mixture data to obtain the closest possible

[3] Examples at https://ccrma.stanford.edu/~gautham/Site/lva_ica_2012.html

reconstruction of the mixture. This flexibility after determining the appropriate dictionary avoids excessive artifacts.

5 Conclusions

We presented a method to perform high quality speech separation using language models in the N-HMM framework. We showed that use of the language model greatly boosts source separation performance when compared to non-negative spectrogram factorization. The methodology was shown for speech but it can be used in other contexts in which high level structure information is available such as incorporating music theory into the N-HMM framework for music separation.

References

1. Lawrence R. Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, 1989.
2. Paris Smaragdis, Bhiksha Raj, and Madhusudana Shashanka. Supervised and semi-supervised separation of sounds from single-channel mixtures. In *Proceedings of the International Conference on Independent Component Analysis and Signal Separation*, Sept. 2007.
3. Gautham J. Mysore, Paris Smaragdis, and Bhiksha Raj. Non-negative hidden Markov modeling of audio with application to source separation. In *Proceedings of the International Conference on Latent Variable Analysis and Signal Separation*, Sept. 2010.
4. Alexey Ozerov, Cedric Fevotte, and Maurice Charbit. Factorial scaled hidden Markov model for polyphonic audio representation and source separation. In *Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, Oct. 2009.
5. Masahiro Nakano, Jonathan Le Roux, Hirokazu Kameoka, Yu Kitano, Nobutaka Ono, and Shigeki Sagayama. Nonnegative matrix factorization with Markov-chained bases for modeling time-varying patterns in music spectrograms. In *Proceedings of the International Conference on Latent Variable Analysis and Signal Separation*, Sept. 2010.
6. Paris Smaragdis and Bhiksha Raj. The Markov selection model for concurrent speech recognition. In *Proceedings of the IEEE International Workshop on Machine Learning for Signal Processing*, August 2010.
7. John R. Hershey, Steven J. Rennie, Peder A. Olsen, and Trausti T. Kristjansson. Super-human multi-talker speech recognition: A graphical modeling approach. *Computer Speech and Language*, 24(1):45–46, 2010.
8. Tuomas Virtanen. Speech recognition using factorial hidden Markov models for separation in the feature space. In *Proceedings of Interspeech*, Pittsburgh, PA, Sept. 2006.
9. Martin Cooke, John R. Hershey, and Steven J. Rennie. Monaural speech separation and recognition challenge. *Computer Speech and Language*, 24(1):1–15, 2010.
10. Emmanuel Vincent, Rémi Gribonval, and Cédric Févotte. Performance measurement in blind audio source separation. *IEEE Transactions on Audio, Speech, and Language Processing*, 14(4):1462–1469, July 2006.