Towards single-channel unsupervised source separation of speech mixtures: The layered harmonics/formants separation-tracking model

Manuel Reyes-Gomez[1], *Nebojsa Jojic*[2], *Daniel P.W. Ellis*[1]

[1] LabROSA, Department of Electrical Engineering, Columbia University
[2] Microsoft Research
[mjr59,dpwe]@ee.columbia.edu jojic@microsoft.com

Abstract

Speaker models for blind source separation are typically based on HMMs consisting of vast numbers of states to capture source spectral variation, and trained on large amounts of isolated speech. Since observations can be similar between sources, inference relies on sequential constraints from the state transition matrix which are, however, quite weak. To avoid these problems, we propose a strategy of capturing local deformations of the time-frequency energy distribution. Since consecutive spectral frames are highly correlated, each frame can be accurately described as a nonuniform deformation of its predecessor. A smooth pattern of deformations is indicative of a single speaker, and the cliffs in the deformation fields may indicate a speaker switch. Further, the log-spectrum of speech can be decomposed into two additive layers, separately describing the harmonics and formant structure. We model smooth deformations as hidden transformation variables in both layers, using MRFs with overlapping subwindows as observations, assumed to be a noisy sum of the two layers. Loopy belief propagation provides for efficient inference. Without any pre-trained speech or speaker models, this approach can be used to fill in missing time-frequency observations, and the local entropy of the deformation fields indicate source boundaries for separation.

1. Introduction

In situations where two or more speakers speak simultaneously, we may wish to be able to separate the speech from the individual speakers. Conventionally, this is referred to as the *speaker-separation* or *source-separation* problem.

A popular approach to speaker separation is through the use of multiple microphones. Solutions typically require at least as many microphones as signal sources, and separation is performed using techniques such as Independent Component Analysis (ICA). This approach does not utilize any knowledge of the statistical characteristics of the signals to be separated, other than the very loose assumption that the various signals are statistically independent [1]. This approach can fail when, for instance, signals are recorded in a reverberant environment, or the degree of overlap and/or the dimensionality of the observations make the blind inference problem irresolvable.

A completely different approach uses extensive prior information about the statistical nature of speech from individual speakers, usually represented by dynamic models [2, 3, 4]. The spectral parameters of the models are composed of hundreds or even thousands of states describing all possible log-spectra of each source to an adequate level of detail. Learning such a large number of parameters from composed signals is practically impossible, so such models are learned using clean speech utterances of the corresponding speaker. The models are used to separate combined speech signals using the "refiltering" technique introduced in [2]. A significant problem with this approach is the requirement for large amounts of training data to accurately capture the complexity and variability of a particular speaker.

Here, we propose a new technique that has some resemblance to both of these approaches, exploiting very general properties of certain audio sources including speech and musical instruments by modeling the evolution of their harmonic components. 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. This layered approach draws on successful applications in computer vision that use layers to account for different sources of variability [5, 6, 7, 8]. 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 $t-1$. As a result, we do not require separate states for every possible spectral configuration, but only a limited set of initial states that can cover the full spectral variety of a source through such transformations. This factoring of the sources of variability results in a model with very few parameters that could be learned from composed data without supervision.

We will first introduce a model that captures the spectral deformation field of the speech harmonics, and show how this can be exploited to interpolate missing observations. Then, we introduce the two-layer model that separately models the deformation fields for harmonic and formant resonance components, and briefly describe a range of

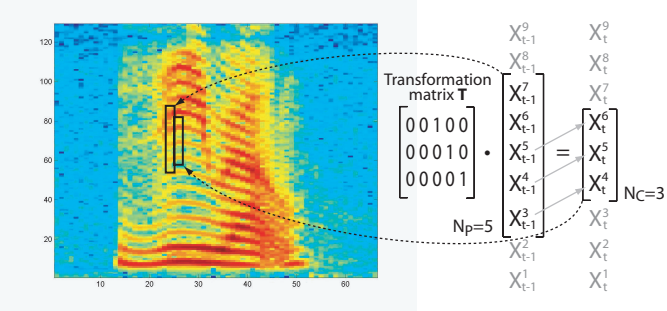

Figure 1: The $N_C = 3$ patch of time-frequency bins outlined in the spectrogram can be seen as an "upward" version of the marked $N_P = 5$ patch in the previous frame. This relationship can be described using the matrix shown.

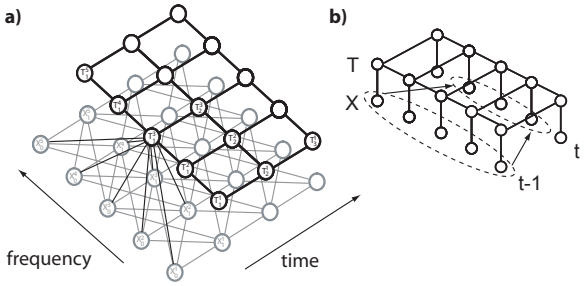

Figure 2: a) Graphical model b) Graphical simplification.

existing applications including semi-supervised source separation. Finally we describe the matching-tracking model including the initial states, and discuss its application to the unsupervised source separation task.

2. Spectral Deformation Model

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, $X_t^k = \sum_{\tau=0}^{N_F-1} w[\tau]x[\tau - t \cdot H]e^{-j2\pi\tau k/N_F}$, 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. We use 32 ms windows with 16 ms hops.

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.

$$\mathbf{X}_t^{[k-n_C,k+n_C]} \approx \mathbf{T}_t^k \cdot \mathbf{X}_{t-1}^{[k-n_P,k+n_P]} \quad (1)$$

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

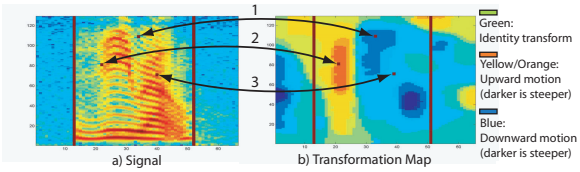

Figure 3: Example transformation map showing corresponding points on original signal.

that point on the time-frequency plane. We use overlapping patches to enforce transformation consistency [8].

Figure 1 uses 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 generative graphical model for a single layer 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. For now, we consider the continuous nodes $\mathcal{X}$ as observed, although below we will allow some of them to be hidden when analyzing the missing data scenario. 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 $\mathbf{T}$ is of the form:

$$\begin{pmatrix} \mathbf{w} & 0 & 0 \\ 0 & \mathbf{w} & 0 \\ 0 & 0 & \mathbf{w} \end{pmatrix} \quad (2)$$

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. Here, $N_C = 3$ and $\mathbf{w}$ is a row vector with length $N_W = N_P - 2$; using $\mathbf{w} = (0\ 0\ 1)$ yields the transformation matrix shown in figure 1. To ensure symmetry along the frequency axis, we constraint N_C, N_P and N_W to be odd. The complete set of $\mathbf{w}$ vectors include upward/downward shifts by whole bins as well as fractional shifts. An example set, containing each $\mathbf{w}$ vector as a row, is:

$$\begin{pmatrix} 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & .25 & .75 \\ 0 & 0 & 0 & .75 & .25 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & .25 & .75 & 0 \\ \dot{.}75 & \dot{.}25 & \dot{0} & \dot{0} & \dot{0} \\ 1 & 0 & 0 & 0 & 0 \end{pmatrix} \quad (3)$$

The length N_W of the transformation vectors defines the supporting coefficients from the previous frame $\mathbf{X}_{t-1}^{[k-n_W,k+n_W]}$ (where $n_W = (N_W - 1)/2$) that can "explain" X_t^k.

For harmonic signals in particular, we have found that a model using the above set of $\mathbf{w}$ vectors with parameters $N_W = 5$, $N_P = 9$ and $N_C = 5$ is very successful at capturing the self-similarity and dynamics of the harmonic structure. The transformations set could, of course, be learned, but in view of the results we have obtained with this predefined set, we defer the learning of the set to future work.

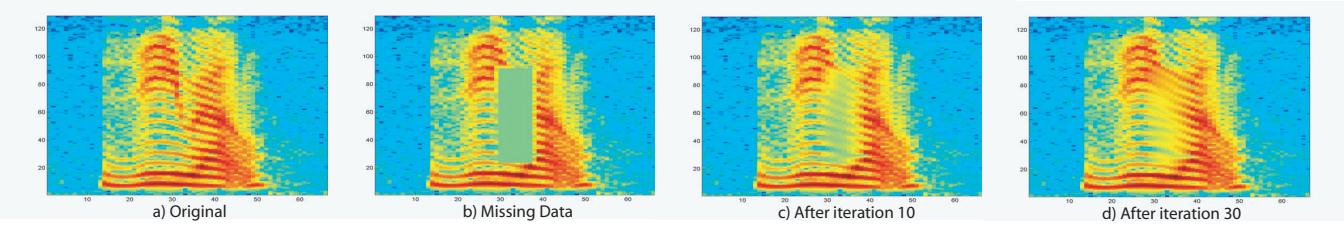

Figure 5: Missing data interpolation example a) Original, b) Incomplete, c) After 10 iterations, d) After 30.

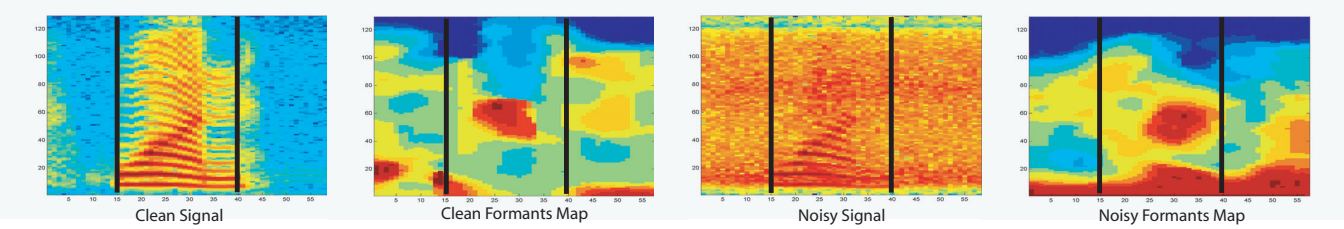

Figure 6: Formant tracking map for clean speech (left panels) and speech in noise (right panels).

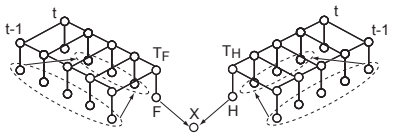

Figure 4: Graphical representation of the two-layer source-filter transformation model.

The clique "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:

$$\psi\left(\mathbf{X}_t^{[k-n_C,k+n_C]}, \mathbf{X}_{t-1}^{[k-n_P,k+n_P]}, T_t^k\right) = \mathcal{N}\left(\mathbf{X}_t^{[k-n_C,k+n_C]}; \mathbf{T}_t^k \mathbf{X}_{t-1}^{[k-n_P,k+n_P]}, \Sigma^{[k-n_C,k+n_C]}\right) \quad (4)$$

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. For the transformation cliques, the horizontal and vertical transition potentials $\psi_{hor}(T_t^k, T_{t-1}^k)$ and $\psi_{ver}(T_t^k, T_t^{k-1})$, are represented by transition matrices.

For observed nodes $\mathcal{X}$, inference consists in finding probabilities for each transformation index at each time-frequency bin. Exact inference is intractable and is approximated using Loopy Belief Propagation [9, 10]. Appendix A gives a quick review of the loopy belief message passing rules, and Appendix B presents the specific update rules for this case. The transformation map, a graphical representation of the *modes* of the transformation node posteriors across time-frequency, provides an appealing description of the harmonics' dynamics as can be observed in figure 3. In these panels, the links between three specific time-frequency bins and their corresponding transformations on the map are highlighted. Bin 1 is described by a steep downward transformation, while bin 3 also has a downward motion but is described by a less steep transformation, consistent with the dynamics visible in the spectrogram. Bin 2, in other hand, is described by a steep upwards transformation. These maps tend to be robust to noise (as shown below), making them a valuable representation in their own right.

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 5, where a rectangular region in the center has been removed and tagged as missing. Loopy belief inference now requires continuous-valued messages, complicating the procedure as explained in Appendix C. The figure shows the interpolated values inferred by the model after a few iterations. The missing-data model will be used below in the two layer source-filter model.

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*, such as the pseudo-periodic glottal flow, perhaps modeled as an impulse train, and a time-varying resonant *filter*, such as the vocal tract, that 'colors' the excitation to produce speech sounds or other distinctions. When the excitation has a spectrum consisting of well-defined harmonics, the overall spectrum is in essence samples of the filter's resonances at the frequencies of the harmonics. Convolution of the source with the filter in the time domain corresponds to multiplying their spectra in the Fourier domain, or an additive relationship in the log-spectral domain. Hence,

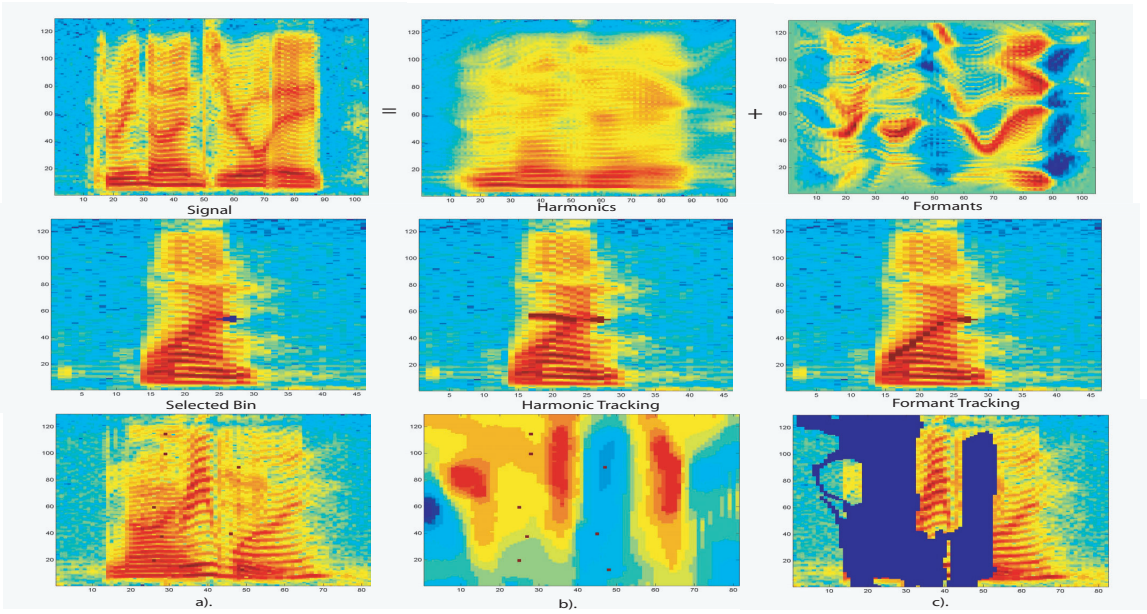

Figure 7: First row: Harmonics/Formants decomposition (posterior distribution means). Row 2: Harmonics/Formants tracking example. The transformation maps on both layers are used to track a given time-frequency bin. Row 3: Semi-supervised Two Speakers Separation. a) The user selects bins on the spectrogram that she believes correspond to one speaker. b) The system finds the corresponding bin on the transformation map. c) The system selects all bins whose transformations match the ones chosen; the remaining bins correspond to the other speaker.

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 4. Variables F and H in the model are hidden, while, as before, X can be observed or hidden. The symmetry between the two layers is broken by using different parameters in each, chosen to suit the particular dynamics of each component. We use transformations with a larger support in the formant layer ($N_W = 9$) compared to the harmonics layer ($N_W = 5$). Since all harmonics tend to move in the same direction, we enforce smoother transformation maps on the harmonics layer by using potential transition matrices with a higher self-loop probabilities. An example of the transformation map for the formant layer is shown in figure 6, which illustrates how these maps can remain relatively invariant to high levels of signal corruption; belief propagation searches for some kind of consistent dynamic structure within the signal, and since additive noise is less likely to have a well-organized structure, it is properties of the speech component that are extracted. 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 (Appendix C), with the addition of the potential function relating the signal X_t^k with its transformation components H_t^k and F_t^k) at each time-frequency bin:

$$\psi(X_t^k, H_t^k, F_t^k) = \mathcal{N}(X_t^k; H_t^k + F_t^k, \sigma^k) \qquad (5)$$

The first row of figure 7 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.

5. Applications

We have built an interactive model that implements formant and harmonics tracking, missing data interpolation, formant/harmonics decomposition, and semi-supervised source separation of two speakers.

Formants and Harmonics Tracking: Analyzing a signal with the two-layer model permits separate tracking of the harmonic and formant 'ancestors' of any given point. The user clicks on the spectrogram to select a bin, and the system reveals the harmonics and formant "history" of that bin, as illustrated in the second row of figure 7.

Semi-Supervised Source Separation: After modeling the input signal, the user clicks on time-frequency bins that appear to belong to a certain speaker. The demo then masks all neighboring bins with the same value in the transformation map; the remaining unmasked bins should belong to the other speaker. The third row of figure 7 depicts an example with the resultant mask and the "clicks" that generated it. Although far from perfect, the separation is good enough to perceive each speaker in relative isolation.

Missing Data Interpolation and Harmonics/Formants Separation: Examples of these have been shown in figures 5 and 7.

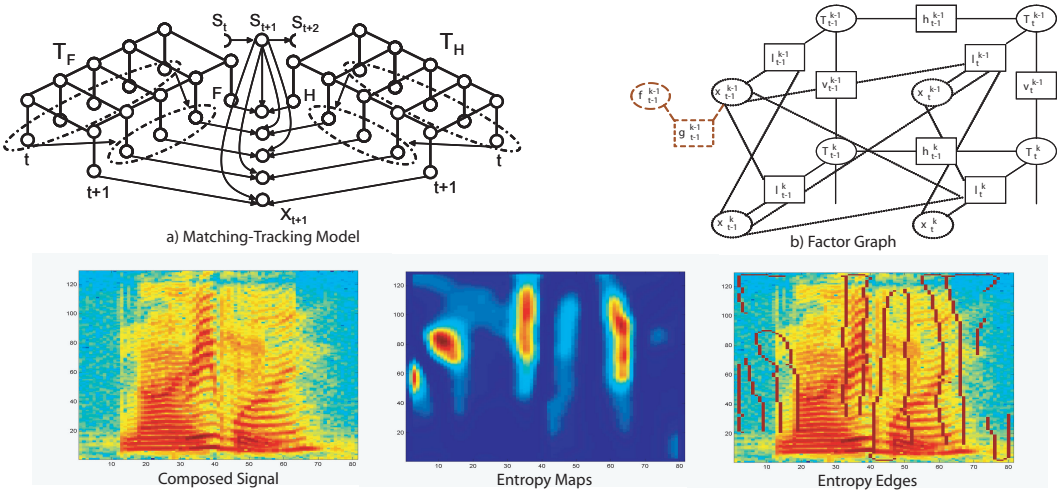

Figure 8: First row: Graphic model and factor graph of the matching-tracking model; Second row: Entropy Map and Entropy Edges

Features for Speech Recognition: The phonetic distinctions at the basis of speech recognition reflect vocal tract filtering of glottal excitation. In particular, the dynamics of formants (vocal tract resonances) are known to be powerful "information-bearing elements" in speech. We believe the formant transformation maps may be a robust discriminative feature to be use in conjunction with traditional features in speech recognition systems, particularly in noisy conditions; this is future work.

6. Matching-Tracking Model

The second row of figure 8 b) illustrates the *entropy* of the distributions inferred by the system for each transformation variable. The third pane shows 'entropy edges', boundaries of high transformation uncertainty. With some exceptions, these boundaries correspond to transitions between silence and speech, or when occlusion between speakers starts or ends. Similar edges are also found at the transitions between voiced and unvoiced speech. High entropy at these points indicates that the model does not know what to track, and cannot find a good transformation to predict the following frames. This motivates the introduction of a new variable in the model to provide a set of explicit "initial states". We refer to this as the "matching-tracking" model, represented graphically in the first row of figure 8. One state/switch variable S_t per time frame is connected as a regular HMM on top of the two tracking layers. But unlike a regular HMM, the S variables have a special "tracking state" in which the model tracks the current observation values through deformations instead of matching it to the most likely template. Hence, the model requires only a small set of states, with a few states representing the pitch of the speaker and a few others for unvoiced sounds.

The source separation problem can be addressed as follows: When multiple speakers are present, each speaker will be modeled in its own layer, further divided into harmonics and formants layers. The idea is to reduce the transformation uncertainty at the onset of occlusions by continuing the tracking of the "old" speaker in one layer at the same time as estimating the initial state of the "new" speaker in another layer – a realization of the "old-plus-new" heuristic from psychoacoustics. This is part of our current research.

7. Conclusions

We have presented a harmonic/formant separation and tracking model that effectively identifies the different factors underlying speech signals. We show that this model has a number of useful applications, many of which have already been implemented in a working real-time demo. Previous research has shown that single-microphone speech source separation is possible given detailed models of the sources, but learning those models in a unsupervised way from composed signals is practically impossible. The model we have proposed in this paper captures the detailed dynamics of speech with only a few parameters, and is a promising candidate for sound separation systems that do not rely on extensive isolated-source training data.

8. Appendices

A: Loopy Belief Propagation

The sum-product algorithm [12] can be used to approximate inference on graphical models with loops. The algorithm update rules applied to the factor graph representation of the model are:

Variable to local function:

$$m_{x \to f}(x) = \prod_{h \in n(x) \setminus f} m_{f \to x}(x) \qquad (6)$$

Local function to variable:

$$m_{f\to x}(x) = \sum_{\sim x} f(X) \prod_{y\in n(f)\setminus x} m_{y\to f}(y) \qquad (7)$$

where $X = n(f)$ is the set of arguments of the function f.

The factor graph for a section of our model is depicted in the top right corner of figure 8. The circles are the variable nodes, representing the hidden variables, and the squares represent the local function nodes, i.e. the potential functions within variable nodes [12]. The solid lines represent the model when the variables X_t^k are observed, and the dotted part is added when they are hidden.

B: Update Rules for the Spectral Deformation Model

When variables X_t^k are observed, there are only discrete messages in the algorithm. Applying the above update rules, we obtain the following forward recursion for the horizontal nodes on the grid:

$$m_{T_t^k\to h_t^k}(T_t^k) = (\sum_{T_{t-1}^k} h_t^k(T_t^k, T_{t-1}^k) m_{T_{t-1}^k\to h_{t-1}^k}(T_{t-1}^k))$$
$$l_t^k(\mathbf{X}_t^{[k-N_C:k+N_C]}, \mathbf{X}_{t-1}^{[k-n_P:k+n_P]}, T_t^k) g(T_t^{k-1}, T_t^{k+1}) \qquad (8)$$

where $g(T_t^{k-1}, T_t^{k+1})$ is the multiplication of the messages coming from the adjacent vertical nodes. A similar backward recursion can also be found. The messages for the vertical chains can be updated through analogous upward/downward recursions.

C: Loopy Belief with Continuous-Valued Messages

The message from function l_s^r of the factor graph in figure 8 to variable X_j^i has the form.

$$m_{l_s^r\to X_j^i}(X_j^i) =$$
$$\int_{\mathbf{y},\mathbf{z}} \frac{1}{C} exp^{\frac{1}{2}(\alpha X_j^i - \Gamma\mathbf{y}+\mathbf{z})' \Sigma^{-1}_{[r-n_C:r+n_C]} (\alpha \mathbf{X}_j^i - \Gamma\mathbf{y}+\mathbf{z})}$$
$$\mathcal{N}(\mathbf{y}; \mu_y, \Sigma_y)\mathcal{N}(\mathbf{z}; \mu_z, \Sigma_z) d\mathbf{y} d\mathbf{z} \qquad (9)$$

Values j and s can be either t or $t-1$, and vector $\mathbf{y}$ is formed by the values on $X_{t-1}^{[r-n_P:r+n_P]}$ other than X_j^i (or the whole vector if $j = t$). Vectors $\mathbf{z}$ and $\mathbf{X}_t^{[r-N_C:r+N_C]}$ have an analogous relationship. Vector α and matrix Γ come from the most likely (or weighted mean) of the transformation matrix used at bin X_s^r. To speed up the process, we approximate $\mathcal{N}(\mathbf{y}; \mu_y, \Sigma_y)\mathcal{N}(\mathbf{z}; \mu_z, \Sigma_z)$ by delta functions $\delta(\mathbf{y} - \mu_y)$ and $\delta(\mathbf{z} - \mu_z)$. Then the messages reduce to: $m_{l_s^r\to X_j^i}(X_j^i) = \frac{1}{C} exp^{\frac{1}{2}(\alpha X_j^i - \Gamma\mu_y+\mu_z)' \Sigma^{-1} (\alpha X_j^i - \Gamma\mu_y+\mu_z)}$.

The posterior probability of node X_t^k, $q(X_t^k)$, is equal to the multiplication of all its incoming messages. We approximate this multiplication with a Gaussian distribution, $q'(X_t^k) = \mathcal{N}(X_t^k; \mu_{X_t^k}, \psi_{X_t^k})$. Minimizing their KL divergence we find:

$$\mu_{X_t^k} = \frac{\sum_{i=1}^{N_C+N_P} \alpha_i' \Sigma_i^{-1} (\Gamma_i \mathbf{y}_i - \mathbf{z}_i)}{\sum_{i=1}^{N_C+N_P} \alpha_i' \Sigma_i^{-1} \alpha_i^{-1}} \qquad (10)$$

The values displayed by the missing data application are these mean values. The variable to local functions have the same form as in equation 10, just subtracting the numerator and denominator factor corresponding to the incoming message from the corresponding function. Since we use diagonal variances, parameters μ_y and μ_z in 9 are found by concatenating the relevant μ_X parameters. When using the two layer model, an extra message comes from node g_t^k adding extra factors in the numerator and denominator of equation 10.

9. Acknowledgments

This work was supported by Microsoft Research and by the NSF under grant no. IIS-0238301. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the NSF.

10. References

[1] A. Hyvärinen, "Survey on Independent Component Analysis", Neural Computing Surveys, 1999.

[2] S. Roweis, "One-microphone source separation", Advances in NIPS, MIT Press, 2000.

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

[4] M. Reyes-Gomez, D. Ellis, and N. Jojic, "Subband audio modeling for single-channel acoustic source separation", Proc. ICASSP, Montreal, 2004.

[5] N. Jojic and B. Frey, "Learning flexible sprites in video layers", Proc. CVPR, 2001.

[6] A. Lenit, A. Zomet, and Y. Weiss "Learning to perceive transparency from the statistics of natural scenes", Proc. NIPS, 2002.

[7] P.H.S. Torr, R. Szeliski, and P. Anandan, "An integrated Bayesian approach in layer extraction from sequences", PAMI, 2001.

[8] N. Jojic, B. Frey, and A. Kannan, "Epitomic Analysis of Appearance and Shape", Proc. ICCV, 2003.

[9] J.S. Yedidia, W.T. Freeman, and Y. Weiss, "Understanding Belief Propagation and its Generalizations", Exploring Artificial Intelligence in the New Millennium, Chapter 8.

[10] Y. Weiss and W.T. Freeman, "Correctness of Belief Propagation in Gaussian Graphical Models of Arbitrary Topology", Neural Computation, V13, No 10, pp 2173-2200, 2001.

[12] F. Kschischang, B. Frey, and H.-A. Loeliger, "Factor Graphs and the Sum-Product Algorithm", IEEE Transactions on information theory, Vol. 47 No. 2, 2001.