Hierarchical Variational Loopy Belief Propagation for Multi-talker Speech Recognition

Steven J. Rennie, John R. Hershey, Peder A. Olsen
IBM T.J. Watson Research Center
Yorktown Heights, N.Y., U.S.A.
(sjrennie, jrhershe, pederao)@us.ibm.com

***Abstract*—We present a new method for multi-talker speech recognition using a single-channel that combines loopy belief propagation and variational inference methods to control the complexity of inference. The method models each source using an HMM with a hierarchical set of acoustic states, and uses the max model to approximate how the sources interact to generate mixed data. Inference involves inferring a set of probabilistic time-frequency masks to separate the speakers. By conditioning these masks on the hierarchical acoustic states of the speakers, the fidelity and complexity of acoustic inference can be precisely controlled. Acoustic inference using the algorithm scales linearly with the number of probabilistic time-frequency masks, and temporal inference scales linearly with LM size. Results on the monaural speech separation task (SSC) data demonstrate that the presented Hierarchical Variational Max-Sum Product Algorithm (HVMSP) outperforms VMSP by over 2% absolute using 4 times fewer probablistic masks. HVMSP furthermore performs on-par with the MSP algorithm, which utilizes exact conditional marginal likelihoods, using 256 times less time-frequency masks.**

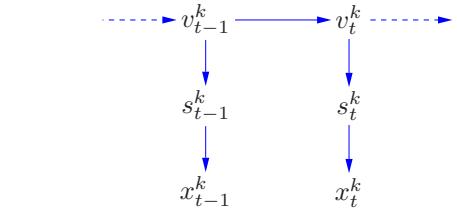
Index Terms: Speech separation, variational inference, loopy belief propagation, factorial hidden Markov models, Iroquois, Max model.

I. Introduction

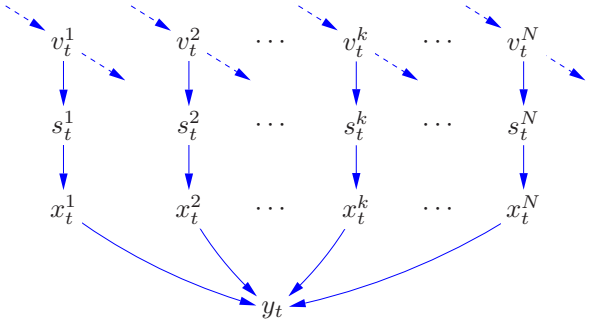
Most existing automatic speech recognition (ASR) research has focused on single-talker recognition. In many scenarios, however, the acoustic background consists of multiple sources of acoustic interference, including speech from other talkers. Such input is easily parsed by the human auditory system, but is highly detrimental to conventional ASR systems.

In [1], a system for separating and recognizing multiple speakers using a single channel is presented. The system won the recently introduced monaural speech separation challenge [2], and even outperformed human listening results on the task. The performance of this system hinges on the separation component of the system, which models speakers using a factorial hidden Markov model (HMM) (see Figure 1). In [3] several approximations are used to make inference in this model tractable, but inference still scales exponentially with the number of sources. When the vocabulary and/or acoustic models of the speakers are large, or there are more than two talkers, more efficient methods are necessary.

In [4] a loopy belief propagation algorithm (MSP) that makes inference scale linearly with language model size was presented. This algorithm, however, still scales exponentially with the number of sources as a function of acoustic model size. This shortcoming was addressed in [5], where a new variational framework for approximate inference using the max interaction model was introduced. Inference in this framework involves computing a set of probabilistic time-frequency masks to separate the sources and approximate their likelihoods. The framework was used to derive the Variational MSP (VMSP) algorithm, which scales linearly with language and acoustic model size, and is currently the best-performing algorithm on the SSC task that scales as such.



(a) Single Speaker Model



(b) Multi-speaker Model

Fig. 1. a) Generative model for the features, x^k, of single speaker: an HMM with grammar states, v^k, sharing common acoustic states, s^k. b) Generative models for hidden features $x_t^1, ..., x_t^N$ of N speakers combined to explain mixed observation features y_t using an interaction model. Dashed arrows indicate the continuation of each Markov chain over time.

VMSP achieves linearity by iteratively conditioning the variational probabilistic time-frequency masks on the acoustic states of a single source, which forces the masks to be shared across all combinations of acoustic states of the other sources during each iteration. In this work, we generalize this variational framework to hierarchical acoustic models. By conditioning the probabilistic masks on hierarchical acoustic states of multiple sources, the resolution of these probabilistic masks and the complexity of acoustic inference can be

precisely controlled. Inference using HVMSP scales linearly with the number of probabilistic masks, and linearly with LM size. The presented Hierarchical VMSP (HVMSP) algorithm generalizes the VMSP algorithm, and reduces to MSP, which utilizes exact conditional marginal acoustic likelihoods, when the probabilistic masks are conditioned on all combinations of the full resolution acoustic states of the sources.

II. Speech Models

We use the model detailed in [3], and depicted in Figure 1(a). The model consists of an *acoustic model* and a *temporal dynamics model* for each speaker. These are combined using an *interaction model*, which describes how the source features generate the observed mixed features (Figure 1(b)).

Acoustic Model: The log-power spectrum $\mathbf{x}^k$ of source k given the discrete acoustic state s^k is modeled as a diagonal covariance Gaussian, $p(\mathbf{x}^k|s^k) = \prod_f \mathcal{N}(x_f^k; \mu_{f,s^k}, \sigma^2_{f,s^k})$, for frequency f. Hereafter we drop the f when it is clear that we are referring to a single frequency. In this paper we use $D_s = 256$ gaussians for each speaker k unless otherwise noted.

Approximate acoustic inference using the max interaction model is done in this paper using hierarchical representation of this model, which decomposes $p(s^k)$ into a hierarchy of acoustic states. This is done by recursively clustering the acoustic model down and storing the probabilistic mappings between the acoustic states at different model resolutions. This process is described in further detail in section VI.

Grammars: The task grammar has $D_v = 506$ states and is represented by a sparse matrix of state transition probabilities, $p(v_t^k|v_{t-1}^k)$. The association between the grammar state v^k and the acoustic state s^k is captured by the transition probability $p(s^k|v^k)$, for speaker k. These are learned from clean training data using inferred acoustic and grammar state sequences.

III. Interaction Model

Here we consider the problem of separating a set of N source signals from a single, additive mixture

$$y(t) = \sum_k x_k(t). \tag{1}$$

The Fourier transform of $y(t)$ is $Y = \sum_k X^k$, and has power spectrum

$$|Y|^2 = \sum_k |X^k|^2 + \sum_{j\neq k} |X^j||X^k| \cos(\theta_j - \theta_k), \tag{2}$$

where θ_k is the phase of source X^k. In the log spectral domain:

$$y = \log\left(\sum_k \exp(x^k) + \sum_{j\neq k} \exp(\frac{x^j + x^k}{2}) \cos(\theta_j - \theta_k)\right),$$

where $x^k \triangleq \log|X^k|^2$ and $y \triangleq \log|Y|^2$. Assuming uniformly distributed source phases, $E\big(|Y|^2\big|\{X^k\}\big) = \sum_k |X^k|^2$. When one source dominates the others in a given frequency band, the phase terms in (2) are negligible. This motivates the *log-sum* approximation, $y \approx \log\sum_k \exp(x^k)$, which can be written in the following form:

$$y = \max_k x^k + \log\left(1 + \sum_i \exp(x^i - \max_k x^k)\right),$$

and historically motivated the *max* approximation to y,

$$y \approx \max_k x^k. \tag{3}$$

The max approximation was first used in [6] for noise adaptation. In [7], the max approximation was used to compute joint state likelihoods of speech and noise and find their optimal state sequence under a factorial hidden Markov model (HMM) of the sources. Recently [8] showed that in fact $E_\theta(\mathrm{y}|x^a, x^b) = \max(x^a, x^b)$ for uniformly distributed phase. The result holds for more than two signals when $\sum_{j\neq k}|X^j| \leq |X^k|$ for some k. In general the max is not the expected value of y for $N > 2$, but can still be used as an approximate likelihood function:

$$p(y|\{x^i\}) = \delta(y - \max_k x^k), \tag{4}$$

where $\delta()$ is the Dirac delta function, and $\{x^i\}$ is the set of all speaker feature variables.

IV. Exact Inference in the Max Model

In this section we review how the joint acoustic state likelihoods of the speakers, $p(y_f|\{s^i\})$, and the conditional expectations of the features of speaker k, $E(x_f^k|\{s^i\})$, are computed at each frequency band f for speaker models with conditionally independent acoustic features. These quantities form the basis of any exact inference strategy.

Let $p_{\mathsf{x}_f^k}(y_f|s^k) \triangleq p(\mathsf{x}_f^k = y_f|s^k)$ for random variable x_f^k, and $\Phi_{\mathsf{x}_f^k}(y_f|s^k) \triangleq p(\mathsf{x}_f^k \leq y_f|s^k) = \int_{-\infty}^{y_f} p(x_f^k|s^k)\mathrm{d}x_f^k$ be the cumulative distribution of x_f^k evaluated at y_f. Further let d_f be a random variable that is equal to k when source k dominates the mixture in frequency band f. The likelihood of state combination $\{s^i\}$ given y is:

$$\begin{aligned} p(y_f|\{s^i\}) &= \sum_k p(y_f, d_f = k|\{s^i\}) \\ &= \sum_k p_{\mathsf{x}^k}(y_f|s^k) \prod_{j\neq k} \Phi_{\mathsf{x}^j}(y_f|s^j). \end{aligned} \tag{5}$$

The probability that source k dominates is then simply:

$$\pi_k \triangleq p(d_f = k|y_f, \{s^i\}) = \frac{p(y_f, d_f = k|\{s^i\})}{p(y_f|\{s^i\})}, \tag{6}$$

and the expected value of x_f^k given $\{s^i\}$ is

$$E(\mathsf{x}_f^k|y_f, \{s^i\}) = \pi_k y + (1 - \pi_k)E(\mathsf{x}_f^k|d_f \neq k, \{s^i\}), \tag{7}$$

where

$$E(\mathsf{x}_f^k|d_f \neq k, \{s^i\}) = \mu_{f,s^k} - \frac{\sigma^2_{f,s^k} p_{\mathsf{x}_f^k}(y_f|s^k)}{\Phi_{\mathsf{x}_f^k}(y_f|s^k)} \tag{8}$$

for guassian $p_{\mathsf{x}_f^k}(y_f|s^k)$. Note that $E(\mathsf{x}_f^k|d_f \neq k, \{s^i\})$ only depends on acoustic state of source k, s^k.

V. The MSP algorithm

In this section we briefly review the max-sum product (MSP) loopy belief propagation algorithm presented in [4].

Inference using the MSP algorithm, roughly speaking, consists of iteratively decoding a single source given the current

estimates of the other sources. More specifically, inference consists of passing messages between random variables of the model depicted in Figure 1(b), according to the following message-passing schedule:

1. Compute approximate grammar likelihoods for source i for all t:

$$\hat{p}(\mathbf{y}_t|v_t^i) = \sum_{s_t^i} p(s_t^i|v_t^i)\hat{p}(\mathbf{y}_t|s_t^i) \tag{9}$$

2. Propagate messages forward for $t = 1..T$ and then backward for $t = T...1$ along the grammar chain of source i:

$$\hat{p}_{\text{fw}}(v_t^i) = \max_{v_{t-1}^i} p(v_t^i|v_{t-1}^i)\hat{p}_{\text{fw}}(v_{t-1}^i)\hat{p}(\mathbf{y}_t|v_{t-1}^i) \tag{10}$$

$$\hat{p}_{\text{bw}}(v_t^i) = \max_{v_{t+1}^i} p(v_{t+1}^i|v_t^i)\hat{p}_{\text{bw}}(v_{t+1}^i)\hat{p}(\mathbf{y}_t|v_{t+1}^i) \tag{11}$$

3. Update the conditional acoustic state prior of source i:

$$\hat{p}(s_t^i) = \sum_{v_t^i} p(s_t^i|v_t^i)\hat{p}_{\text{fw}}(v_t^i)\hat{p}_{\text{bw}}(v_t^i) \tag{12}$$

The arguments of the maximization in $\hat{p}_{\text{fw}}(v_t^i)$ are stored for all t so that the current MAP estimate of the grammar states of sources can be evaluated at the end of each iteration. This procedure is iterated for a specified number of iterations or until the MAP estimates of all sources converge. The Bayes net of our model (Figure 1(b)) has loops so there is no guarantee of convergence.

Surveying the message updates, we can see that combinations of grammar states are never considered: the MAP grammar state sequences of each speaker are estimated *independently* of one another *given* the current estimates of their marginal grammar state likelihoods, $\{\hat{p}(\mathbf{y}_t|v_t^i)\}$. Temporal inference on the grammar chains (step 2) therefore requires just $O(TD_v^2)$ operations per source, per iteration, rather than the $O(TD_v^{N+1})$ operations required to do exact inference. The algorithm, however, requires that the conditional acoustic marginal likelihoods of the speaker currently being decoded be computed:

$$\hat{p}(\mathbf{y}|s^k) = \sum_{\{s^j:j\neq k\}} \prod_{j\neq k} \hat{p}(s^j) \prod_f p(y_f|\{s^i\}) \tag{13}$$

In general this computation requires at least $O(D_s^N)$ operations per source, where D_s is the number of acoustic states per source, because all possible combinations of the acoustic states of the sources must be considered. In the case of the max model, unfortunately, if the features have more than one dimension, this is also the case. Under the max model, however, the likelihood in a single frequency band (5) consists of N terms, each of which *factor* over the states of the sources. This unique property can be exploited to efficiently approximate the marginal state likelihoods of the sources.

VI. Variational Inference in the Max Model

In this work, we generalize the variational framework presented in [5] to hierarchical acoustic models, which allows us to fully control the efficacy and complexity of the approximate inference procedure.

A. Hierarchical Acoustic Models

The hierarchical acoustic models for each speaker have the form:

$$p(\{l_i\}, x_f) = p(l_0)\prod_{i=1}^{L} p(l_i|l_{i-1})p(x_f|l_L) \tag{18}$$

where the $\{l_i\} = \{l_0, l_1, .., l_N\}$ are discrete random variables and i denotes the hierarchy level. The hierarchy was trained by successively clustering the GMM for level $i+1$ down from K_{i+1} to $K_i = K_{i+1}/B$ gaussians starting from level L, where B is a chosen "branching factor". The clustering was performed by minimizing a variational approximation to the KL-divergence [9], between the original and clustered GMM, using the algorithm described in [10]. The variational parameters of this algorithm provide the mapping between the original and clustered GMM components, $p(l_i|l_{i+1})$. Note that the hierarchy of states is not in general tree-structured, and that in the final model only the gaussians at the leaves of the tree are retained.

In this work, for a given source k, we reduce the hierarchical model to two levels during inference: one level for the low resolution states $c^k = l_i^k$ for a chosen i, which the probablistic time-frequency masks will condition on, and another for the leaf states $s^k = l_L^k$. The acoustic model for source k is then given by

$$p(c^k, s^k, x) = p(c^k)p(s^k|c^k)p(x^k|s^k), \tag{19}$$

where:

$$p(c^k) = \sum_{l_{0:i-1}^k} \prod_{i'=1}^{i} p(l_{i'}^k|l_{i'-1}^k), \tag{20}$$

$$p(s^k|c^k) = \sum_{l_{i+1:L-1}^k} \prod_{i'=i+1}^{L} p(l_{i'}^k|l_{i'-1}^k), \tag{21}$$

$p(x^k|s^k) = p(x^k|l_L)$, and $l_{0:i-1}^k = \{l_0^k, l_1^k, .., l_{i-1}^k\}$.

B. Exploiting acoustic hierarchies to compute arbitrarily tight variational bounds on the probability of mixed data

The log probability of the data given a particular combination of acoustic states $\{s^i\}$ under the max model is:

$$\begin{aligned}\log p(\mathbf{y}|\{s^i\}) &= \sum_f \log p(y_f|\{s^i\}) \\ &= \sum_f \log \sum_{d_f} p(y_f, d_f|\{s^i\})\end{aligned} \tag{22}$$

Where $p(y_f, d_f|\{s^i\})$ is the probability that source d dominates frequency band f and generates data y_f. Introducing variational masking distributions $q(d_f|\{c^i\})$ that condition of the low-resolution states of the sources, we can write the following bound for $\log p(\mathbf{y}|\{s^i\})$:

$$\begin{aligned}\log p(\mathbf{y}|\{s^i\}) &\geq \sum_f \sum_{d_f} q(d_f|\{c^i\}) \log \frac{p(y_f, d_f|\{s^i\})}{q(d_f|\{c^i\})} \\ &= \log \hat{p}(\mathbf{y}|\{s^i\})\end{aligned} \tag{23}$$

which holds for any $q(d_f|\{c^i\})$ by Jensen's inequality. It is easily verified that if $q(d_f|\{c^i\}) = p(d_f|\{s^i\})$ the bound

$$q(s^k|c^k) \propto p(s^k|c^k)\exp\left(\sum_f q(d_f=k|c_k)\log p_{\mathbf{x}_d^k}(y_d|s^k) + (1-q(d_f=k|c_k))\log\Phi_{\mathbf{x}_d^j}(y_d|s^k)\right) \quad (14)$$

$$q(\{c^i\}) \propto \prod_j p(c^j)\exp(-\sum_k \mathcal{D}(q(s^k|c^k)||p(s^k|c^k)) + \sum_f \mathcal{H}(q(d_f|\{c^i\})) + \sum_k q(d_f=k|\{c_i\})\left[E_{q(s^k|c^k)}[\log p_{\mathbf{x}_d^k}(y_d|s^k)] + \sum_{j\neq k} E_{q(s^j|c^j)}[\log\Phi_{\mathbf{x}_d^j}(y_d|s^j)]\right] \quad (15)$$

$$q(d_f=k|\{c^i\}) \propto \exp\left(E_{q(s^k|c^k)}[\log p_{\mathbf{x}_d^k}(y_d|s^k)] + \sum_{j\neq k} E_{q(s^j|c^j)}[\log\Phi_{\mathbf{x}_d^j}(y_d|s^j)]\right) \quad (16)$$

$$\log\hat{p}(\mathbf{y}|s^k) = E_{q(\{c^i\}|s^k)}[H\{q(d_f|\{c^i\})\}] + q(d_f=k|s^k)\log p_{\mathbf{x}_f^k}(y_f|s^k) + (1-q(d_f=k|s^k))\log\Phi_{\mathbf{x}_d^k}(y_f|s^k) + \sum_{j\neq k}\sum_{c^j} q(d_f=j,c^j|d,s^k)E_{q(s^j|c^j)}[\log p_{\mathbf{x}_f^j}(y_f|s^k)] + \sum_{c^j}(q(c^j|s^k) - q(d_f=j,c^j|d,s^k))E_{q(s^j|c^j)}[\log\Phi_{\mathbf{x}_d^j}(y_f|s^j)] \quad (17)$$

Box 1: Variational updates for the HVMSP algorithm. Each update increases the lower bound on the likelihood. The approximate marginal acoustic state log likelihoods for source k under the bound are also depicted. These are used to decode source k. The decode updates the acoustic state priors of source k, and the process is repeated for all i, and for multiple iterations.

is *tight*. Conditioning the probabilistic masks on the low-resolution states of the sources reduces the number of masks M from D_s^N to $\prod_{k=1}^N D_{c^k}$, where D_{c^k} is the number of low-resolution acoustic states used for source k. These masks could be constructed from the low-resolution GMMs of the sources given the data using (6). In this work we optimize the masks to maximize the probability of the observed mixed data under a unified variational framework for estimating the posterior distribution of the hidden variables of the sources.

In this paper we assume the following form for the posterior distribution of the unobserved variables of the model:

$$q(\{s^i\},\{c^i\},\{d_f\}) = q(\{c^i\})\prod_k q(s^k|c^k)\prod_f q(d_f|\{c^i\}) \quad (24)$$

This form models correlation in the posterior distribution of the low-resolution states of the sources (whose resolution can be arbitrarily set) and ignores correlation between the high resolution states of the speakers to make inference tractable.

Using this surrogate posterior, we can lower-bound the probability of the data as follows:

$$\begin{aligned}\log p(\mathbf{y}) &= \log\sum_{\{c^i\},\{s^i\}}\prod_k p(c^k,s^k)p(\mathbf{y}|\{s^i\}) \\ &\geq \sum_{\{c^i\},\{s^i\}} q(\{c^i\})\prod_k q(s^k|c^k)\log\frac{\prod_k p(c^k,s^k)p(\mathbf{y}|\{s^i\})}{q(\{c^i\})\prod_k q(s^k|c^k)} \\ &\geq -D(q(\{c^i\})\prod_k q(s^k|c^k)\,||\,\prod_k p(c^k,s^k)) + \\ &\qquad E_{q(\{c^i\})\prod_k q(s^k|c^k)}[\log\hat{p}(\mathbf{y}|\{s^i\})] \end{aligned} \quad (25)$$

where $D(q||p)$ is the relative entropy between q and p. The first bound follows again from Jensen's inequality, and the second follows from (23).

C. Computing variational acoustic likelihoods

Differentiating the lower-bound (25) w.r.t. the parameters of q and enforcing normalizing constraints leads to the set of closed-form but coupled update rules depicted in box 1, which are iterated to optimize the lower-bound and identify q. The expression for the approximate marginal log likelihood $\log\hat{p}(\mathbf{y}|s^k)$ under q for source k is also given. Because the state prior fed into this likelihood estimation algorithm during inference may be incorrect, we want to be able to extract the likelihoods of acoustic states with an estimated posterior probability of zero. To do this requires that the q distribution have the following form: $q(\{c^i\},\{s^i\},\{d_f\}) = q(s^k)q(c^k|s^k)q(\{c^i\}_{i\neq k}|c^k)\prod_{j\neq k}q(s^j|c^j)\prod_f q(d_f|\{c^i\})$, which is an equivalent representation, but allows for the extraction of *all* of the high resolution acoustic likelihoods directly from the update for $q(s^k)$. The chosen form of q decouples inference over the high resolution acoustic states of the sources: combinations of high resolution states are never considered. Acoustic inference scales as $O(NM + D_s\sum_k D_{c^k})$ per iteration over the updates for q, where $M = \prod_k D_{c^k}$ is the number of probabilistic time-frequency masks. The first term typically dominates the second one. The HVMSP algorithm is identical to the MSP algorithm, but approximates the conditional marginal likelihood (13), which is $O(D_s^N)$, with the approximation (17), which is $O(NM)$, each time a marginal grammar state likelihood (9) needs to be computed during inference (see section V for further details).

HVMSP, therefore, scales *linearly* with the number of probabilistic time-frequency masks per frame. The source features, furthermore, can be reconstructed in time linear in the number of time-frequency masks. The MMSE estimate of the source

features under the variational posterior is:

$$E_q[\mathsf{x}_f^k|\mathbf{y}] \;=\; q(d_f = k)y_f \;+$$
$$E_{q(c^k)}[(1 - q(d_f = k|c^k))E_{q(s^k|c^k)}[\mu_{s^k} - \frac{\sigma^2_{s^k} p_{\mathsf{x}^k}(y|s^k)}{\Phi_{\mathsf{x}^k}(y|s^k)}]] \tag{26}$$

which is analogous to the MMSE estimate when exact inference is done in the max model by averaging (7) over $p(\{s^i\}|y)$.

VII. Experiments

Table I summarizes the error rate performance of our multi-talker speech recognition system on the 0 dB portion of the SSC task [2] as a function of separation algorithm. Also depicted are the number of probablistic time-frequency masks utilized by each algorithm on a per-frame basis, which correlates directly with the computational complexity of acoustic inference. In all cases, the loopy belief propagation message passing schedule was executed for 10 iterations, and in the case of VMSP and HVMSP, the variational updates were iterated 10 times each time the conditional marginal grammar likelihood (9) of a source was computed. In these experiments the factors of q were initialized to their priors, with the exception of the time-frequency masks, which were initialized to be uniformly distributed. MMSE estimates of the features of the source receiving message (9) were reconstructed using (26) for VMSP and HVMSP, each time this message was sent. In the case of MSP and the Joint Viterbi algorithms, the conditional MMSE estimates of the speaker features given the MAP grammar sequences of the sources were used to do reconstruction. The reconstructed speaker signals were then fed into a conventional ASR system that does speaker-dependent labeling [3] for recognition. In all cases oracle speaker ids and gains were utilized. Note that the presented system implements a multi-talker speech recognition system, but better recognition results were obtained by doing post-reconstruction recognition as described, possibly because the shared set of gaussians in the separation model is not discriminative enough: as we increase the number of gaussians in the separation system the WER discrepancy decays rapidly.

Looking at the results, we can see that when the probabilistic masks are conditioned on just 8 low-resolution states per source (64 masks total), HVMSP scores 28.4%, which is 2% absolute better than the VMSP result, which utilizes 256 masks per marginal likelihood calculation. When the probabilistic masks are conditioned on just 16 low resolution states per source (256 masks total), HVMSP amazingly performs as well as MSP, which utilizes 65536 masks in total.

Table II depicts WER results obtained using the HVMSP algorithm for 3 speaker separation as a multi-talker decoder. Here the HVMSP separation algorithm uses $D_s = 1024$ high resolution acoustic states per source to improve recognition accuracy, eliminating the need for post-separation recognition processing. Exact acoustic under the model involves computing $1024^3 > 10^9$ acoustic masks, which is intractable. Using only 4096 masks HVMSP achieves an impressive 34.0% error rate on the three-talker data set.

Algorithm	# Masks/Frame	WER
Humans	?	27.7
Joint Viterbi	$256^2 = 65536$	**22.4**
MSP	$256^2 = 65536$	**25.6**
VMSP	256	30.4
HVMSP	$2^2 = 4$	37.6
	$4^2 = 16$	32.3
	$8^2 = 64$	28.4
	$16^2 = 256$	**25.2**

TABLE I

WER (LETTER AND DIGIT) AS A FUNCTION OF ALGORITHM AND NUMBER OF PROBABILISTIC TIME-FREQUENCY MASKS PER FRAME ON THE 0 DB PORTION OF THE SSC TASK. IN ALL CASES ORACLE SPEAKER IDENTITIES AND GAINS WHERE USED. HVMSP OUTPERFORMS VMSP BY OVER 2% ABSOLUTE USING 4 TIMES LESS TIME-FREQUENCY MASKS. HVMSP FURTHERMORE, PERFORMS ON-PAR WITH MSP, WHICH COMPUTES EXACT CONDITIONAL ACOUSTIC MARGINALS, USING 256 TIMES LESS TIME-FREQUENCY MASKS. THE PERFORMANCE DISCREPANCY IS MEASUREMENT NOISE (A SINGLE ERROR). THE VITERBI ALGORITHM SCALES EXPONENTIALLY WITH LM SIZE. ALL OTHER ALGORITHMS SCALE LINEARLY WITH LM SIZE. RESULTS EXCEEDING HUMAN PERFORMANCE ARE BOLDED.

# Masks $M = D_{c_f} * D_{c_b}^2$	Target Speaker (F)	Masker 1 (M)	Masker 2 (F)	Overall
$16 * 4^2 = 256$	42.9	31.1	42.9	38.9
$16 * 8^2 = 1024$	38.5	33.0	41.0	37.5
$1024 * 1^2 = 1024$	40.0	28.0	37.0	35.0
$16^3 = 4096$	34.5	30.4	37.1	34.0

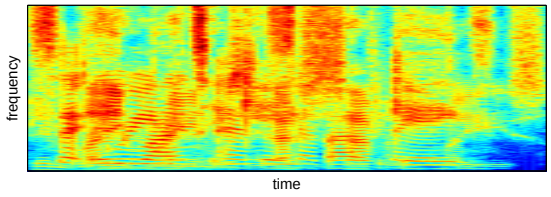
TABLE II

WER (LETTER AND DIGIT) AS A FUNCTION OF THE NUMBER OF TIME-FREQUENCY MASKS USED BY HVMSP FOR SYNTHETIC MIXTURES OF 3 SPEAKERS (100 UTTERANCES). HERE THE HVMSP SEPARATION ALGORITHM IS USED AS A MULTI-TALKER DECODER, AND USES $D_s = 1024$ HIGH RESOLUTION ACOUSTIC STATES PER SOURCE TO IMPROVE RECOGNITION ACCURACY. THE MASKS CONDITION ON D_{c_f} LOW-RESOLUTION ACOUSTIC STATES OF THE FOREGROUND SOURCE WHOSE LIKELIHOOD IS CURRENTLY BEING APPROXIMATED, AND D_{c_b} LOW-RES. STATES OF THE OTHER SOURCES. THE SNR OF THE TARGET SPEAKER IS 0 DB. THE AVERAGE SNR OF THE MASKING SPEAKERS IS -4.8 DB. IN ALL CASES, ORACLE SPEAKER IDENTITIES, GAINS, AND GRAMMAR MODELS WERE USED. DE-MIXED UTTERANCES FROM THE SSC TEST SET WERE MIXED DIRECTLY ON TOP OF EACH ANOTHER TO CONSTRUCT THE MIXTURES. EXACT INFERENCE UNDER THE MODEL INVOLVES COMPUTING OVER ONE BILLION ACOUSTIC MASKS, WHICH IS INTRACTABLE. USING ONLY 4096 MASKS HVMSP ACHIEVES AN IMPRESSIVE 34.0% ERROR RATE ON THE THREE TALKER DATA SET.

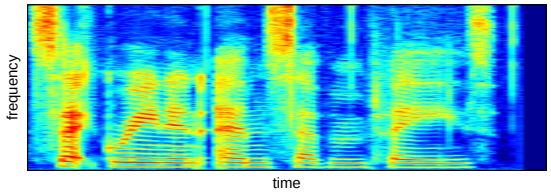
Figure 2 depicts separation results for a synthetic mixture of four sources. Here the speech models utilized by HVMSP to generate this result have $D_s = 1024$ high-resolution acoustic states per source. Exact inference using the max model given these source models involves computing over a trillion time-frequency masks per frame. Here only $M = D_{c_f} D_{c_b}^3 = 16 * 4^3 = 1024$ masks per frame are used to separate and decode the sources. In this example all four speakers were decoded correctly.

The fact we can so precisely control the complexity of acoustic inference using the presented hierarchical framework is a distinguishing property of HVMSP. While several variational algorithms for model-based analysis of mixed feature data exist, most of them are restricted to recovering a uni-modal estimate of the posterior distribution of the features.

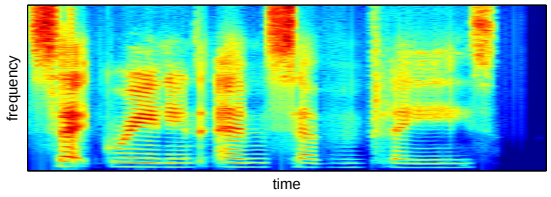
An important and direction of future work will be to optimize



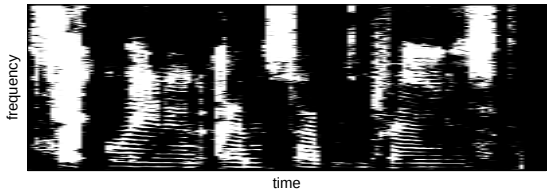
(a) Log Power Spectrogram of Mixture of 4 Speech Sources



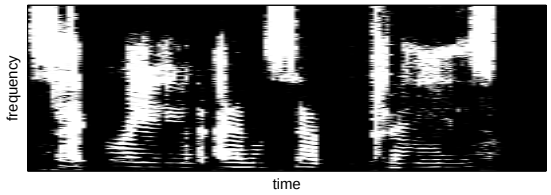
(b) Log Power Spectrogram of Speaker 4



(c) Estimated Log Power Spectrogram of Speaker 4



(d) True Power-ratio Mask



(e) Estimated Power-ratio Mask

Fig. 2. Separation results for a synthetic mixture of four sources, which were generated as described in figure II. The SNRs of the target and masking sources are 0 dB and -7 dB, respectively. The log power spectrum of the mixed signal, and the true and estimated log power spectrum of a masking source, source 4, are depicted, as are the estimated and actual power-ratio masks for speaker 4, which were computed as $r = \exp(x^4)/\sum_k \exp(x^k)$. Note that this power ratio ignores phase interactions, which are inconsistent with the use of soft binary masks. In this example all four speakers were decoded correctly.

and characterize the speed and performance of HVMSP. Preliminary experiments indicate that HVMSP far outperforms exact inference with the low-resolution acoustic model, and that iterating the variational updates improves performance substantially over using masks derived directly from the low-resolution model, but the performance discrepancies and trade-offs of these approaches need to be more fully characterized. We are also currently working on a variant of HVMSP that, rather than fixing the resolution of the acoustic states that the probabilistic time-frequency masks condition on, does a variational search of the acoustic state hierarchy of the sources during inference to further improve the speed-performance characteristics of HVMSP. This approach differs from a standard hierarchical search methods in that the hierarchy is expanded based on the estimated *best* scoring gaussians at full resolution, as opposed to the best scoring gaussians at the current resolution of the search expansion. The former approach (tree-expansion based on the best scoring paths) has been shown to be a more effective search strategy than the latter (tree-expansion based on the average score of the paths originating at a given node) in computational approaches to games like Go. It will be interesting to see if these results hold in the context of searching acoustic hierarchies. Of course a hybrid approach that does expansions based on the low-resolution gaussians early in the search and then switches to a variational mode of search may yield the best speed-performance trade-off.

In any case, these results are exciting because, while this technology is still far from mature enough to be deployed and many practical challenges remain, they demonstrate that the approach of modelling multiple acoustic sources in the environment to achieve robust ASR can be made feasible.

REFERENCES

[1] John R. Hershey, Steven J. Rennie, Peder A. Olsen, and Trausti T. Kristjansson, "Super-human multi-talker speech recognition: A graphical modeling approach," *Computer Speech and Language*, 2009.

[2] M. Cooke, J. R. Hershey, and S. J. Rennie, "The speech separation and recognition challenge," *Computer Speech and Language*, 2009.

[3] J. Hershey, T. Kristjansson, S. Rennie, and P. Olsen, "Single channel speech separation using layered hidden Markov models," *NIPS*, pp. 593–600, 2006.

[4] S. J. Rennie, J. R. Hershey, and P. A. Olsen, "Single-channel speech separation and recognition using loopy belief propagation," *ICASSP*, 2009.

[5] S. J. Rennie, J. R. Hershey, and P. A. Olsen, "Variational loopy belief propagation for multi-talker speech recognition," *INTERSPEECH*, 2009.

[6] A. Nádas, D. Nahamoo, and M. Picheny, "Speech recognition using noise-adaptive prototypes," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 37, pp. 1495–1503, 1989.

[7] A.P. Varga and R.K. Moore, "Hidden Markov model decomposition of speech and noise," *ICASSP*, pp. 845–848, 1990.

[8] M.H. Radfar, R.M. Dansereau, and A. Sayadiyan, "Nonlinear minimum mean square error estimator for mixture-maximisation approximation," *Electronics Letters*, vol. 42, no. 12, pp. 724–725, 2006.

[9] John Hershey and Peder Olsen, "Approximating the Kullback Leibler divergence between gaussian mixture models," in *ICASSP*, Honolulu, Hawaii, April 2007.

[10] Pierre L. Dognin, John R. Hershey, Vaibhava Goel, and Peder A. Olsen, "Refactoring acoustic models using variational density approximation," in *ICASSP*, April 2009, pp. 4473–4476.