Factorial Models and Refiltering for Speech Separation and Denoising

Sam T. Roweis
Department of Computer Science, University of Toronto, roweis@cs.toronto.edu

Abstract

This paper proposes the combination of several ideas, some old and some new, from machine learning and speech processing. We review the max approximation to log spectrograms of mixtures, show why this motivates a "refiltering" approach to separation and denoising, and then describe how the process of inference in factorial probabilistic models performs a computation useful for deriving the masking signals needed in refiltering. A particularly simple model, factorial-max vector quantization (MAXVQ), is introduced along with a branch-and-bound technique for efficient exact inference and applied to both denoising and monaural separation. Our approach represents a return to the ideas of Ephraim, Varga and Moore but applied to auditory scene analysis rather than to speech recognition.

1. Sparsity & Redundancy in Spectrograms

1.1. The LOG-MAX Approximation

When two clean speech signals are mixed additively in the time domain, what is the relationship between the individual log spectrograms of the sources and the log spectrogram of the mixture? Unless the sources are highly dependent (synchronized), the spectrogram of the mixture is almost exactly the *maximum* of the individual spectrograms, with the maximum operating over small time-frequency regions (fig. 2). This amazing fact, first noted by Roger Moore in 1983, comes from the fact that unless e_1 and e_2 are both large and almost equal, $\log(e_1 + e_2) \approx \max(\log e_1, \log e_2)$ (fig. 1a). The *sparse* nature of the speech code across time and frequency is the key to the practical usefulness of this approximation: most narrow frequency bands carry substantial energy only a small fraction of the time and thus it is rare that two independent sources inject large amounts of energy into the same subband at the same time. (Figure 1b shows a plot of the relative energy of two simultaneous speakers in a narrow subband; most of the time at least one of the two sources shows negligible power.)

1.2. Masking and Refiltering

Fortunately, the speech code is also *redundant* across time-frequency. Different frequency bands carry, to a certain extent, independent information and so if information in some bands is suppressed or masked, even for significant durations, other bands can fill in. (A similar effect occurs over time: if brief sections of the signal are obscured, even across all bands, the speech is still intelligible; while also useful, we do not exploit this here.) This is partly why humans perform so well on many monaural speech separation and denoising tasks. When we solve the cocktail party problem or recognize degraded speech, we are doing structural analysis, or a kind of "perceptual grouping" on the incoming sound. There is substantial evidence that the appropriate subparts of an audio signal for use in grouping may be narrow frequency bands over short times. To generate these parts computationally, we can perform multiband analysis – break the original speech signal $y(t)$ into many subband signals $b_i(t)$ each filtered to contain only energy from a small portion of the spectrum.

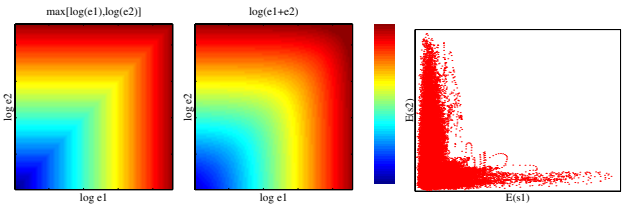

Figure 1: *(left) Relationship between log of sum and max of logs; each function's value is shown using the color scale indicated in the middle. Significant differences occur only when $e_1 \approx e_2$ and both are large. (right) Relative energy of two sources in a single subband; few points appear on the diagonal.*

The basic idea of *refiltering* [1, 2] is to separate or denoise sources by selectively reweighting the $b_i(t)$ obtained from multiband analysis of the original mixed or corrupted recording. Crucially, unlike in unmixing algorithms, the reweighting is not constant over time; it is controlled by a set of *masking signals*. Given a set of masking signals, denoted $\alpha_i(t)$, a clean source $s(t)$ can be recovered by modulating the corresponding subband signals from the original input and summing:

$$\underbrace{s(t)}_{\text{estimated source}} = \overbrace{\alpha_1(t)}^{\text{mask 1}}\underbrace{b_1(t)}_{\text{sub-band 1}} + \ldots + \overbrace{\alpha_K(t)}^{\text{mask K}}\underbrace{b_K(t)}_{\text{sub-band K}} \quad (1)$$

The $\alpha_i(t)$ are gain knobs on each subband that we can twist over time to bring bands in and out of the source as needed. This performs masking on the original spectrogram. (An equivalent operation can be performed in the frequency domain by making a conventional spectrogram of the original signal $y(t)$ and modulating the magnitude of each short time DFT while preserving its phase: $s^w(\tau) = \mathcal{F}^{-1}\{\alpha^w \| \mathcal{F}\{y^w(\tau)\} \| \angle \mathcal{F}\{y^w(\tau)\}\}$ where $s^w(\tau)$ and $y^w(\tau)$ are the w^{th} windows (blocks) of the recovered and original signals, α_i^w is the masking signal for subband i in window w, and $\mathcal{F}[\cdot]$ is the DFT.) This approach, illustrated in figure 3, forms the basis of many CASA systems[2, 3]. The basic intuition is to "gate in" subbands deemed to have high signal to noise and to be part of the source we are trying to separate and "gate out" subbands when they are deemed to be noisy or part of another source.

For any specific choice of masking signals $\alpha_i(t)$, refiltering attempts to isolate a *single clean source* from the input signal and suppress all other sources and background noises. Different sources can be isolated by choosing a different set of masking signals. Although, in general, masking signals are real-valued, positive quantities that may take on values greater than unity, in practice the (strong) simplifying assumption that $\alpha_i(t)$ are binary and constant over a timescale τ of roughly 30ms can be made. This assumption is physically unrealistic, because the energy in each small region of time-frequency never comes entirely from a single source. However, for small numbers of sources,

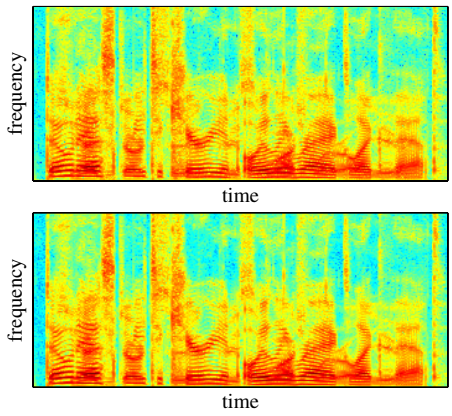

Figure 2: *(left) Log spectrogram of a mixture of two sources. (right) Elementwise maximum (within each time-frequency bin) of log spectrograms of original sources.*

this approximation works quite well[1], in part because of the effect illustrated in figure 1b. Refiltering can also be thought of as a highly nonstationary Wiener filter in which both the signal and noise spectra are re-estimated at a rate $1/\tau$; the binary assumption is equivalent to assuming that over a timescale τ the signal and noise spectra are nonoverlapping. It is a fortunate empirical fact that refiltering, even with piecewise constant binary masking signals, *can* cleanly separate sources from a single mixed recording.

2. Multiband grouping as a statistical pattern recognition problem

Since refiltering for separation and denoising is indeed possible if the masking signals are well chosen, the essential problem is: how can the $\alpha_i(t)$ be computed automatically from a single input recording? The goal is to group together regions of the spectrogram that have high signal-to-noise and belong to the same auditory object. Fortunately, natural auditory signals—especially speech—exhibit a lot of regularity in the way energy is distributed across the time-frequency plane. Grouping cues based on these regularities have been studied by psychophysicists and are hand built into many CASA systems. The approach advocated in this paper is to use statistical learning methods to discover these regularities from a large amount of speech data and then to use the learned models to compute the masking signals for new signals in order to perform refiltering.

2.1. MAXVQ: Factorial-Max Vector Quantization

It is often advantageous to model complicated sensory observations using a number of separate but interacting causes. One general way to pursue this modeling idea is to have a fixed number M of vector quantizers (or mixture models), each of which proposes an output, and then have some way of combining the output proposals into a final observation. Motivated by the observation above regarding the MAX approximation to log spectrograms of mixtures, we propose such a model, called Factorial-Max Vector Quantization (MAXVQ), which uses the MAX operation to combine outputs from the various causes. The model has a bank of M independent vector quantizers, each of which stochastically selects a prototype with which to model the observation vector. The final output vector is a noisy composite of the set of proposed prototypes, obtained by taking the *elementwise maximum* of the set and adding nonnegative noise.

The MAXVQ model is useful in situations where there are multiple "objects", "sources" or "causes" in the world but there is some kind of occlusion or sparseness governing how the sources interact to produce observations. For example, as noted above, in clean speech recordings, the log spectrogram of a mixture of speakers is almost exactly the elementwise maximum of the log spectrograms of the individual speakers. For noisy mixtures of speech signals, each clean speaker and each noise source can be thought of as an independent cause contributing to the observed signal. We will use the short-time log power in linearly spaced narrow frequency bands as our vectors when analyzing speech with this model.

Formally, MAXVQ is a latent variable probabilistic model for D-dimensional data vectors $\boldsymbol{x}$. The model consists of M vector quantizers, each with K_m codebook vectors $\boldsymbol{v}_m^k$. Latent variables $z_m \in \{1 \dots K_m\}$ control which codebook vector each vector quantizer selects. Given these selections, the final output $\boldsymbol{x}$ is generated as a noisy version of the elementwise maximum of the selected codewords. If we assume that the each vector quantizer chooses its codebook entries independently with fixed rates π_m^k, then the model can be written as:

$$\begin{aligned} p(z_m = k|\pi) &= \pi_m^k \qquad m \in \{1 \dots M\}, k \in \{1 \dots K_m\} \\ p(\boldsymbol{z}) &= \prod_m p(z_m), \quad \boldsymbol{z} = (z_1, \dots, z_M) \\ a_d &= \arg\max_m (v_{md}^{z_m}) \\ p(x_d|a_d, \boldsymbol{v}, \Sigma) &= \mathcal{N}^+(x_d|v_{ad}^{z_a}, \Sigma_{ad}) \\ p(\boldsymbol{x}|\boldsymbol{v}, \Sigma, \pi) &= \sum_{\boldsymbol{z}} p(\boldsymbol{z}|\pi) p(\boldsymbol{x}|\boldsymbol{z}, \boldsymbol{v}, \Sigma) \end{aligned}$$

were z_m are latent variables, $\boldsymbol{v}_m^k$ are the codebook vectors, Σ_{md} are noise variances (shared across k), and M, K_m are structural size parameters chosen to control complexity. The distribution $\mathcal{N}^+$ is the positive side of a Gaussian.

MAXVQ can be thought of as an exponentially large mixture of positive Gaussians having $(\prod_m K_m)$ components, with the mean of each component constrained to be the elementwise max of some underlying parameters $\boldsymbol{v}$. This technique, of representing an exponentially large codebook using a factorial expansion of a small number of underlying parameters has been very influential and successful in recent machine learning algorithms (e.g. transformed mixtures, multiple-cause VQ).

This model can also be extended through time to generate a Factorial-Max Hidden Markov Model [1, 4]. There are some additional complexities, and the details of the heuristics used for inference are slightly different but in our experience the frame-independent MAXVQ model performs almost as well and so for simplicity, we will not discuss the full HMM model.

2.2. Parameter Estimation from Isolated Sources

Given some isolated (clean) recordings of individual speech or noise sources, we can estimate the codebook means v_d^k, noise variances Σ_d and the selection probabilities π^k associated with the source's model by training a mixture density or a vector quantizer on the columns of a short-time narrowband log spectrogram. Some care must be taken in training to properly obey the nonnegativity assumption on the noise and to avoid too many codebook entries (mixture components) representing low energy (silent) segments (which are numerous in the data).

2.3. Inference for Refiltering

The key idea in this paper is that the process of inference (i.e. deducing the values of the hidden variables given the parameters

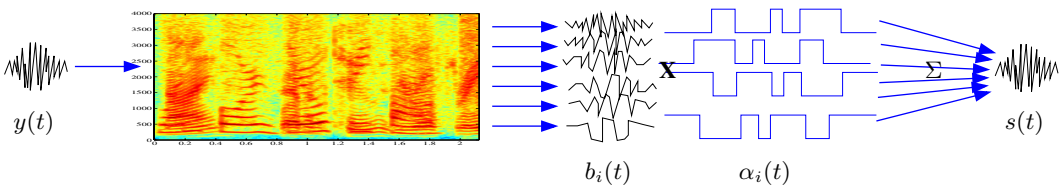

Figure 3: *Refiltering for separation and denoising. Multiband analysis of the original signal $y(t)$ gives sub-band signals $b_i(t)$ which are modulated by masking signals $\alpha_i(t)$ (binary or real valued between 0 and 1) and recombined to give an estimated source $s(t)$.*

and observations) in a MAXVQ model performs a computation which is extremely useful for computing the masking signals required to perform refiltering for denoising or separation. Because the number of possible joint settings of the hidden selection variables $\boldsymbol{z}$ is exponentially large, we are usually only interested in finding the single most likely (MAP) setting of $\boldsymbol{z}$ given $\boldsymbol{x}$ or the N-best settings. (For unsupervised learning and likelihood computations we may also be interested in efficiently summing over all possible joint settings of $\boldsymbol{z}$ to compute the marginal likelihood of a given observation $\boldsymbol{x}$.) Computing these Viterbi settings (or the sum) is intractable either by direct summation or by naive dynamic programming because of the factorial nature of the model. We must resort to branch-and-bound algorithms for efficient decoding or else approximations (e.g. variational methods) to estimate likely settings of $\boldsymbol{z}$.

Once we have computed the MAP (or approximate) setting of $\boldsymbol{z}$, we can use this to estimate the refiltering masking signals as follows: for each (overlapping) frame of the input spectrogram, set the masking signal to unity for every frequency at which the output proposed by the model corresponding to the source to be recovered is the maximum proposal over all models. Other frequencies have their masks set to zero. Actual refiltering is then performed by retaining the phase from the spectrogram of the original (noisy/mixed) recording, applying the (binary) masking signals to the log magnitude of each frequency, and reconstituting the clean signal using overlap-and-add reconstruction. The windowing function used to compute the original spectrogram must be known (or estimated) in order to remove its effect properly during refiltering.

2.4. Branch-and-Bound for Efficient Inference

As discussed above, naive computation of the MAP joint settings of the hidden selection variables in MAXVQ is exponentially expensive. Fortunately, there is a clever branch and bound trick which can be used, based on the following observation: if $z_m = k$, we can *upper bound* the log likelihood we can achieve on a data case $\boldsymbol{x}$, no matter what values the other $z_{m' \neq m}$ take on. The bound $\log p(\boldsymbol{x}|z_m = k) \leq B_{mk}$ is constructed as follows (using constant Σ for simplicity):

$$B_{mk} = -\frac{1}{2}\sum_d [x_d - v_{md}^k]_+^2 - \frac{D}{2}\log|\Sigma| - \log \pi_m^k \qquad (2)$$

where $[r]_+$ takes the max of zero and r. The intuition is that either $\boldsymbol{v}_m^k$ is *greater than* $\boldsymbol{x}$ along a certain dimension d of the output, in which case the error will be at least $(x_d - v_{md}^k)^2$ or else it is *less than* $\boldsymbol{x}$ along dimension d in which case the error on that dimension could potentially be zero.

This bound can be used to quickly search for the MAP setting of $\boldsymbol{z}$ given $\boldsymbol{x}$ as follows. For each $m \in \{1 \ldots M\}$ and each $k \in \{1 \ldots K_m\}$, compute the bound B_{mk}. Initially set the guess of the best configuration to the settings with the best bounds: $z_m^* = \arg\min_k B_{mk}$ and compute the true likelihood achieved by that guess: $\ell^* = \log p(\boldsymbol{x}|\boldsymbol{z}^*)$. Now, for each $m \in \{1 \ldots M\}$, we can eliminate all k for which $B_{mk} < \ell^*$. In other words, we can definitively say that certain codebook choices are impossible for certain models, independent of what other models choose because they would incur a minimum error worse that what has already been achieved. Now, for each m, and for all possible settings of k that remain for that m, systematically evaluate $\log p(\boldsymbol{x}|\boldsymbol{z})$ and if it is less than ℓ^*, eliminate the setting. If the likelihood is greater than ℓ^*, we accept it as the new best setting and reset $\boldsymbol{z}^*$ and ℓ^*; we also re-eliminate all settings of k that are now invalid because of this improved bound, and repeat until all settings have been either pruned or checked explicitly. This method is guaranteed to find the *exact* MAP setting, but it comes with no guarantees about its time complexity. In practice, however, we have found it to prune very aggressively and almost always find the MAP configuration in reasonable time.

3. Experiments

As an illustration of the methods presented above, we performed simple denoising and separation experiments using TIMIT prompts read by a single speaker and noise (babble) from the NOISEX database. Narrowband spectrograms we constructed from isolated, clean training examples of the speaker and noise. (Signals were downsampled to 12.5kHz, frames of length 512 were used with Hanning windows and frame shifts of 64 samples, resulting in one 257-vector of log energies each 5ms representing the signal over the last 40ms.) A simple vector-quantization codebook with 512 codewords was trained on the speech and one with 32 codewords was trained on the noise. Approximately 5 minutes of speech (with low energy frames eliminated) and 100 seconds of noise were used for training. A modified k-means algorithm which includes split-and-merge heuristics for finding good local optima was used. (We have also experimented with training "scaled" vector quantizers which cluster onto rays in the input space rather than on points, although this technique was not used in the results below.) The trained models were then used to perform MAXVQ inference on previously unseen test data, using the branch-and-bound technique. Based on this inference, refiltering was performed as described above to recover clean/isolated sources. In the denoising experiment, a 6 second speech segment was linearly mixed with 6 seconds of babble noise at 0dB SNR (equal power). Figure 4 shows the results of denoising with MAXVQ and also with a simple spectral subtraction trained on the same isolated noise sample as used for the VQ model. For the separation experiment, two different utterances, spoken by the same speaker, were mixed at equal power and the speech model was used (symmetrically) to perform MAXVQ inference. The results of this monaural separation are shown in figure 5. Of course, these results do not represent competitive performance on either denoising or separation tasks; they are merely a proof of concept that the marriage of refiltering and inference in factorial models can be used for powerful speech processing tasks.

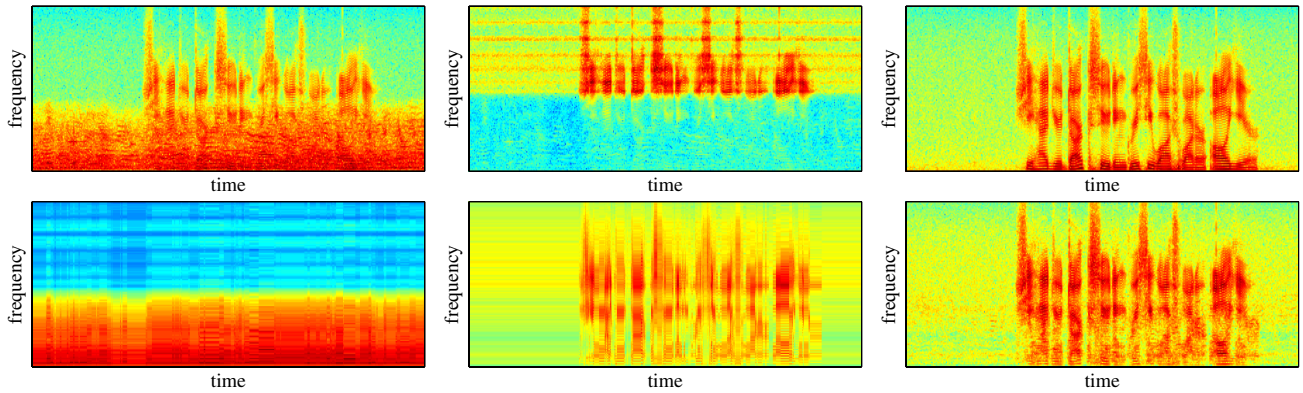

Figure 4: *Denoising using MAXVQ. Clockwise from top left: noisy input, spectral subtraction estimate (trained on isolated noise), original clean source, MAXVQ estimate after exact branch-and-bound inference and refiltering (trained on isolated speech and noise), proposed codebook output sequence from speech model, proposed codebook output sequence from noise model.*

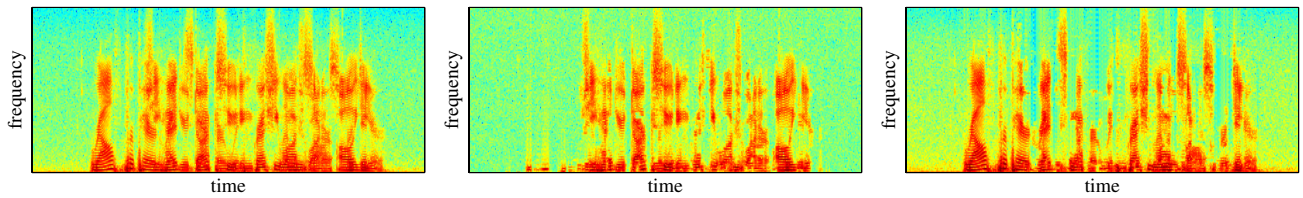

Figure 5: *Monaural separation using MAXVQ. (left) mixed input of two different utterances spoken by the same speaker. (middle, right) MAXVQ estimates of original utterances after exact branch-and-bound inference and refiltering (trained on isolated speech).*

4. Discussion, Related & Future Work

In this paper, we have argued that the *refiltering* approach to separation and denoising can be successfully achieved by using the inference step in a factorial model to provide the masking signals. Varga and Moore [4] proposed a factorial model for spectrograms (focusing on the factorial nature and using the log-max approximation) as did Gales and Young [5] (focusing on the combination operation) but these models were used for speech recognition in the presence of noise only, and not for refiltering to do separation and denoising. In a series of papers, Green et.al. [2] have studied masking (refiltering) for denoising, but do not employ factorial model inference as an engine for finding masking signals. There have also been several approaches to monaural separation and denoising that operate mainly in the time domain, without using refiltering or factorial models. Cauwenberghs [6] investigated separation based on maximizing periodic coherence; Wan and Nelson [7] use nonlinear autoregressive networks and extended Kalman filtering.

Our work here and previously[1] is closest in spirit to that of Ephraim et.al. [8] who model speech using a HMM and noise using an AR model and then attempt to approximately infer the clean speech by alternating between Wiener filtering to find the noise and Viterbi decoding in the HMM. Logan and Moreno [9] also investigated the use of factorial HMMs for modeling speech and found standard HMMs to be just as good, but they did not compose their model using the MAX of two underlying models; rather they learned separate parameters for each combination of states. Reyes et.al. [10] investigated factorial HMMs for separation but using multi-channel inputs. The main challenge for future work is to develop techniques for learning from only mixed/noisy data, without requiring clean, isolated examples of individual sources or noises at training time.

5. References

[1] S. Roweis (2002) *One Mic. Source Separation*, NIPS13.

[2] P. Green, J. Barker, M.P. Cooke & L. Josifovski (2001) *Handling Missing and Unreliable Information in Speech Recognition*, AISTATS'01.

[3] G.J. Brown & M.P. Cooke (1994) *Computational auditory scene analysis.* Computer Speech and Language, v8

[4] A.P. Varga & R.K. Moore (1990) *Hidden Markov model decomposition of speech and noise*, ICASSP'90.

[5] M.J.F. Gales & S.J. Young (1996) *Robust continuous speech recognition using parallel model combination*, IEEE Trans. Speech & Audio Processing v.4.

[6] G. Cauwenberghs (1999) *Monaural separation of independent acoustical components*, ISCAS'99.

[7] E.A. Wan & A.T. Nelson (1998) *Removal of noise from speech using the dual EKF algorithm*, ICASSP'98.

[8] Y. Ephraim, D. Malah & B.H. Juang (1989) *On the Application of Hidden Markov Models for Enhancing Noisy Speech*, IEEE Trans. Acoust., Speech and Sig. Proc., v. 37

[9] B.T. Logan & P.J. Moreno (1998) *Factorial HMMs for Acoustic Modeling* ICASSP'98.

[10] M. Reyes, B. Raj & D. Ellis (2003) *Multi-channel Source Separation by Factorial HMMs*, ICASSP'03.

Acknowledgments

STR is funded in part by NSERC Canada. Thanks to Lawrence Saul and Chris Harvey for helpful discussions and comments and to Mazin Rahim and Rob Shapire for organizing the special session in which this paper appeared.