Detailed graphical models for source separation and missing data interpolation in audio

Manuel J. Reyes-Gomez[1], **Nebojsa Jojic**[2] **and Daniel P.W. Ellis**[1]
[1] LabROSA, Department of Electrical Engineering, Columbia University [2] Microsoft Research

Methods for blind source separation based only on general properties such as source independence encounter difficulties when the degree of overlap and/or the dimensionality of the observations make the blind inference problem unresolvable. Such situations require additional constraints on the form of the individual sources, motivating the development of models able to capture in detail the consistency and variability of a single source's sound. With single-chain HMMs this requires a very large number of states, making the models marginally practical. Here we present two approaches to factorize the variability in detailed models. First is a coupled subband model, where each source signal is broken into multiple frequency bands, and separate but coupled HMMs are built for each band requiring many fewer states per model. In order to avoid the unnatural state combinations that would arise from independent models for each band (as in the multiband speech models of [1] and [2]), we couple adjacent bands resulting in a grid-like model (fig. 1) for the full spectrum. Exact inference of such a model is intractable, but we have derived an efficient approximation based on variational methods.

We can use these subband models to separate source mixtures by combining them in a factorial model and estimating Viterbi state alignment for each component (again using an efficient variational approximation). A time-frequency mask calculated as the elementwise maximum of the state means at each frame can extract energy almost completely dominated by one of the sources, as in [3]. Our approach outperforms both in degree of separation and computational time its full spectra (traditional HMMs) counterpart. Figure 3 (a) shows the spectrogram of a mixture of two sources; panel (b) shows a separation mask, and panel (c) shows the resulting masked elements corresponding to one source, with many cells missing. Although the human perceptual system is quite tolerant of these deletions, recognizers trained on clean signals will be seriously disturbed by such distortion, motivating our second model for efficient interpolation in time-frequency representations.

This model represents a signal with a limited number of states plus a transformation mechanism that models each frame of a spectrogram as a combination of a representative state and a transformation of the previous frame. The transformation is done by using a set of matrices T_t^k that relate a vector of $N+1$ time-frequency coefficients centered around the k^{th} bin at frame t, $X_t^{[k-N/2,k+N/2]}$ with a vector of $M+1$ coefficients centered around the k^{th} bin at frame $t-1$, $X_{t-1}^{[k-M/2,k+M/2]}$. Although this model could also be used for source separation as above, we use it here for interpolating missing spectrogram values from observed ones. The resulting model is a two layer Markov random field with an upper layer representing the transformation nodes and a lower layer representing the energies in time-frequency cells (fig. 2). The model is intractable but inference is done using loopy belief propagation. We applied our transformation model to interpolate the missing bins in fig. 3 (c); panel (d) shows the result after 15 iterations, and panel (e) after 30 iterations. Our interactive demo visualizes the interpolation process in real-time.

This work was supported by Microsoft Research and by the NSF under grant NSF-IIS-0238301.

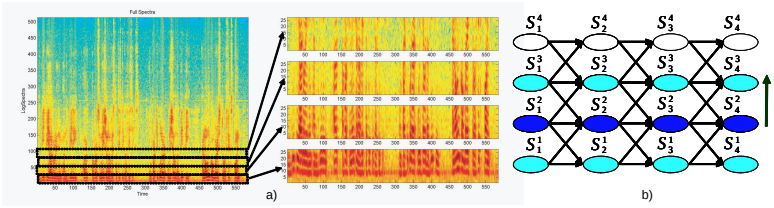

Figure 1: (a) Spectrogram partitioning and (b) multiband model.

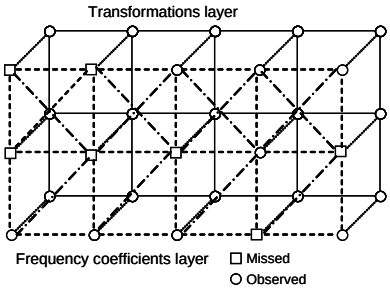

Figure 2: Interpolation Model

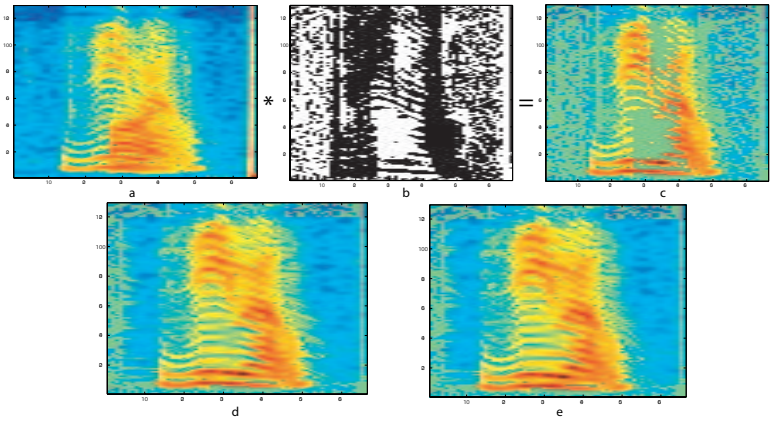

Figure 3: Spectral masking and interpolation of missing data

References

[1] H. Bourlard and S. Dupont. Subband-based speech recognition. In *Proc. ICASSP*, 1997.

[2] N. Mirghafori. *A Multiband Approach to Atomatic Speech Recognition*. PhD thesis, Dept. of EECS, UC Berkeley, 1998.

[3] S. Roweis. Factorial models and refiltering for speech separation and denoising. In *Proc. EuroSpeech*, Geneva, 2003.