Signal separation by efficient combinatorial optimization

Anonymous Author(s)
Affiliation
Address
City, State/Province, Postal Code, Country
email

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 draw a parallel between this function and the mutual information function, which is known to be submodular, and propose the use of tractable combinatorial optimization techniques, in particular Queyranne's algorithm, suited to optimization of symmetric submodular 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. 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 [4]. Queyranne's algorithm is than 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 eaxmple, 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 assumed to be such that each mixed signal component $x_{\mathbf{i}}$ comes from one of the two individual sources. 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 [6], 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 4, 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. To our knowledge processing of these kind of single microphone mixtures is not possible with previous related research. 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 [5]. Section 3 provides a brief background on the single microphone blind source separation research. In Section 4 we 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 5. Finally, in Section 6 we present experimental results.

2 Queyranne's algorithm and signal segmentation

In [5] it was shown that several types of clustering criteria can be reduced to functions that can be optimized using Queryanne algorithm [6], whose complexity is $O(|V|^3)$. In particular, [5] 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 [5].

Unfortunately, as opposed to $f(S)$ in (2), the new criterion $-\log p(X|S)$ is symmetric, but not a submodular function. For a function to be submodular, increasing a set size should progressively lead to smaller and smaller increases in the function as new elements are added:

$$f(S \cup \{\mathbf{j}\}) - f(S) \geq f(S \cup \{\mathbf{i}\} \cup \{\mathbf{j}\}) - f(S \cup \{\mathbf{i}\}). \tag{5}$$

It would be sufficient for $-\log p(X_S)$ to be submodular, as the Queyranne algorithm optimizes the symmetric part of a given submodular function, which for $-\log p(X_S)$ is our optimization criterion $-\log p(X_S) - \log p(X_{V \setminus S})$. But, plugging $-\log p(X_S)$ into the above definition would lead to the conclusion that $p(X_\mathbf{j}|X_S) \leq p(X_\mathbf{j}|X_S \cup X_\mathbf{i})$, which is not generally true. While extra information $X_\mathbf{i}$ may decrease the uncertainty about $X_\mathbf{j}$, the conditional probability of a particular observation $X_\mathbf{j}$ can be higher without the knowledge of $X_\mathbf{i}$, if it so happens that the observed X_i contradicts X_j under the model. In practice, however, we expect that when S is close to a subset of the true solution, this will be unlikely and the new elements will be more and more in accordance with the rest of the set as it grows.

Due to these two reasons (relationship to entropy and the typically diminishing punishment on likelihood $\log p(X_S)$ as more elements are added), we expect that minimizing $-\log p(X))$ from (1) using the Queyranne algorithm is a viable strategy in practice.

3 Background on blind source separation

The separation of speech mixtures into individual sources using a single microphone is a hard problem that has generated a lot of interest in the research community. Current approaches attempt to disambiguate the log-spectra representation of the mixture on a time-frequency bin basis. Each bin is subjected to analysis and tagged as belonging to one of the individual sources. The large combinatorial space created by the analysis of the signal in such a small resolution posses a significant challenge to systems attempting to do this kind of separation.

The time-frequency representation of speech signals, the spectrogram, is very sparse: most narrow frequency bands carry substantial energy only during a small fraction of time. Therefore it is rare to encounter two independent sources with large amounts of energy at the same frequency band at the same time. Figure 1 shows a spectrogram representation of a speech signal, where each column depicts the energy content across frequency in a short-time window or time-frame. The value in each cell is actually the log-magnitude of the short-time Fourier transform in deciBels:

$$x_t^k = 20 \log \left(abs \left(\sum_{\tau=0}^{N_F-1} w[\tau] x[t \cdot H + \tau] e^{-j 2\pi \tau k / N_F} \right) \right) \tag{6}$$

where t is the time-frame index, k indexes the frequency bands, N_F is the size of the discrete Fourier transform, H is the hop between successive time-frames, $w[\tau]$ is the N_F-point short-time window, and $x[\tau]$ is the original time-domain signal. The time-frequency masking approach exploits the sparseness characteristic of the time-frequency representation by assigning each time-frequency bin x_t^k (eq. 6) from the mixture signal to one and only one of the sources. Each source in the mixture has a correspondent binary mask with the same dimensions as the time-frequency representation of the mixed signal, where "*one*" in a given source mask indicates that the corresponding time-frequency bin in the spectrogram of the mixture signal belongs to the corresponding speaker. The individual sources are later resynthesized using the masks and the spectrogram and phase information from the original mixed signal.

The large combinatorial space created by the analysis of the signal at such a fine resolution poses a great challenge to systems attempting to do such a separation. In [1] the combinatorial search is restricted by the use of pretrained speaker models, which limits the applicability of the approach to mixtures of sources whose individual properties are known in great detail. In [2], a training session is required to choose the right parameters for a spectral clustering algorithm. Finding clusters among the set of all time-frequency bins requires huge matrices that pose significant numerical problems.

Empirically, time-frequency cells belonging to any particular source occur in large clumps (local regions), and highly-intelligible separation can be achieved by limiting the masks to consist of relatively large, locally-consistent regions of labeling. Moreover the intelligibility of the experiments was greatly increased when the missing regions that were masked by the competing signal [4]. In this paper, we propose a dominant speaker source separation approach, where the spectogram is first segmented using the deformable spectogram model [4] into regions where only one speaker dominates the mixture. This model exploits the high correlation between adjacent frames of the spectogram, shown in many audio signals including speech and musical instruments to model successive spectra as *transformations* of their immediate predecessors. The model can also detect when

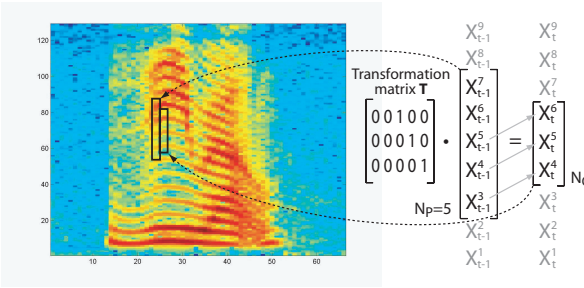

Figure 1: The $N_C = 3$ patch of time-frequency bins outlined in the spectrogram can be seen as an "upward" version of the marked $N_P = 5$ patch in the previous frame. This relationship can be described using the matrix shown.

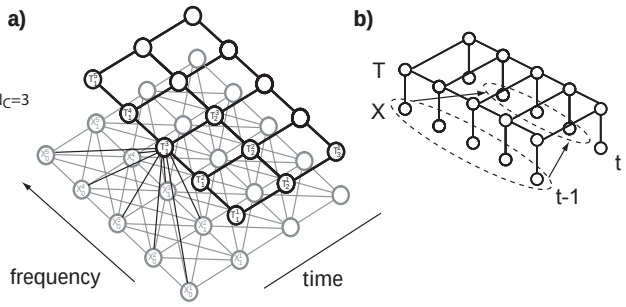

Figure 2: a) Graphical model b) Graphical simplification.

prediction of frames from their context is not possible, marking those frames as boundaries between regions with different patterns on the signal's energy dynamics [4].

Clustering these larger regions is a task several degrees simpler than clustering individual time frequency bins. If the speakers have sufficiently dissimilar characteristics, such as when they belong to different genders, the regions can be clustered [5] using spectral clustering techniques – a set of algorithms that rely on the eigen-structure of a similarity matrix between the points to be clustered to partition the points so that points in the same cluster have high similarity while points in different clusters have low similarity [2]. However, in cases such as [2] where a spectral clustering approach was used to separate speakers using speaker dependant similarity features (e.g. pitch), these techniques fail to cluster dominant regions belonging to the same speaker when the different speakers' voices are similar.

A plausible alternative is to apply a model-based clustering approach where constraints on the form that the mixture components can take are encoded in models which capture the statistical distributions of the features of the mixture components. Previous related work has used probabilistic speech models to guide single microphone source separation tasks [1], however the models used in that approach were speaker-dependant models which would also encounter problems if the speakers are sufficiently similar. Furthermore, the use of pretrained speaker models greatly limits the applicability of the approach.

In this work, we propose to evaluate the quality of the speaker segmentations by the use of a generic speech model, like the one obtained through the training of a regular speech recognizer, trained with clean speech from sources other than the ones that are intended to be separated. Even with the local preclustering of individual time frequency bins into N dominant speaker regions for m speakers, an exhaustive search for the best partition will require $m^N/(m-1)!$ evaluations of the speech recognizer, a task that quickly becomes unbearable for most values of N and m. To tackle this problem we perform the search using the Q-clustering algorithm [5] which can approximate the optimal solution in polynomial time.

4 Deformable spectrograms and dominant speaker segmentation

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 [4]. Figure 1 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.

$$\vec{X}_t^{[k-n_C,k+n_C]} \approx \vec{T}_t^k \cdot \vec{X}_{t-1}^{[k-n_P,k+n_P]} \tag{7}$$

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 1 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. 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 2. 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. Each $N_C \times N_P$ transformation matrix $\vec{T}$ is of the form:

$$\begin{pmatrix} \vec{w} & 0\ 0 \\ 0 & \vec{w} & 0 \\ 0\ 0 & \vec{w} \end{pmatrix} \tag{8}$$

i.e. each of the N_C cells at time t predicted by this matrix is based on the same transformation of cells from $t-1$, translated to retain the same relative relationship. The complete set of $\vec{w}$ vectors is composed by upward/downward shifts of whole bins, which when applied to the matrix format on eq.8, results in a transformation matrix similar to the one in 1, .i.e. $\vec{w}$ = [0 0 1] in the figure. Given the probabilistic nature of the model an apparently limited set of transformation, like the ones described by pure shifts, can have a wide range of representational capabilities. For example a transformation like the one described by [0 0 0 .25 .75] with probability one is equivalent to trasformations described by [0 0 0 1 0] and [0 0 0 0 1] with probabilities 0.25 and 0.75 respectively. The length N_W of the transformation vectors defines the supporting coefficients from the previous frame $\vec{X}_{t-1}^{[k-n_W,k+n_W]}$ (where $n_W = (N_W - 1)/2$) that can "explain" x_t^k. The "local-likelihood" potential between the time-frequency bin x_t^k, its relevant neighbors in frame t, its relevant neighbors in frame $t-1$, and its transformation node T_t^k has the following form:

$$\psi\left(\vec{X}_t^{[k-n_C,k+n_C]}, \vec{X}_{t-1}^{[k-n_P,k+n_P]}, T_t^k\right) = \mathcal{N}\left(\vec{X}_t^{[k-n_C,k+n_C]}; \vec{T}_t^k \vec{X}_{t-1}^{[k-n_P,k+n_P]}, \Sigma^{[k-n_C,k+n_C]}\right) \tag{9}$$

The diagonal matrix $\Sigma^{[k-n_C,k+n_C]}$, which is learned, has different values for each frequency band to account for the variability of noise across frequency bands.

Two Layer Source-Filter Transformations

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 X as the sum of variables F and 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, where variables F and H in the model are hidden, while, as before, X can be observed or hidden. The two layers are iteratively and approximately inferred using loopy belief propagation as described in [4]. The first row of figure 3 shows the decomposition of a speech signal into harmonics and formants components, illustrated as the means of the posteriors of the continuous hidden variables in each layer. Figure 3 Second Row a) shows the spectrogram on the first row with deleted regions;in b) the regions have been filled via inference in a single-layer model. Notice that since the formant motion does not follow the harmonics, the formants are not captured in the reconstruction and hence the need for two layers. In c) the two layers are first decomposed and then each layer is filled in; the figure shows the addition of the filled-in version in each layer.

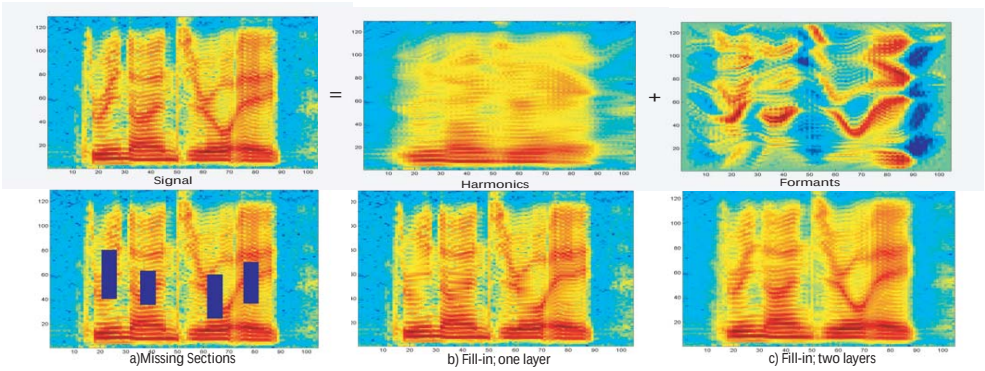

Figure 3: First Row.- Harmonics/Formants decomposition (posterior distribution means). Second Row (a) Spectrogram with deleted (missing) regions. (b) Filling in using a single-layer transformation model. (c) Results from the two-layer model.

Subband transition boundary detection

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. Figure 4, shows the transition boundaries obtained by the model for a female-female mixture of two speakers.

Dominant speaker transition boundary detection

In [5], for a set of 200 artificially mixed mixtures of two speakers, divided in four categories: 50 female-female, 50 male-male, 50 male-female and 50 with the same speaker voicing different utterances, the correlation between these transition boundaries and the actual boundaries where changes on dominant speakers ocurr was measured by the use of the basic measures used in evaluating search strategies: precision = 97.25 and recall = 63.88. Recall is the ratio of the number of relevant records, in this case the ground truth boundaries, retrieved to the total number of records. Precision is the ratio of the number of relevant records retrieved to the total number of records retrieved, both relevant and irrelevant. The precision values were homogeneous between the different kind of mixtures, even for the ones with the same speaker. The model does well regardless of the nature of the speakers because it discovers interruptions in the energy pattern of the signal without relying on any source dependant features. On the other hand, precision results are not as good. This is in part, because transitions between voiced and unvoiced data for the same speaker are also detected but they are not considered a "ground truth" dominant speaker change boundary. Two alternative approaches for the dominant speaker transition boundary detection are also presented, one using a strong speaker dependant feature such as pitch and another using a minimum description length criteria to segment the regions, both approaches considerable underperform the detection obtained by our proposed approach.

5 Clustering single-speaker regions

Since the deformable spectrogram-based segmentation has high recall values we can be pretty certain that the signal is segmented into single-speaker regions (albeit a large number). Clustering these regions is a task several degrees simpler than clustering individual time frequency bins. When the speakers are sufficiently dissimilar the regions can be clustered using spectral clustering methods that use similarity measures within the regions to be clustered. However when the speakers are very similar like in the extreme case of figure 4 that depicts the dominant speaker segmentation of the spectogram of a mixture of the same speaker voicing different utterances, there is not an affinity based clustering method capable of grouping the regions that correspond to the different utterances.

Therefore, as argued above, we instead use a semantic clustering approach that consist of evaluating the quality of the speaker segmentations by the use of a generic speech model. The idea behind this approach is that the feature constraints encoded in the acoustic model along with the language model constraints can efficiently help to discard poor segmentations. However exhaustive search for the best partition will require $m^N/(m-1)!$ evaluations of the speech recognizer, for a mixture of m speakers presegmented into N dominant speaker regions. To tackle this problem we perform the search using the Q-clustering algorithm [4] which terminates in polynomial time.

In the following, R_i denotes a small single-speaker region. V denotes the complete spectrogram composed of N non overlapping single-speaker regions, $V = \sum_{i=1}^{N}(R_i)$. $R_i^{'} = V \backslash R_i$ denotes all the regions in V but R_i. S denotes a union of individual regions $S = \sum_{i \in G}(R_i)$, and $S' = V \backslash S$, denotes all the regions in V but the ones in S. $\mathcal{L}(S) = logp(X_S|theta)$ denotes the loglikelihood obtained when the decoder is applied to the signal part X_S (marginalizing over the rest of the spectrogram as hidden). As indicated above, regions R_i are obtained by segmentation into regions that do not violate the smoothness constraints of the deformable spectrogram model, and $p(X_S|theta)$ is based on a generic speech model trained on a large corpus of data as described in the next section. $\mathcal{L}^T(S) = \mathcal{L}(S) + \mathcal{L}(V \setminus S)$ denotes the total loglikehihood given the decoder for the spectogram partition $P = (S, V \setminus S)$.

Start with the N original regions.
$N_{new} = N$;
While $N_{new} \geq 2$.
  $S = \emptyset$.
  N_{tested} = 0;
  While $N_{tested} \leq N_{new} - 2$.
    For all $R_i \in V \setminus S$
      Compute $\mathcal{L}^T(S + R_i)$
    end
    $R_i \leftarrow argmax_{R_j \in V \setminus S}\ (\mathcal{L}^T(S + R_j))$; $S \leftarrow S + R_j$;
    $N_{tested} = N_{tested} - 1$
  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 $(R_i, V \setminus R_i)$ on 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.

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 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.

Each mixture was first pre-segmented using the deformable spectogram model into regions with smooth energy patterns. Then, the algorithm in Section 5 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 5 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

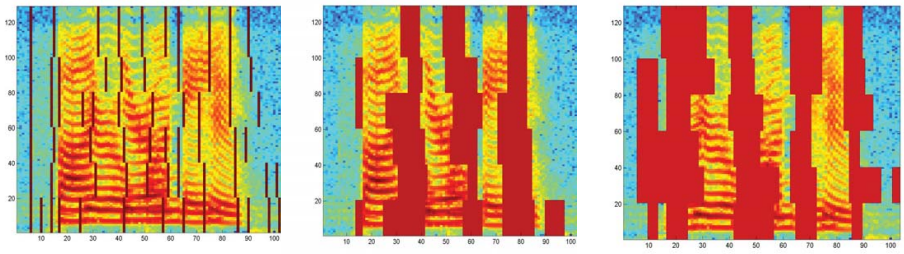

Figure 4: Detected boundaries: resulting in over 60 presegmented regions and estimated partition. (Presegmented silence regions are including in both signals)

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	82.17%
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. Figure shows an example of such a mixture with its correspondent partitionP_{est}. In the suplemental material, we provide wav files with examples of speech separation using our algorithm.

References

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

[2] F. Bach and M. Jordan, "Blind one-microphone speech separation: A spectral learning approach", NIPS, 2004.

[3] M. Cooke, "10+1 perspectives on speech separation and identification in listeners and machines", Workshop on Speech Separation and Comprehension in Complex Acoustic Environments, 2004.

[4] M. Reyes-Gomez, "Statistical Graphical Models for Scene Analysis, Source Separation and other Audio Applications", Ph.D. Thesis, Columbia University, 2005.

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

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