Modelling Sound Dynamics Using Deformable Spectrograms: Segmenting the Spectrogram into Smooth Regions

Manuel Reyes-Gomez
LabROSA
Department of Electrical Engineering
Columbia University
mjr59@ee.columbia.edu

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

Daniel P.W. Ellis
LabROSA
Department of Electrical Engineering
Columbia University
dpwe@ee.columbia.edu

Abstract

Speech and other natural sounds show high temporal correlation and smooth spectral evolution punctuated by a few, irregular and abrupt changes. We model successive spectra as *transformations* of their immediate predecessors, capturing the evolution of the signal energy through time. The speech production model is used to decomposed the log-spectrogrum into two additive layers, which are able to separately explain and model the evolution of the harmonic excitation, and formant filtering of speech and similar sounds. We present results on a speech recognition task, that suggest that the model discovers a global structure on the dynamics of the signal's energy that helps to alleviate the problems generated by noise interferences. The model is also used to segment mixtures of speech into dominant speaker regions on a unsupervised source separation task. Results on: http://www.ee.columbia.edu/~mjr59/def_spec.html

1 Introduction

In many audio signals including speech and musical instruments, there is a high correlation between adjacent frames of their spectral representation. Using the common source-filter model for such signals, we devise a layered generative graphical model that describes these two components in separate layers: one for the excitation harmonics, and another for resonances such as vocal tract formants. Our approach explicitly models the self-similarity and dynamics of each layer by fitting the log-spectral representation of the signal in frame t with a set of transformations of the log-spectra in frame $t-1$. Early developments of this work was presented in [1]. In this paper we present results on applications that were

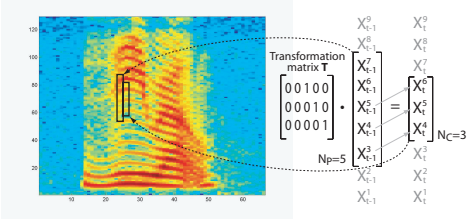

Figure 1: The $N_C = 3$ patch of time-frequency bins outlined in the spectrogram can be seen as an "upward" version of the marked $N_P = 5$ patch in the previous frame. This relationship can be described using the matrix shown.

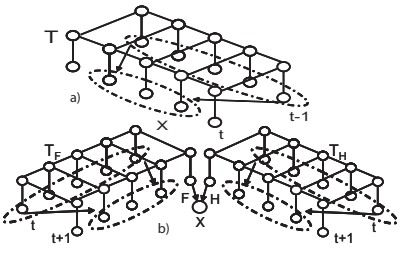

Figure 2: a) Model graphical simplification b) Two layers model.

mentioned as future work on those previous publications.

2 Spectral Deformation Model

Many audio signals, including speech and musical instruments, have spectral representations that show high correlation between adjacent frames. We propose a model that 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.

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. The value in each cell is actually the log-magnitude of the short-time Fourier transform; in decibels, $x_t^k = 20log(abs(\sum_{\tau=0}^{N_F-1} w[\tau]x[\tau - t \cdot H]e^{-j2\pi\tau k/N_F}))$, 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. 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.

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

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, 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 factor graph representation [2] of a section of the model, is shown in figure 3. A factor graph has a variable node (circle) for each variable X_i, and a factor node (square) for each local function ψ_{X_S}. Nodes $\mathcal{X} = \{x_1^1, x_1^2, ..., x_t^k, ..., x_T^K\}$ represent all the time-frequency bins in the spectrogram. For now, we consider the continuous nodes $\mathcal{X}$ as observed. Discrete nodes $\mathcal{T} = \{T_1^1, T_1^2, ..., T_t^k, ..., T_T^K\}$ index the set of transformation matrices used which consist in simple upward and downward shifts.

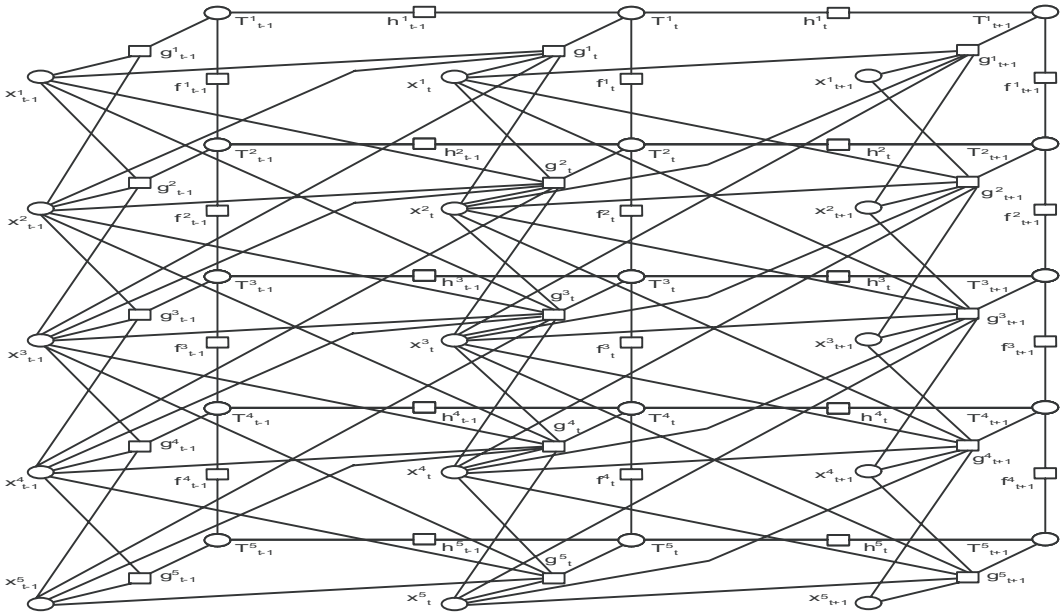

Figure 3: Factor Graph of the Model

The potential: (Function node g_t^k on figure 3)

$$\psi\left(\vec{X}_t^{[k-n_C,k+n_C]}, \vec{X}_{t-1}^{[k-n_P,k+n_P]}, T_t^k\right) = \mathcal{N}\left(\vec{X}_t^{[k-n_C,k+n_C]}; \vec{T}_t^k \vec{X}_{t-1}^{[k-n_P,k+n_P]}, \Sigma^{[k-n_C,k+n_C]}\right) \quad (2)$$

imposes restrictions on the data, such that eq (1) is satisfied as well as possible.

The diagonal matrix $\Sigma^{[k-n_C,k+n_C]}$, which is learned, accounts for the variability of noise across frequency bands. The pairwise horizontal and vertical potentials $\psi_{hor}(T_t^k, T_{t+1}^k)$ and $\psi_{ver}(T_t^k, T_t^{k+1})$, ($h_t^k$ and f_t^k, respectively, on figure 3), favored changes in transformations that are consistent with the current motion of the energy.

When nodes $\mathcal{X}$ are fully observed, inference of the transformations is intractable, and it is approximated using Loopy Belief Propagation [3, 4], which consists on applying the sum-product algorithm message-passing rules on models with loops [2]. The sum-product algorithm computes marginal functions using the distributive law to simplify the summations and reuse intermediate partial sums. The update rules for those messages are defined as: Variable x to local function f message : $m_{x\to f}(x) = \prod_{h\in g_x\setminus f} m_{h\to x}(x)$, where g_x represents all the functions that have x as one of its arguments. Local function f to variable x message: $m_{f\to x}(x) = \sum_{\sim x} f(X) \prod_{y\in n(f)\setminus x} m_{y\to f}(y)$, where $X = n(f)$ is the set of arguments of the function f and $\sum_{\sim x}$ represents the summations of all the arguments in X, excepting x. Variable to function messages can be interpreted as the "belief" that the variable has, of itself, given the values of all its other functions. Function to variable messages can be interpreted as the "belief" that the function has with respect to the variable's state, given the states of all the other variables in the function's argument.

When variable nodes x_t^k are observed, all messages $m_{x_k^t\to g_i^j}$ consist of trivial identity mes-

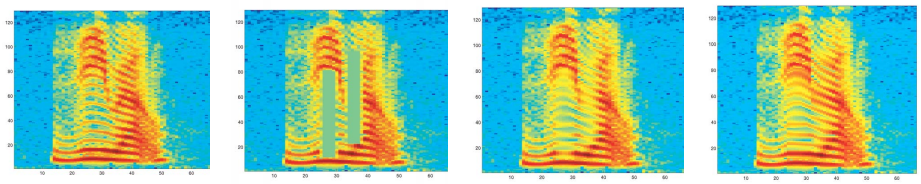

Figure 4: Missing data interpolation example a) Original, b) Incomplete, c) After 10 iterations, d) After 30.

sages. Then the only messages we need to compute are the ones that go through the T_t^k variable nodes.

We first run messages through all the bins in a given frame t for all frames. Later, we run messages through all the bins on a given frequency for all frequencies. Applying the "belief propagation" formulas on these "'chains"' results in forward/backward, upward/downward recursions similar to the ones obtained in HMMs but with weighted local likelihood that takes into account the match to the local observation as well as the "beliefs" from the neighboring chains. The use of HMM-like recursions make the inference procedure quite fast. Full derivations in [5].

3 Inferring Missing Data

If a certain region of cells in the spectrogram are missing, the corresponding nodes in the model become hidden. This is illustrated in figure 4, where regions of the spectrogram have been removed. Inference of the missing values is performed again using belief propagation. Now, for the missing values $\hat{x}_k^t$, continuous messages $m_{\hat{x}_k^t \to g_i^j}$ and $m_{g_i^j \to \hat{x}_k^t}$ have to be computed as well.

The posteriors of the hidden continuous nodes are represented using Gaussian distributions, the missing sections on figure 4 part b), are filled in with the means of their inferred posteriors, figure 4 part c), and d). The complete derivations of the continuous messages and the posterior probabilities can be found in [5].

4 Two Layer Source-Filter Transformations

Many sound sources, including voiced speech, can be successfully regarded as the convolution of a broad-band *source excitation* and a filter that 'colors' the excitation to produce speech sounds or other distinctions. Given that convolution of the source with the filter in the time domain corresponds to multiplying their spectra in the Fourier domain, or adding in the log-spectral domain. 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, as illustrated in figure 2 b), a graphical simplification, of the layered model. Where each layer with the structure depicted in figure 2 a) corresponds to a factor graph like the one on figure 3.

Variables F and H in the model are hidden, while, X can be observed or hidden. Inference in this model is more complex, but the actual form of the continuous messages is essentially the same as in the one layer case, with the addition of the potential function relating the signal x_t^k with its transformation components. The two layers are iteratively estimated as described on [5].

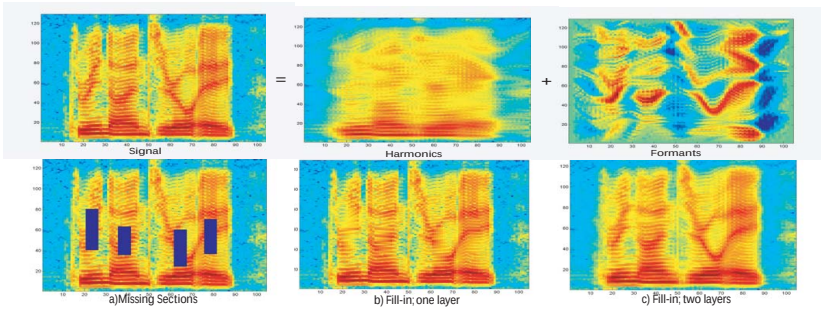

Figure 5: First Row.- Harmonics/Formants decomposition (posterior distribution means). Second Row (a) Spectrogram with deleted (missing) regions. (b) Filling in using a single-layer transformation model. (c) Results from the two-layer model.

The first row of figure 5 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 5 second Row a) shows spectrogram on the first row with deleted regions; notice that the two layers have distinctly different motions. In b) the regions have been filled via inference in a single-layer model and the formants are not captured in the reconstruction. 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 laye

Features	CLEAN	SNR20	SNR15	SNR10	SNR5
PLP12+delta	.94	2.3	4.1	7.9	12.2
PLP12+delta+(FTM1)	.98	2.3	3.4	6.8	11.1
PLP12+delta+(FTM2)	1.3	2.5	4.2	9.7	12.5

Table 1: Word Error Rate % obtained with different sets of features as a function of SNR in dB.

5 Speech Recognition Results:

The dynamics of formants are known to be powerful "information-bearing elements" in speech.

Using the formant transformation posteriors, we calculate the expected formant transformation maps, the expected transformation at each bin, which captures information about the global dynamics of the formants.

We computed two sets of transformation maps: one using formants obtained with our model, and another with formants obtained using cepstral smoothing. For the latter we only require a single layer model. We then use features derived from these maps in combination with standard features in a speech recognizer

To convert the formant transformation maps into features suitable for the recognizer, we applied mel-scale filtering and a discrete cosine transform to decorrelate and reduce the dimensionality of the final feature vectors.

We used the Aurora-2 noisy digits database for our experiments. Results at different SNR levels are shown in table 1. Features derived from formant transformation maps obtained using (one,two) layers are referred to as ("FTM1","FTM2").

Using Perceptual Linear Prediction (PLP) features combined with FTM1 features, the rec-

ognizer performance remains about the same as the standard features alone when the signal has high SNR values, but when the SNR decreases the new features improve the word error rate (WER) by as much as 19.5% relative for the 15 dB SNR ("SNR15") condition. We believe that when the signals are relatively clean, a local analysis of the energy dynamics, is sufficient to effectively disambiguate the words. However as the interference becomes larger a more global model of the energy dynamics, such as the formants transition maps, can reduce the influence of local energy variations due to the noise.

Using FTM2 features do not improve the performance of the recognizer. This may be because the layers cannot be separated when the two layers have parallel dynamics. However, independent modeling of transformation maps for both layers is important for other applications such as the source separation.

6 Unsupervised Dominant Source Separation

The separation of speech mixtures into its individual sources using a single microphone is a very hard and interesting problem. Current approaches include attempts to segregate a time-frequency representation on a bin-by-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 at such fine resolution poses a great challenge to systems attempting to do such a separation. In [6] the combinatorial search is restricted by the use of pretrained speaker models, which limits the applicability of the approach. In [7], a training session is required to choose the right parameters for a spectral clustering algorithm. Finding clusters among the set of all bins requires huge matrices that pose significant numerical problems. On the other hand, other research had shown that an intelligible separation can be done by grouping those regions of the spectrogram where a given speaker is more dominant than the others [8]. The problem is how to find those speaker-dominant regions.

6.1 Subband Matching-and-Tracking model

Prediction of frames from their context is not always possible such as when there are transitions between silence and speech, transitions between voiced and unvoiced speech or transitions between dominant speakers in the case of speech mixtures, so we need a set of states to represent these unpredictable frames explicitly. We will also need a "switch" variable that will decide when to "track" (transform) and when to "match" the observation with a state. Figure 6 a), shows the spectrogram of a mixture of two speakers, notice that the magnitude of the interference, when a change of dominant speaker ocurrs is not uniform across all the spectrum. Then we require a model that can "track" in some sections of the spectra while "matching" in others. The spectrogram is divided in R subbands, $\mathcal{K}_r$ = $[k^r_{min}, k^r_{max}]$ defines the range of frequencies encompassed by subband r. At each band r and at each time frame, discrete variables S^r_t and C^r_t are connected to all frequency bins in that subband. S^r_t is a GMM containing the means and the variances of the states for that subband. When variable C^r_t is equal to 0, the model is in "tracking mode" on subband r; a value of 1 designates "matching mode" on subband r. Inference is done again using loopy belief propagation. When $p(C^r_t) \approx 1$ a switch on the dominant speaker is detected at frame t on subband r.

7 Segmentation Results

We ran experiments on 200 artificially mixed mixtures of two speakers: 50 female-female, 50 male-male, 50 male-female and 50 same speaker with different utterances. Since we are artificially mixing the signals we can find the dominant speaker boundaries. Defining $y^i_{k,t}$ as the abs-spectra coefficient at frequency k at frame t for speaker i. We compute,

$Y^i_{r,t}$, the energy of i speaker at subband r at frame t as: $Y^i_{r,t} = \sum_{k \in \mathcal{K}_r} y^i_{k,t}$. We then find three regions for the composed spectrogram, $\mathcal{R}_1 = y^s_{k,t} \forall (k,t)$ such that $k \in \mathcal{K}_r$ and 10*log10($Y^1_{r,t}/Y^2_{r,t}$) $\geq$ 3db, $\mathcal{R}_2 = y^s_{k,t} \forall (k,t)$ such that $k \in \mathcal{K}_r$ and 10*log10($Y^1_{r,t}/Y^2_{r,t}$) $\leq$ -3db and $\mathcal{R}_0 = y^s_{k,t} \forall (k,t)$ such that $k \in \mathcal{K}_r$ and -3db $\geq$ 10*log10($Y^1_{r,t}/Y^2_{r,t}$) $\leq$ 3db; $\mathcal{R}_1$ defines those regions where speaker 1 clearly "dominates" speaker 2 for a margin of at least 3db, $\mathcal{R}_2$ is the correspondent region for speaker 2. $\mathcal{R}_0$ defines those regions where there is not a clear dominant speaker.

We then define two types of dominant speaker boundaries: hard boundaries correspond to the boundaries between regions $\mathcal{R}_1$ and $\mathcal{R}_2$ and soft boundaries that correspond to regions $\mathcal{R}_0$ found between $\mathcal{R}_1$,$\mathcal{R}_2$ regions. We require our model to detect a switch in either of the two frames bordering the hard edges and to detect a switch anywhere on the regions defined by the soft edges.

The segmentation results using the subband deformable spectrograms segmentation can be observed in the first part of table 2.

Type of Mixture	Female Female	Male Male	Female Male	Same Speaker	Female Female	Male Male	Female Male	Same Speaker
Deformable Spectrograms Segmentation					Pitch-Bic Segmentation			
Recall	96.64	97.94	97.51	96.88	68.47	66.19	71.46	61.49
Precision	62.80	62.37	61.14	69.18	39.94	38.92	42.04	36.55

Table 2: Recall/Precision results

The recall values are high without substantial differences between the different kind of mixtures. 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. The precision results are not as good. This is because transitions between voiced and unvoiced data for the same speaker are also detected as well as mismatches within the same speaker like when there are abrupt variations in the motion of both layers.

Table 2, also shows the segmentations results obtained using the a subband version of Bayesian Information Criteria (*BIC*) procedure used to detect boundaries in personal audio archives [9]. We use a subband pitch estimation as the feature for the BIC segmentation system, its parameteres were set such that for each mixture the number of resulting segments would be equivalent to the number obtained through the subband deformable spectrogram model.

Since the deformable spectrograms based segmentation has high recall values we can be pretty certain that the signal is segmented in dominant speaker regions. Even with a few false positives clustering these regions is a task several degrees simpler than clustering individual bins.

Figure 6 shows an example of the segmentation results in a composed signal.

7.1 Clustering Regions by patch similarity

We used the spectral clustering algorithm proposed in [10]. We first cluster regions within the same subband and later we cluster regions between bands. The entries for the affinities matrix A for the i and j regions is defined as:

A_{ij} = exp (-$\|Di,j\|^2/2\sigma^2$) if $i \neq j$, where $D_{i,j}$ is the summation of the n time-frequency patches taken from regions i and j with the minimum distances divided by n. A_{ii} is directly set to zero. When clustering within subbands we used $n = 3$, when clustering between bands we used $n = 10$; This similarity matrix does not depend on pitch, therefore even regions with similar pitch can be clustered if they show other sources of dissimilarity like

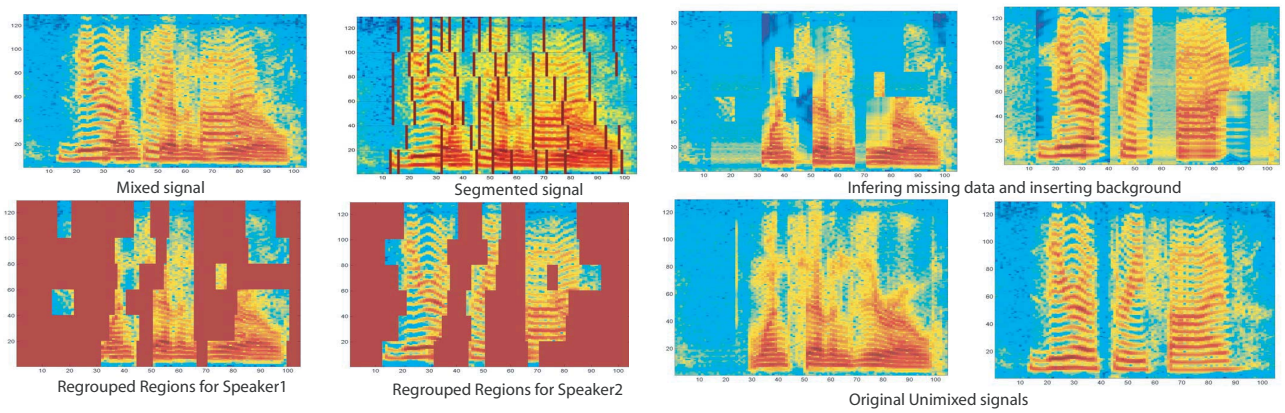

Figure 6: The first row shows the original mixed signal. Second row shows the regrouped signals by the spectral clustering algorithm. The third rows shows the reconstructions of the "'missing"' regions. Fourth row shows the original signals prior to the mixing.

prosody or style.

Figure 6 shows the results after reconstructing the missing regions.

7.2 Inference of Missing Data

Once we have cluster the segments, we can use the model to infer the masked sections. Figure 6 shows an example of the reconstruction.

Here we keep the transformation maps of both layers for the regions that the desired speaker dominates, while relearning the transformation maps for the regions that wered masked by the other speaker. The reconstruction here is not freely done as in the missing information examples shown before. Since we do have constraints of what the data can be given that we can observe the mixed signal on those regions. Moreover restrictions on the structure that the reconstructed signal may take have to be inforced to prevent the reconstruction to follow the structure of the competing speaker. Results on: http://www.ee.columbia.edu/~mjr59/def_spec.html

References

[1] N. J. M. Reyes-Gomez and D. Ellis, "Deformable spectrograms," in *AISTATS*, Barbados, 2005.

[2] B. F. F. Kschischang and H.-A. Loeliger, "Factor graphs and the sum-product algorithm," *IEEE Transactions on information theory*, vol. 47, 2001.

[3] W. F. J.S. Yedidia and Y. Weiss, "Understanding belief propagation and its generalizations," in *Exploring Artificial Intelligence in the New Millennium*, 2001.

[4] Y. Weiss and W. Freeman, "Correctness of belief propagation in gaussian graphical models of arbitrary topology," *Neural Computation*, vol. 13, pp. 2173–2200, 2001.

[5] N. J. M. J.Reyes Gomez and D.Ellis, "Deformable spectrograms," in *Microsoft Tech Report*, www.research.microsoft.com/∼manuelrg.

[6] S. Roweis, "Factorial models and refiltering for speech separation and denoising," in *Proc. EuroSpeech*, Geneva, 2003.

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

[8] M. Cooke, "10+1 perspectives on speech separation and identification in listeners and machines," Montreal, November 2004, presentation at the AFOSR/NSF Workshop on Speech Separation and Comprehension in Complex Acoustic Environments. [Online]. Available: http://labrosa.ee.columbia.edu/Montreal2004/s-overview.html#cooke

[9] D. Ellis and K. Lee, "Minimal-impact audio-based personal archives," in *SAPA2004*, Korea, 2005.

[10] M. J. Andrew Y. Ng and Y. Weiss, "On spectral clustering: Analysis and an algorithm," in *NIPS 14,, 2002*, 2002.