Signal separation by efficient combinatorial optimization

Manuel Reyes-Gomez
Microsoft Research
Redmond,WA.
manuelrg@microsoft.com

Nebojsa Jojic
Microsoft Research
Redmond,WA.
jojic@microsoft.com

Abstract

We present a formulation of the source separation problem as the minimization of a symmetric function defined on fragments of the observed signal. We prove the function to be posimodular and propose the use of tractable combinatorial optimization techniques, in particular Queyranne's algorithm, suited to optimization of symmetric submodular and posimodular functions. While these ideas can be applied to any signal segmentation problem (e.g., image or video segmentation), we focus here on unsupervised separation of sources in mixed speech signals recorded by a single microphone. The optimization criterion is the likelihood under a generative model which assumes that each time-frequency bin is assigned to one of the two speakers, and that each speaker's utterance has been generated from the same generic speech model. This assumption is made given that the time-frequency representation of speech signals is very sparse. The optimization can then be performed over all possible assignments of the time-frequency bins to the two speakers. Even though the algorithm requires polynomial time, it is still too slow for large signals. Therefore, we first oversegment the spectrogram into a large number of segments which do not violate the deformable spectrogram model [1]. Queyranne's algorithm is then constrained to search only over unions of these segments, rather than all possible signal fragments. We show that this technique leads to blind separation of mixed signals where both speakers are of the same gender and very similar spectral characteristics.

1 Introduction

This paper is concerned with analysis of multidimensional signals $X = \{x_{\mathbf{i}} : \mathbf{i} \in V\}$, where V is the domain of the signal. For example, a 255×255 image has 2-D indices $\mathbf{i} = (i, j) \in [1..255] \times [1..255]$. An audio spectrogram also has 2-D time-frequency indices $\mathbf{i} = (t, f) \in [1..T] \times [1..F]$, where T is the number of time samples, and F is the number of frequency bins in the representation. We consider a class of signals drawn from the (trainable) joint probability distribution $p(Y|\theta)$, and study an observed mixture X of two signals (sources) of this class. The mixing is approximated assuming that each mixed signal component $x_{\mathbf{i}}$ comes from one of the two individual sources. This assumption is made given that the time-frequency representation of speech signals is very sparse: since most narrow frequency bands carry substantial energy only during a small fraction of time and therefore is rare to encounter two independent sources with large amounts of energy at the same frequency band at the same time.

If we choose a set $S \subset V$ as the set of observed elements to be assigned to the first source, then the log likelihood of the observed signal given the assignment S is:

$$\log p(X|S) = \log p(X_S|\theta) + \log p(X_{V \setminus S}|\theta), \tag{1}$$

where $X_A = \{x_{\mathbf{i}} : \mathbf{i} \in A\}$, and so X_S and $X_{V \setminus S}$ constitute a partition of the signal into two fragments. Note that $p(X_A|\theta) = \sum_{X_{V \setminus A}} p(X|\theta)$, and that the above log likelihood is a symmetric set function ($\log p(X|S) = \log p(X|V \setminus S)$), as the two sources are assumed to follow the same probability distribution.

We will consider signal segmentation as a search for the partition that maximizes this likelihood. For this purpose, we propose the use of Queyranne's algorithm [2], which has the complexity $O(|V|^3)$. The complexity can be reduced if the signal comes presegmented into a large number of smaller regions R_i, $i \in 1..N$ and the search is limited to the unions of these regions. In that case the algorithm has the complexity $O(|N|^3)$.

In our experiments, we focused on separating sources in mixed speech signals, for which we propose the use of the deformable spectrograms model to provide pre-segmentation as described in Section 5, and the use of a generic speech model trained using the HTK library to define the speech model $p(X|\theta)$. In this way, the source separation is driven both by the semantics of the inferred speech as well as the lower level features. On 100 same gender mixtures we obtained overall word recognition rate of 82.17%. This error was measured on the output transcriptions obtained from the generic speech recognizer applied to each signal partition. These transcriptions, are in fact the inferred hidden variables in the models $p(X_S|\theta)$ and $p(X_{S \setminus V}|\theta)$ for optimal S. We expect that this general strategy can be applied to other types of natural signals.

The paper is organized as follows. The next section provides background the Queyranne's algorithm, and illustrates the relationship between our optimization criteria and the submodular functions optimized in [3]. In Section 5 we briefly describe deformable spectrograms and propose the use of this representation for discovering small regions dominated by a single speaker. These regions are clustered using an algorithm which is based on the Queyranne's algorithm, and described in detail in Section 2. Finally, in Section 6 we present experimental results.

2 Queyranne's algorithm and signal segmentation

In [3] it was shown that several types of clustering criteria can be reduced to functions that can be optimized using Queryanne algorithm [2], whose complexity is $O(|V|^3)$. In particular, [3] shows that separating sites in genetic sequences into two clusters so that the mutual information between clusters is minimized can be performed exactly using this algorithm. Their optimization criterion can also be shown as equivalent to the minimal description length criterion:

$$f(S) = H(X_S) + H(X_{V \setminus S}), \tag{2}$$

where

$$H(X_A) = -\sum_{X_A} p(X_A) \log p(X_A), \tag{3}$$

is the entropy of the observations at indices in A. The task of separating sequence sites is defined as finding the partition (S,$V \setminus S$), for which the sum of the two entropies is minimized, and to estimate the entropy multiple genetic sequences are observed under the assumption that a single partition should work for all sequences. The optimization criterion is a symmetric and submodular function, and so Queryanne algorithm can be used to find optimal S in $O(|V|^3)$ time. The resulting segmentation guarantees, that X_S and $X_{V \setminus S}$, over the observed sequences, are as independent of each other as possible. The entropy $H(X_A)$ is clearly related to log likelihood. To estimate an entropy of a signal piece S for a class of signals X^k sampled from a distribution $p(X|\theta)$, we can use:

$$H(X_A) \simeq -\sum_k \log p(X_A^k|\theta), \tag{4}$$

where samples X_A^k are used as an empirical distribution instead of the true distribution. If the empirical distribution truly matches the model distribution, the entropy estimate will be correct. Thus, the MDL criterion $f(S)$ can be thought of as a negative of the log likelihood criterion $-\log p(X|S)$, where only a single mixed signal is observed, rather than an ensemble of consistently mixed signals, as was the case in the genetics application in [3].

As opposed to $f(S)$ in (2), the new criterion $-\log p(X|S)$ is symmetric, but not a submodular function. However it is a posimodular function.

For a function $f(S)$ to be posimodular the following should hold:

$$f(A) + f(B) \geq f(A - B) + f(B - A). \tag{5}$$

Plugging $f(A) = -log(P(X_A))$ on the above yields a posimodular inequality.

Proof

For $A = C + D$, $B = E + D$ and $D = A \cap B$. Then $f(A) = -log(p(X_C, X_D))$ and $f(B) = -log(p(X_D, X_E))$.

$$-log(p(X_C, X_D)) + -log(p(X_D, X_E)) \geq -log(p(X_C)) + -log(p(X_E)). \tag{6}$$

$$p(X_D \mid X_C)p(X_D \mid X_E) \leq 1. \tag{7}$$

An as it is shown in [4], the Queyranne's algorithm is exact for posimodular functions.

We denote R_i $i \in [1..N]$, as N non overlapping regions of V. i.e. $V = \sum_{i=1}^{N}(R_i)$, $R_i^{'} = V \backslash R_i$ as all the regions in V but R_i, S as a union of individual regions $S = \sum_{i \in G}(R_i)$, and $S' = V \backslash S$, as all the regions in V but the ones in S. $\mathcal{L}(S) = logp(X_S|\theta)$ as the loglikelihood of signal part X_S under a certain model (marginalizing over the rest of the signal as hidden) and $\mathcal{L}_T(S) = \mathcal{L}(S) + \mathcal{L}(V \setminus S)$ as the total loglikehihood for partition $P = (S, V \setminus S)$ under the same model.

The practical implementation of the algorithm works as follows.

Initializing R_i to the smallest possible elements in V, i.e. $R_i = x_i$
$N_{new} = N$;
While $N_{new} \geq 2$.
$S = \emptyset$. (Starting with no partition at all)
N_{tested} = 0;
While $N_{tested} \leq N_{new} - 2$.
For all $R_i \in V \setminus S$
Compute $\mathcal{L}^T(S + R_i)$ (Testing gain in partition loglikelihood if R_i is switched).
end
$R_i \leftarrow argmax_{R_j \in V \backslash S}\ (\mathcal{L}^T(S + R_j))$;
$S \leftarrow S + R_i$; (Switching the region for which the gain is maximal).
$N_{tested} = N_{tested} - 1$
end
By this point there is only two original regions untested R_l and R_k.
$R_i \leftarrow max(\mathcal{L}^T(R_l), \mathcal{L}^T(R_k))$
Place $(R_i, V \setminus R_i)$ in the list of possible solutions
Merge regions R_l and R_k into a single R_m.
$SetN_{new} \leftarrow N_{new} - 1$ and reindex original regions.
end
Choose the best solution from the list of possible solutions.

The algorithm has a complexity of $O(N^3)$.

3 Using a generic speech model with composed signals

The Queyranne's algorithm is good as it works for any $p(X_S|\theta)$ regardless of its complexity. Given that speech has a clear structure, it is often modeled using hidden Markov models HMMs, which are plausible to alternative segmentation solutions such as the one described by factorial HMMs. A generic speech model can be build from a database of single speaker utterances by training individual HMMs for each basic unit in the vocabulary and later concatenating the individual HMMs according to the restrictions imposed by an specific language model.

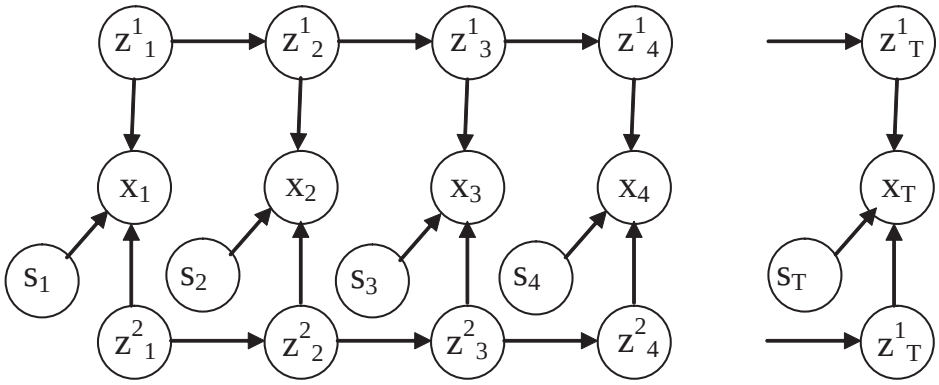

Figure 1: Factorial HMM)

To model a composed signal each one of the sources in the mixture is modeled using the generic speech model while defining the composed output as a combination of the output from the individual HMMs as defined by a mask variable S. This composed model is known as a factorial HMM an is depicted in figure 1.

The joint likelihood of the model for variables $X = [x_1, x_2, .., x_T]$, $S = [s_1, s_2, .., s_T]$,Z^1 = $[z_1^1, z_2^1, .., z_T^1]$ and Z^1 = $[z_1^1, z_2^1, .., z_T^1]$ is defined as:

$$P(X, Z^1, Z^2, S) = \prod_{t=1}^{T} P(x_t \mid z_t^1, z_t^2, s_t) P(z_t^1 \mid z_t^1) P(z_{t-1}^2 \mid z_{t-1}^2) \tag{8}$$

For the oclussion model assummed in this paper, binary mask s_t defines partitions S and $V \setminus S$. Non $zero$ values on s_t define S while $zero$ values define $V \setminus S$.

Inference of the model requires the optimization of function $Q(Z^1, Z^2, S)$ [5], in equation

$$\mathcal{L}(Q, \theta) = max_Q(\sum Q(Z^1, Z^2, S) * log(\frac{P(X, Z^1, Z^2, S)}{Q(Z^1, Z^2, S)})) \tag{9}$$

Is well know that the function $Q(Z^1, Z^2, S)$ that maximizes the above equation corresponds to the posterior probability $P(Z^1, Z^2, S \mid X)$, we considered two exact factorizations.

1. Q_F = $Q(S \mid Z^1, Z^2)Q(Z^1, Z^2)$ and
2. Q_Q = $Q(Z^1, Z^2 \mid S)Q(S)$

The use of the first kind would result in an inference procedure similar to the one needed for a regular HMM with L^2 states, where L is the number of states in the original generic speech model. As described in [6]

For the later case, given the oclussion model: $Q(Z^1, Z^2 \mid S)$ = $Q(Z^1 \mid S)Q(Z^2 \mid S)$ and that probabilites $P(X \mid Z^1, Z^2, S)$ are defined by mixtures of gaussians with diagonal covariances. Equation 9 can be effectively decoupled optimizing each of the chains independently with individual observation dictaded by the partition given by S ($S = 1$) and $V \setminus S$ ($S = 0$), just as defined by eq. 1, linking Q_Q to the Queyranne's algorithm.

4 Algorithm complexity for the source separation task

The algorithm separation complexity for the two considered types of $Q(Z^1, Z^2, S)$ factorization are:

1. $O(Q_Q)$ = $(FT)^3(2L^2)(T)$
2. $O(Q_F)$ = $2L^3(T)$

Where F is the number of frequency bins in the representation, T the number of frames and L the number of states of the generic speech model.

Given that the number of total elements in the mask is FT, the Queryanne algorithm will require $(FT)^3$ iterations to find the optimal solution. Computing the log-likelihood of each partition requires computing viterbi aligments over the two chains, each with L^2T complexity.

The factorial optimization requires one viterbi alignment over a HMM with L^2 states with a $2L^3(T)$ complexity [6].

Clearly both approaches are intractable given that factors FT^3 and L^3 become unmanageable for any practical values for F, T and L. Therefore both approaches require some sort of approximation.

For Queyranne's algorithm case given that time-frequency cells belonging to any particular source occur in large clumps (local regions), and as it has been empirically demostrated highly-intelligible separation can be achieved by limiting the masks to consist of relatively large, locally-consistent regions of labeling [1], the time-frequency bins are first locally grouped in consistent regions before applying the Queryanne algorithm. We use the deformable spectrograms model (section 5) [7] to find a set of N locally consistent regions for each composed signal, where $N << FT$. Furthermore as it will be discussed in the experimental section, in practice the actual number of evaluations for each chain is much smaller than the theoretical N^3 number, being more in the vicinity of $0.05 * N^3$.

For the regular factorial case the magnitude of L^3 depends entirely in the size of the vocabulary used in the training of the generic speech model, for most practical applications the value of L can easily be found in the thousands. In this case the approximation is usually done by beam search [8], where the large state space is first locally limited to a small subspace of states that achive high local likelihood.

It is clear that for small vocabulary applications doing the full factorial search is a more viable option than recurring to the Queryanne algorithm, however for short utterances for applications with large vocabularies, which would be the case for most practical application, the use of the Queryanne algorithm is a much better option.

Since both approaches rely on approximations they are both prone to errors. However, as discussed in [7] the deformable spectrogram model achieves a high recall value when identifying regions dominated by a single source with error due to noise resulting in false positives (over segmentation) rather than omitions of true positives. The beam search approach in other hand is very susceptible to local noise, given that local noise can divert the search to the wrong local search subspace, an error that can be easily further propagated in the subsequent frames. The potential problem are specially critical if the number of states is quite large since a workable subspace will represent just a very limited set of the possible local matches reducing in great manner the probability of obtained the correct alignment.

5 Deformable spectrograms

Many audio signals have spectral representations that show high correlation between adjacent frames. The deformable spectrogram model discovers and tracks the nature of such correlation by finding how the patterns of energy are transformed between adjacent frames and how those transformations evolve over time. The model was introduced and presented in detail in [1]. Figure 2 shows a narrow band spectrogram representation of a speech signal, where each column depicts the energy content across frequency in a short-time window, or time-frame. Using the subscript C to designate current and P to indicate previous, the model predicts a patch of N_C time-frequency bins centered at the k^{th} frequency bin of frame t as a "transformation" of a patch of N_P bins around the k^{th} bin of frame $t-1$, i.e.

$$\vec{X}_t^{[k-n_C,k+n_C]} \approx \vec{T}_t^k \cdot \vec{X}_{t-1}^{[k-n_P,k+n_P]} \tag{10}$$

where $n_C = (N_C - 1)/2$, $n_P = (N_P - 1)/2$, and T_t^k is the particular $N_C \times N_P$ transformation matrix employed at that point on the time-frequency plane. Figure 2 shows an example with $N_C = 3$ and $N_P = 5$ to illustrate the intuition behind this approach. The selected patch in frame t can be seen as a close replica of an upward shift of part of the patch highlighted in frame $t-1$. This "upward" relationship can be captured by a transformation matrix, such as the one shown in the

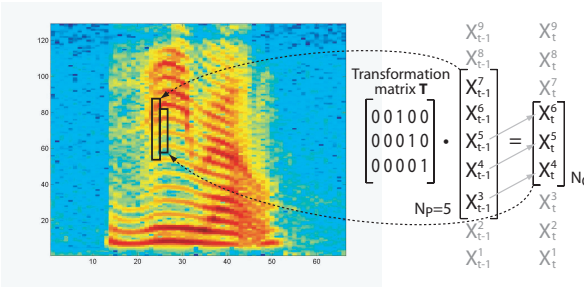

Figure 2: The $N_C = 3$ patch of time-frequency bins outlined in the spectrogram can be seen as an "upward" version of the marked $N_P = 5$ patch in the previous frame. This relationship can be described using the matrix shown.

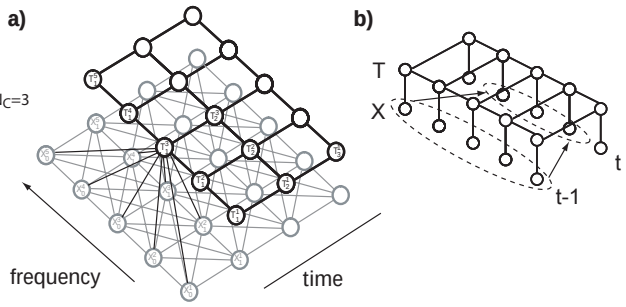

Figure 3: a) Graphical model b) Graphical simplification.

figure. The patch in frame $t-1$ is larger than the patch in frame t to permit both upward and downward motions. The proposed model finds the particular transformation, from a discrete set of transformations, that better describes the evolution of the energy from frame $t-1$ to frame t around each one of the time frequency bins x_t^k in the spectrogram. The model also tracks the nature of the transformations throughout the whole signal to find useful patterns of transformation. The generative graphical model is depicted in figure 3. Nodes $\mathcal{X} = \{x_1^1, x_1^2, ..., x_t^k, ..., x_T^K\}$ represent all the time-frequency bins in the spectrogram. Considering the continuous nodes $\mathcal{X}$ as observed or hidden when parts of the spectogram are missing, discrete nodes $\mathcal{T} = \{T_1^1, T_1^2, ..., T_t^k, ..., T_T^K\}$ index the set of transformation matrices used to model the dynamics of the signal. Many sound sources, can be regarded as the convolution of a broad-band *source excitation*, and a time-varying resonant *filter*, therefore the overall spectrum is in essence the convolution of the source with the filter in the time domain, which corresponds to multiplying their spectra in the Fourier domain, or adding in the log-spectral domain. Hence, we model the log-spectra $\mathcal{X}$ as the sum of variables $\mathcal{F}$ and $\mathcal{H}$, which explicitly model the formants and the harmonics of the speech signal. The source-filter transformation model is based on two additive layers of the deformation model described above.

Prediction of frames from their context is not always possible such as when there are transitions between silence and speech or transitions between voiced and unvoiced speech, or when smooth regions on the energy patterns of a single source are disrupted due to interference from a new source. Given that the magnitude of the interference is not uniform across all the spectrum, the model is extended to detect "vertical" (synchronized) sections of the spectogram, composed by a band of n adjacent time frequency bins on a given time frame, where the model cannot efficiently "track" the energy dynamics from the context, labeling the frame section as a transition boundary. Second row of figure 4, shows the transition boundaries obtained by the model for a female-female mixture of two speakers.

6 Experimental Results

A generic speech recognizer was trained using HTK with over 3000 clean speech signals from over 50 different female speakers from the Aurora database, which is composed of utterances of sequences of three to six continuous digits. We built individual HMMs for each of the eleven words in the vocabulary corresponding to digits: 'one', 'two', 'three', 'four', 'five', 'six', 'seven', 'eight', 'nine', 'oh' and 'zeros', as well as a 'silence' and a 'short pause' models. Each digit HMM had 16 states, the 'silence' model had three states and the 'short pause' one state. Each state in turn was comprised of 7 mixtures of Gaussians with diagonal covariances.

We tested our approach on 100 artificially mixed signals from two female speakers each one uttering a sequence of three continuous digits. The speakers were not present in the training set used to train the recognizer.

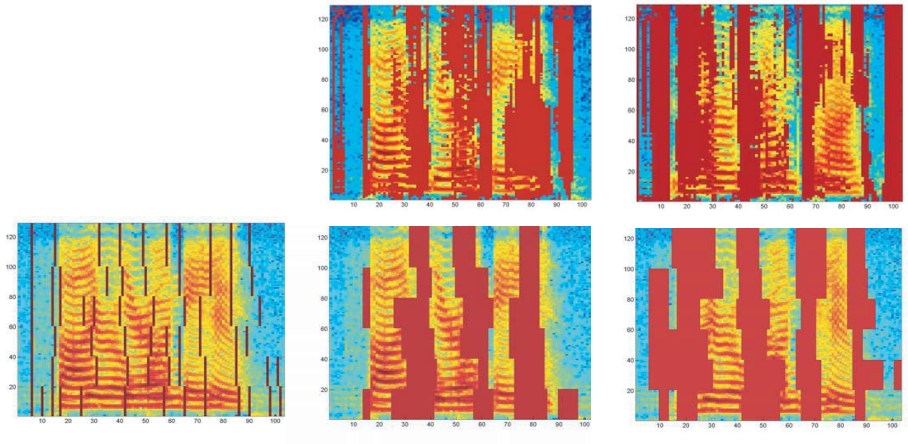

Figure 4: Detected boundaries: resulting in over 60 presegmented regions and estimated partition. (Presegmented silence regions are including in both signals)

Each mixture was first pre-segmented using the deformable spectogram model into regions with smooth energy patterns. Then, the algorithm in Section 2 was applied to each oversegmented signal to obtain the best partition of the two sources.

Before continuing to the evaluation of the partitions, we briefly discuss the computation expense of the algorithm. From 2 it can be seen that the algorithm requires up to N^3 evaluations under the speech decoder. This is quite a reduction from 2^N evaluations needed for the exhaustive search, and this makes this algorithm possible to evaluate. In fact, taking a closer look to the algorithm it becomes apparent that many of those evaluation are repeated and so recording the indexes of the original regions already tested in a hash table greatly reduces the actual number of evaluations needed. First part of table 1 shows the mean and the standard deviation of the ratio between the actual number of evaluations used to complete the algorithm for each mixture and the expected N^3 number of evaluations. The total number of calls to speech recognizer was only around 5% of the worst case N^3 calls.

Computational Cost		
Num. Evaluations Ratio	Mean	Std
Actual Number/N^3	0.054	0.011
Performance Evaluation		
Partition	LogLikelihood	Word Recog. Rate
P_{est}	-7.1220e+003	79.83%
P_{opt}	-7.3487e+003	83.50%

Table 1: Computational cost and performace evaluation

Given that the signals were artificially mixed we could obtain the "optimal" grouping of the dominant speaker regions by asigning each region to the speaker for which the amount of energy contained in its individual source is greater. We called this partition P_{opt}. Table shows performance comparisons for both set of partitions P_{est} and P_{opt}. The first column shows the mean for the partition loglikelihood for all mixture. In each single one of the mixtures the loglikelihood of partition P_{est} is greater than the loglikelihood obtained from partition P_{opt}, which indicates both that the optimization algorithm is working well, and that the generic model is under-trained. Second row shows the word recognition rate over the 600 hundred decoded digits, 3 per independent source over the 100 mixtures.

The test set included a few mixtures containing the *same* speaker uttering two different digits sequences. The word error rate on those mixtures is consistent with the one obtained for the complete test set. Second row of figure 4 shows an example of such a mixture with its correspondent

partitionP_{est}. Given the simplicity of the vocabulary for this task, complete inference of the factorial HMM is relatively feasible in the first row of figure 4 the partition obtained through the complete factorial inference for this mixture is shown, as can be seen the partitions are quite similar. In the suplemental material, we provide wav files with examples of speech separation using our algorithm.

7 Conclusions

An efficient algorithm for optimal signal segmentation through the maximization of the likelihood of the partition given a model of the individual sources was presented. Even though the presented algorithm reduces the combinatorial search space for a mixture of two speakers from 2^{FT} for the brute force approach to a more manageable FT^3, the complexity required for the specific task of separating speech mixtures is still quite large to be tractable, requiring then a reduction on the number of possible partitions by presegmenting the spectogram in local smooth regions through the use of the deformable spectrograms model.

The use of a generic speech model makes possible the segmentation of mixtures with very similar spectral characteristics thanks to the semantic constraints embedded in the model. But it it precisely the structure of speech models what permits alternatives to the presented algorithm through the formation of composed speech models. But even these alternatives require approximations for most practical application requiring a descent size vocabulary. These approximation are potentially more prone to errors than the approximations required by our approach.

In future research we will apply the algorithm presented in this paper to tasks involving larger vocabularies to properly test the performance of our approach compared to the one obtained using beam search techniques.

References

[1] M. Reyes-Gomez, N. Jojic and D. Ellis "Deformable Spectrograms", AISTATS, 2005.

[2] M. Queyranne. "Minimizing symmetric submodular functions", Math. Programming, 1998.

[3] M. Narasimhan and N. Jojic, "Q Clustering", NIPS, 2005.

[4] H. Narayanan "A note on the minimization of symmetric and general submodular functions" Discrete Applied Mathematics, September 2003.

[5] R. Neal and G. Hinton "A View of the EM Algorithm that Justifies Incremental, Sparse, and other Variants", Learning in Graphical Models, Kluwer.

[6] Zoubin Ghahramani and Michael I. Jordan "Factorial Hidden Markov Models", Advances in Neural Information Processing Systems, 1997.

[7] M. Reyes-Gomez, N. Jojic and D. Ellis "Modelling Sound Dynamics Using Deformable Spectrograms: Segmenting the Spectrogram into Smooth Regions", Asilomar, 2006.

[8] S. Roweis, "Factorial Models and refiltering for Speech Separation and Denoising", Proc. EuroSpeech, Geneva, 2003.