STEREOPHONIC SPECTROGRAM SEGMENTATION USING MARKOV RANDOM FIELDS

Minje Kim[1], *Paris Smaragdis*[2], *Glenn G. Ko*[3] *and Rob A. Rutenbar*[4]

University of Illinois at Urbana-Champaign, Department of Computer Science[1,2,4]
University of Illinois at Urbana-Champaign, Department of Electrical and Computer Engineering[2,3]
Adobe Systems Inc.[2]
{minje[1], paris[2], giko[3], rutenbar[4]}@illinois.edu

ABSTRACT

There is a good amount of similarity between source separation approaches that use spectrograms captured from multiple microphones and computer vision algorithms that use multiple images for segmentation problems. Just as one would use Markov random fields (MRF) to solve image segmentation problems, we propose a method of modeling source separation using MRFs, and then solving such problems via common MRF inference methods. To this end, as a preprocessing, we convert stereophonic spectrograms into a integrated form based on their inter-channel level differences (ILD), which is a procedure analogous to getting a disparity map from stereo images for matching problems. Given the ILD matrix as an observed image, we estimate latent labels which stand for the responsibility of each spectrogram's time/frequency bin to a specific sound source. It is shown that the proposed method shows reasonable separation performance in a variety of mixing environments including online separation and moving sources. We expect this new way of formulating source separation problems to help exploit advantages of probabilistic graphical models and the recent advances in low-power, high-performance hardware suited for such tasks.

***Index Terms*—** Blind Source Separation, Markov Random Fields, Probabilistic Graphical Model, Gibbs Sampling

1. INTRODUCTION

The topic of sound source separation has attracted a lot of research in the audio signal processing and machine learning communities. Using the multichannel formulation, multiple recordings are obtained from an array of microphones, and they constitute combinations of all the sound sources with varying delays and magnitudes according to the mixing room environment.

When little or no information is provided about the source or the mixing process, this problem is often called blind source separation (BSS) [1]. The goal of BSS is to recover the sources from the observed mixtures given some source-specific assumptions, for instance, statistical independence.

There have been two major approaches to BSS: inverting the mixing process and masking. In the first case one can separate sources from a mixture if their mixing process is represented by a known mixing matrix, where each element of that matrix represents a scaling factor between a particular pair of source and a recording. One of the most successful techniques for doing this uses independent component analysis (ICA), which operates on the assumption that the original source signals are statistically independent to each other [2]. In the case where the mixing process includes room reverberations, the problem is often reformulated as a convolutive mixture problem which can be solved using a variety of methods [3]. The basic setup of these kinds of problems is the *overdetermined* case which needs at least as many microphones as the number of sources.

However, in practice we often have less microphones than sources or even only one microphone. This is referred to as *underdetermined / single-channel* mixing problem. This is a harder problem than the overdetermined case as it involves less amount of information to estimate the mixing matrix, which is apt to be ill-posed. Another common BSS approach, time-frequency masking [4, 5, 6], is one solution to this kind of problem setting. Binary masking methods seek time/frequency binary masking values, which denote whether each recorded spectrogram's pixels belong to one source or the other. This process sidesteps some of the convolutive BSS problems since it does not model the mixing process explicitly, and instead makes the assumption that the spectrogram of each source is sparse enough not to have significant energy at the same time/frequency bins as the spectrograms of the other sources.

One recent approach to BSS factorizes the magnitude spectrogram into two lower-ranked matrices: basis vectors and corresponding encodings [7]. If the basis vectors contain source-specific spectral information, we can group them, and

This work was supported in part by the C2S2 Focus Center (one of the six research centers funded under the Focus Center Research Program, which is a Semiconductor Research Corporation program), and by the Intel Science and Technology Center for Embedded Computing (ISTC-EC).

978-1-4673-1026-0/12/$31.00 ©2012 IEEE

then reconstruct each source by multiplying only the corresponding group of basis vectors and corresponding encodings. Nonnegative matrix factorization (NMF) [8, 9] has been used for this task as its nonnegative constraints and sparse representation fit well with the magnitude spectrograms of audio signal. We can also classify these NMF-based techniques as masking methods, because the source spectrogram is reconstructed by multiplying each input mixture spectrogram pixel with the pixel-wise proportion of the recovered source to the recovered mixture, which can be seen as a soft masking process.

The proposed work can be seen as a masking method, too. We concentrate on the situation where the number of mixture signals is less than the sources, for instance cellphone with only two microphones. To estimate the mask values, we propose a novel source separation method utilizing a class of probabilistic graphical models called Markov random fields (MRFs). An MRF has been widely used to solve various problems in the field of computer vision, natural language processing and bioinformatics. In this paper, we apply MRF to model an underdetermined source separation problem and perform inference to find the most probable source labels for all recorded time/frequency bins. Though this maximum *a posteriori* (MAP) estimation problem is designed to be similar to image segmentation problems, where users are preferred to mark regions and assist the labeling, it is basically an unsupervised pixel clustering problem without the any input from users. Due to that difference, we propose a parameter readjustment procedure that corresponds to M-step of EM algorithm and replaces the need for user assistance.

The paper is organized as follows. In section 2 we define BSS as a Bayesian labeling task. A quick overview of MRF and an inference method, Gibbs sampling, is given in 3 and 3.1. Section 4 covers formulation of the energy functions of the proposed MRF. The experimental setup is explained in section 5.1. It includes all the assumptions we made about the geometric configuration of sources and sensors in order to perform and evaluate our method. Finally, we discuss our experimental results in 5.2.

2. BSS AS A BAYESIAN LABELING TASK

In order to make this introduction more accessible we will formulate the our model using only a two-microphone formulation, however the introduced technique can scale to more microphones if needed. Let X^L and X^R be the observed magnitude spectrograms of the mixture signals, x^L and x^R as recorded from two microphones. The goal of binary masking is to estimate the stereophonic target source $S^{1,L}$ and $S^{1,R}$, by masking X^L and X^R with appropriate masking matrices using

$$S^{1,L} \approx \hat{S}^{1,L} = X^L \odot C, \qquad S^{1,R} \approx \hat{S}^{1,R} = X^R \odot C. \quad (1)$$

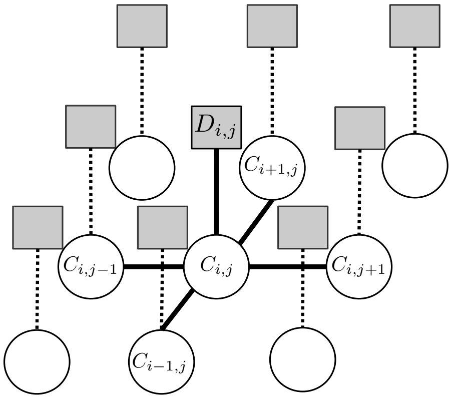


Fig. 1. A pair-wise MRF example for BSS using ILD input D.

Consequently, we can easily get the sum of all the other interference sources, $S^{2,L}$ and $S^{2,R}$, by using the complement of C,

$$\begin{aligned} S^{2,L} &\approx \hat{S}^{2,L} = X^L \odot (1 - C), \\ S^{2,R} &\approx \hat{S}^{2,R} = X^R \odot (1 - C). \end{aligned} \quad (2)$$

In (1) and (2) , C is a binary mask that has the same size as the input spectrograms X^L and X^R. Multiplication $\odot$ is performed in an element-wise manner. It is common to assume that each element of C, which can be indexed with frequency and time domain indices, i and j, is binary, $C_{i,j} \in \{0, 1\}$. However, we can relax this constraint by allowing it to be a soft mask: $0 \leq C_{i,j} \leq 1$.

We can think of this problem as a Bayesian labeling problem, where we want to find labels C that maximize the posterior probability given the observed data D,

$$\arg\max_C P(C|D) = \arg\max_C P(D|C)P(C).$$

Where we construct the input matrix D by extracting pixel-wise inter-channel level differences (ILD) from the two magnitude spectrograms using the following equation,

$$D_{i,j} = 10 \log \frac{(X^L_{i,j})^2}{(X^R_{i,j})^2}.$$

Note that this feature extraction process is comparable to getting a disparity map from stereo images in the stereo matching problem in vision. A more straightforward analogy would be calculating source-specific delays from the two recordings, but we used ILD instead of delays since time/frequency bin delays are not directly observable through spectrograms.

3. MARKOV RANDOM FIELDS

MRF is a class of undirected graphical models, which provides an effective way to model complex systems as simpler local subsets and provides intuitive structure for modeling probability distributions [10]. The vertices of the MRF correspond to a set of binary variables that make up the binary mask we seek to find. We represent that set of vertices with the labeling variables $C_{i,j} \in \mathcal{V}$. We also define a set of pairwise edges, $\mathcal{E}$, resulting in a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$.

Each node is associated with a node potential $\phi(C_{i,j}, D_{i,j})$ while an edge between $C_{i,j}$ and its neighbor $C_{k,l}$ has a corresponding edge potential $\phi(C_{i,j}, C_{k,l})$, where (k,l) is a pair of indices that are included in the set of neighbors of (i,j), $\mathcal{N}_{i,j}$. Figure 1 depicts an MRF with four neighbors. Note that $D_{i,j}$ are constant observations.

Those potentials consist of *unnormalized probabilities* since they can be seen as functions that actually refer to conditional probabilities after the proper normalization:

$$\begin{aligned} P(D|C)P(C) &= \prod_{i,j} P(D_{i,j}|C_{i,j}) \prod_{k,l \in \mathcal{N}_{i,j}} P(C_{i,j}|C_{k,l}) \\ &= \frac{1}{Z} \prod_{i,j} \phi(C_{i,j}, D_{i,j}) \prod_{k,l \in \mathcal{N}_{i,j}} \phi(C_{i,j}, C_{k,l}), \end{aligned} \tag{3}$$

where the first equality implies Markov property, and Z represents a normalization constant. We can think of the node potential as corresponding to the conditional probability of getting the observed value $D_{i,j}$ given the label, while the edge potential is related to the probability of $C_{i,j}$ given its neighbors.

Originating from statistical physics, MRFs are often described in terms of energy instead of probability distribution. Hence, we formulate the MRF labeling problem as a sum of different energy formula as shown below,

$$\epsilon(C, D) = \sum_{i,j} \epsilon(C_{i,j}, D_{i,j}) + \sum_{k,l \in \mathcal{N}_{i,j}} \epsilon(C_{i,j}, C_{k,l}), \tag{4}$$

by taking the negative logarithm of corresponding potential functions in (3). Note that now we need to minimize the objective function instead of maximizing the posterior probability.

3.1. MRF Inference

Although an MRF has a very compact description, various dependencies and local interactions among the variables can get extremely complex. Therefore, inference is often intractable. As a result, there has been extensive research on using approximate inference methods on MRFs, that closely approximate the original distribution yet simplify computations.

In this paper we do not limit the way of solving this MRF problem to a specific inference method. In fact, we checked that several well-known inference algorithms, such as graph cuts and loopy belief propagation, give converged results. However, we presents the inference results from Gibbs sampling [11] as we need the marginal probability of each node to be used as soft masks. Given random variables x and y, the main idea behind the Gibbs sampler is that it is much easier to find the marginal distributions $p(x)$ and $p(y)$ using a sequence of conditional distributions $p(x|y)$ and $p(y|x)$ than it is by using joint distribution $p(x,y)$ and integrating over each variable separately. The Gibbs sampler generates a sample of a desired variable from its conditional probability given all the other variables, or only its neighbors with Markov property.

After sufficient iterations, the sampler accumulates a number of samples that provide converged approximate marginal distributions of the desired variables. The probability of each label represents a time/frequency-specific soft mask. Note that we can make a hard decision by selecting the label value with maximum probability. In this work, we also compare the hard and soft decision results to help readers anticipate separation results with discrete labels from other inference methods, such as graph cuts.

4. MRF SETUPS FOR SOURCE SEPARATION

For a given set of ILD values in the form of a matrix D, the goal of BSS in terms of Bayesian labeling, is to infer random variables C. Using the results in the energy form in (4), we seek a combination of labels that minimizes the total cost of energy terms.

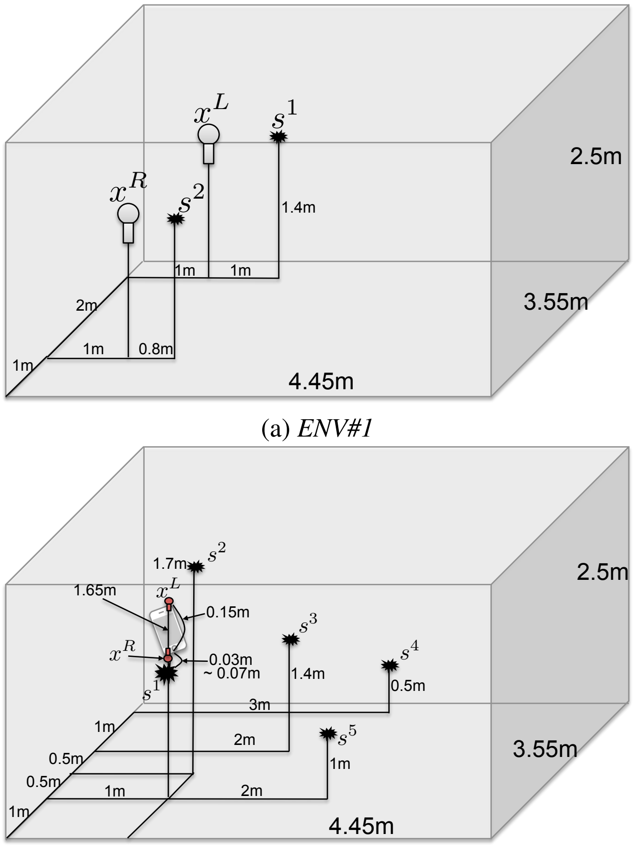
To this end, two cost terms are defined: data and smoothness costs. We simply assume that the probability of observing a ILD value, $D_{i,j}$, follows a source-specific Gaussian distribution similarly to [6]. Thus, the data cost is defined with Euclidean distances,

$$\epsilon(C_{i,j}, D_{i,j}) = \left\{ \begin{array}{rl} (\mu_0 - D_{i,j}(1 - C_{i,j}))^2/\sigma_0^2 & \text{if } C_{i,j} = 0 \\ (\mu_1 - D_{i,j}C_{i,j})^2/\sigma_1^2 & \text{otherwise} \end{array} \right. ,$$

between the sample means of the source-wise ILDs, μ_0 and μ_1, and the given observation, $D_{i,j}$, with pre-defined variances, σ_0^2 and σ_1^2. We can also define a smoothness cost based on the common assumption that the neighboring pixels tend to belong to the same source group:

$$\epsilon(C_{i,j}, C_{k,l}) = (C_{i,j} - C_{k,l})^2/\sigma_{\mathcal{N}}^2,$$

where $\sigma_{\mathcal{N}}^2$ is a pre-defined variance of the distribution of neighbors. In this work, we used a simple neighboring system with four adjacent nodes, which proved sufficient in demonstrating the advantage of employing prior information through pair-wise neighboring. In this case, we can interpret the vertical edges as representing the continuity of broadband noises of a source, e.g. speech consonants, or smearing of frequency sub-bands. On the other hand, horizontal edges are responsible for the temporal steadiness of the more temporally continuous sources, e.g. vowels. See section 5.2 to see the separation performance improvement by introducing this set of dependencies.



(a) *ENV#1*

(b) *ENV#2* and *ENV#3*

Fig. 2. Pictorial representation of the simulated mixing environments. In (b), channel noises are omitted.

Table 1. Separation results in the *ENV#1* situation.

		Mixture	Hard		Soft	
			Node	Node +edge	Node	Node +edge
SDR	s^1	-1.40	7.32	8.14	8.52	8.92
	s^2	1.40	8.72	9.54	9.92	10.32
ISR	s^1	15.94	10.45	10.33	11.72	10.86
	s^2	17.52	14.52	17.34	14.28	18.17
SIR	s^1	-1.12	13.28	16.96	12.08	17.61
	s^2	1.65	11.60	11.79	12.29	12.20
SAR	s^1	231.16	12.16	13.03	14.90	14.26
	s^2	231.16	13.63	14.97	16.35	16.22

As used in computer vision, it is common for image segmentation applications to request users to provide an initial seed segmentation. For example, users can mark certain foreground and background pixels with different colors to help consist initial guesses about underlying distributions, respectively [12]. Once this is done, the initial guesses are considered as constants during the inference procedure. However, in the proposed model this user intervention is not realistic for several reasons: users are unfamiliar to spectrograms, there are unclear object boundaries, the translucent nature of the mixtures makes this a difficult hand-labeling process, this is hard to impose on a real-time system, etc.

We therefore adopt an EM-like approach to appropriate estimation of sample means and variances. We start with rough guesses of the parameters based on the geometric information of sensors and sources. For instance, in a cellphone situation, we can expect that the owner's mouth is roughly located at a known spot in relation to the available microphones. Conversely, ambient sound sources will be probably randomly distributed around the space, so they tend to construct a flat distribution of ILD centered around zero. Although these guesses provide at least a crude estimate of the needed parameters, we can improve them after inference similarly to performing an M-step in the EM-algorithm:

$$\mu_0 = \frac{1}{N}\sum_{i,j} D_{i,j}(1-C_{i,j}), \qquad \sigma_0 = \frac{1}{N}\sum_{i,j}(D_{i,j}-\mu_0)^2$$

$$\mu_1 = \frac{1}{N}\sum_{i,j} D_{i,j}C_{i,j}, \qquad \sigma_1 = \frac{1}{N}\sum_{i,j}(D_{i,j}-\mu_1)^2,$$

where N is the number of pixels. Note that the labels $C_{i,j}$ are the running estimate from the previous inference. We do the inference again using the updated data cost based on the new parameters, and repeat until convergence.

5. EMPIRICAL RESULTS

5.1. Experimental Setups

- Input signals: we chose two speech signals, one male and one female, from the TIMIT corpus [13] as our source signals. They are sampled with 16,000Hz sampling rate and encoded with 16bit PCM.
- Mixing process: we evaluated the proposed model with five different convolutive mixing environments:
  - *ENV#1*: Two sources with two corresponding sensors. This assumes that there is no noise or interference, but it is still a realistic convolutive mixture with simulated room reverberations [14]. See Figure 2 (a).
  - *ENV#2*: A comprehensive mixture of a dominant source and five interferences in the cellphone environment. We assume that there are two microphones on the opposite diagonal ends of a cellphone. The goal is to separate the dominant speech out of the convolutive mixture with various ambient interferences: another speech, a blender, a rolling can, a washer, and additional microphone noises. The distance between the main source and the main sensor is 3cm. See Figure 2 (b) for the detailed configuration.

Table 2. Separation results in the *ENV#2* and *ENV#3* situation.

		Mixture		Hard				Soft			
				3cm		7cm		3cm		7cm	
		3cm	7cm	Node	Node +edge	Node	Node +edge	Node	Node +edge	Node	Node +edge
SDR	Dominant source	6.75	0.06	15.35	16.71	8.23	9.93	15.72	17.13	8.08	10.42
	Interferences	-6.75	-0.06	8.59	9.94	8.16	9.85	8.95	10.35	8.01	10.34
ISR	Dominant source	24.80	18.17	28.55	26.75	22.35	21.20	28.73	26.83	22.38	21.13
	Interferences	11.49	18.20	11.01	14.78	9.95	12.90	10.82	14.73	9.05	12.97
SIR	Dominant source	6.83	0.20	17.90	21.73	10.09	12.97	17.72	21.67	9.14	12.98
	Interferences	-6.37	0.08	23.23	20.25	24.01	21.59	23.36	20.35	23.30	21.44
SAR	Dominant source	226.46	219.36	19.38	18.96	13.57	13.48	20.65	19.69	15.58	14.54
	Interferences	226.46	219.36	11.52	11.61	11.98	12.64	12.35	12.32	12.49	13.70

Table 3. Separation results in the *ENV#4* situation. In the second column, 'S' and 'I' stand for the dominant source and interferences, respectively.

		Mixture	Hard		Soft	
			Batch	Online	Batch	Online
SDR	S	14.57	20.44	22.97	21.22	23.38
	I	-14.57	5.86	8.39	6.64	8.80
ISR	S	34.62	23.51	35.42	24.41	35.39
	I	6.30	15.98	13.76	15.81	13.73
SIR	S	14.62	30.83	28.46	30.63	28.42
	I	-13.58	8.39	20.83	9.27	20.73
SAR	S	274.44	24.19	24.75	25.09	25.39
	I	274.44	9.64	9.43	10.40	10.02

Table 4. Separation results in the *ENV#5* situation.

		Mixture	Hard	Soft
SDR	Dominant Source	12.68	19.41	19.78
	Noise	-12.68	6.58	6.94
ISR	Dominant Source	30.73	31.25	31.23
	Noise	5.07	12.58	12.48
SIR	Dominant Source	12.75	25.83	25.72
	Noise	-11.43	18.18	18.14
SAR	Dominant Source	238.93	20.89	21.45
	Noise	238.93	6.34	6.86

  - *ENV#3* Same as *ENV#2*, but now the dominant source is 7cm away.
  - *ENV#4*: Same as *ENV#2*, but now the dominant source is moving. It starts to speak 1cm away from the main microphone, and then changes its position to 10cm away.
  - *ENV#5*: Same as *ENV#2*, but now the channel noise is the lone interference. This environment is to check whether the Gaussian assumption about the ILDs holds in a very simple case, such as a very quiet environment.

- STFT: size of each frame and number of FFT were set to 1024 points. We adopted sine square window with 50% overlap to avoid block artifacts after applying masks.

5.2. Separation Performance

We adopted an objective separation quality evaluation method, *BSS_EVAL*, proposed in [15]. Because dereverberation is not a goal of this BSS system, we used the room-filtered sources instead of clean source signals as the input of *BSS_EVAL*. Hence, clean source signals are convolved with room impulse responses of the configuration in Figure 2. We use the signal-to-distortion ratio (SDR) which represents the presence of both interference and artifacts, which are the main elements that can have negative effects on subsequent processing, such as speech recognition.

- *ENV#1*: Table 1 shows the separation results under the mixing environment *ENV#1*. Although the soft decision mask generally outperforms the hard decision mask, the hard decision mask results are still acceptable. Note also that the edge potentials with the simple four-neighboring system improve separation quality.

- *ENV#2* and *ENV#3*: We can also check that the proposed method works well in the simulated cellphone environment in Table 2. Furthermore, SDR improvements of the dominant source from the mixture SDR lie between +8 and +10.5 regardless of the source position, which support the robustness of the separation scheme. Note that we assume a uniform smoothness cost for the sum of interferences as we have no preference about their relationship to neighbors.

- *ENV#4*: We tested an online version of the proposed algorithm where we deal with only the latest $N+1$ spectra at a time to construct an MRF. We set $N = 10$ for this experiment (0.384 seconds). First, we bypass the first N frames without separation, but set aside them to separate $(N + 1)$'th frame. After we collect first $(N + 1)$ frames, we build the first MRF. Once we have converged labels of the $(N+1)$ spectrums, we can start masking the $(N+1)$'th spectrum. Then, we discard the first one of the $(N + 1)$ spectrums. Remaining 2nd to $(N + 1)$'th spectrums are used to build a successive MRF when we acquire $(N + 2)$'th spectrum.

  Note that we can initialize parameters and labels of current MRF using the previous results to expedite convergence. This is rational since the first N spectra of the current MRF and the last N spectra of the previous one point to the same input data. If we can tolerate the first N unseparated frames, this mechanism provides an online (or real-time with proper hardware support) separation.

  To stress the performance of the online approach we also radically changed the position of the dominant source from 1 cm to 10 cm in *ENV#4*. The 'batch' approach results in an inaccurately aggregated estimation of the dominant source distribution, which unnecessarily averages all the movements. However, the 'online' method can adjust the model to these dynamics, so that it could earn more than 2 dB improvement in SDR in Table 3

- *ENV#5*: One can argue that this model seems to be overly complex for a simpler situation, where there are not enough interferences to consist a distinctively flat Gaussian distribution. We addressed this situation by adding only channel noises to the reverberated dominant source. Based on the assumption that the supplementary microphone is cheaper to produce more noise than the main one, we doubled the variance of Gaussian noise of the supplementary channel. As the results in Table 4 show, the proposed method also works in the simple Gaussian noise case. However, it is true that applying the BSS system degrades the sound quality if the level of channel noise is negligibly low.

6. CONCLUSION

In this paper, we proposed an MRF-based BSS system, which can be seen as an ILD matrix segmentation scheme. The proposed EM-like update along with the conventional MRF inference provided reasonable separation results in various mixing environments. Specifically, we verified that a relatively small amount of spectra (less than one second) are enough to get good separation results even when the source is moving.

The MRF formulation is especially attractive for building real-time systems, .e.g on cell phones, since such algorithms can be very efficiently implemented on special hardware with relatively limited cost. Our future work includes exploring the hardware options that can facilitate such algorithms and producing low-power hardware that can be easily embedded in common devices.

7. REFERENCES

[1] J. F. Cardoso, "Blind signal separation: Statistical principles," *Proceedings of of the IEEE, Special Issue on Blind Identification and Estimation*, vol. 86, no. 10, pp. 2009–2025, Oct. 1998.

[2] A. Hyvärinen, J. Karhunen, and E. Oja, *Independent Component Analysis*. John Wiley & Sons, Inc., 2001.

[3] P. Smaragdis, "Blind separation of convolved mixtures in the frequency domain," *Neurocomputing*, vol. 22, pp. 21–34, 1998.

[4] N. Roman, D. L. Wang, and G. J. Brown, "Speech segregation based on sound localization," *Journal of the Acoustical Society of America*, vol. 114, no. 4, pp. 2236–2252, 2003.

[5] Ö. Yilmaz and S. Rickard, "Blind separation of speech mixtures via time-frequency masking," *IEEE Transactions on Signal Processing*, vol. 52, no. 7, pp. 1830–1847, 2004.

[6] M. Kim, S. Beack, K. Choi, and K. Kang, "Gaussian mixture model for singing voice separation from stereophonic music," in *Audio Engineering Society Conference: 43rd International Conference*, Kyoto, Japan, 9 2011.

[7] P. Smaragdis and J. C. Brown, "Non-negative matrix factorization for polyphonic music transcription," in *Proceedings of IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, New Paltz, NY, 2003, pp. 177–180.

[8] D. D. Lee and H. S. Seung, "Learning the parts of objects by non-negative matrix factorization," *Nature*, vol. 401, pp. 788–791, 1999.

[9] ——, "Algorithms for non-negative matrix factorization," in *Advances in Neural Information Processing Systems (NIPS)*, vol. 13. MIT Press, 2001.

[10] D. Koller and N. Friedman, *Probabilistic Graphical Models: Principles and Techniques*. MIT Press, 2009.

[11] S. Geman and D. Geman, "Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 6, pp. 721–741, 1984.

[12] Y. Y. Boykov and M. P. Jolly, "Interactive graph cuts for optimal boundary & region segmentation of objects in n-d images," in *Proceedings of the International Conference on Computer Vision (ICCV)*, 2001, pp. 105–112.

[13] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, N. L. Dahlgren, and V. Zue, "Timit acoustic-phonetic continuous speech corpus," *Linguistic Data Consortium, Philadelphia*, 1993.

[14] E. Vincent and D. R. Campbell, "Roomsimove matlab toolbox," 2008. [Online]. Available: http://www.irisa.fr/metiss/members/evincent/software

[15] E. Vincent, C. Fevotte, and R. Gribonval, "Performance measurement in blind audio source separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 4, pp. 1462–1469, 2006.