Variational Loopy Belief Propagation for Multi-talker Speech Recognition

Steven J. Rennie, John R. Hershey, Peder A. Olsen

IBM T.J. Watson Research Center
{(sjrennie, jrhershe, pederao)@us.ibm.com}

Abstract

We address single-channel speech separation and recognition by combining loopy belief propagation and variational inference methods. Inference is done in a graphical model consisting of an HMM for each speaker combined with the max interaction model of source combination. We present a new variational inference algorithm that exploits the structure of the max model to compute an arbitrarily tight bound on the probability of the mixed data. The variational parameters are chosen so that the algorithm scales linearly in the size of the language and acoustic models, and quadratically in the number of sources. The algorithm scores 30.7% on the SSC task [1], which is the best published result by a method that scales linearly with speaker model complexity to date. The algorithm achieves average recognition error rates of 27%, 35%, and 51% on small datasets of SSC-derived speech mixtures containing two, three, and four sources, respectively, using a single audio channel.

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

1. 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 interpreted by the human auditory system, but is highly detrimental to conventional ASR.

In [2], a system for separating and recognizing multiple speakers using a single channel is presented. The system won the recently introduced monaural speech separation challenge [1], and even outperformed human listening results on the task. The performance of this system hinges on the separation component of the system, which models each speaker by a layered, factorial hidden Markov model (HMM). 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 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.

In this paper, we present a model-based algorithm for multi-talker speech separation and recognition using a single channel, which combines loopy belief propagation and variational inference methods to scale linearly with acoustic and language model size. The method is based upon a new variational framework for approximating the acoustic likelihoods of the sources using the max interaction model. The framework allows us to compute an arbitrarily tight bound on the probability of the data. Optimizing the bound involves computing a set of probabilistic masks that define what frequency bins are dominated by each source. By iteratively conditioning the masks on the acoustic states of single source, considering combinations of source states is avoided. In this sense the algorithm bears resemblance to missing-feature methods, which infer probabilistic masks to isolate a target speaker, but do not explicitly model the other sources in the environment.

\[ y(t) = \sum_k x_k(t). \]  
\[ |Y|^2 = \sum_k |X|^2 + \sum_{j \neq k} |X|^2 |X|^2 \cos(\theta_j - \theta_k). \]

Figure 1: a) Generative model (GM) for the features, \( x^a \), of single source: an HMM with grammar state, \( v^a \), sharing common acoustic states, \( s^a \). b) GM of mixed features for two sources. The source models are combined with an interaction model to explain the data. Here \( x^a \) and \( x^b \) have been integrated out.

2. 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 (Figure 1(a)). 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 \( x^a \) of source \( k \) given the discrete acoustic state \( s^a \) is modeled as a diagonal covariance Gaussian, \( p(x^a|s^a) = \prod_f \mathcal{N}(x^a_f | \mu_f(s^a_k), \sigma_f(s^a_k)) \), 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 \( D_x = 256 \) gaussians per speaker unless otherwise noted.

**Grammars:** The task grammar is represented by a sparse matrix of state transition probabilities, \( p(v^a_i | v^a_{i-1}) \). The association between the grammar state \( v^a \) and the acoustic state \( s^a \) is captured by the transition probability \( p(s^a_k | v^a_i) \), for speaker \( k \). These are learned from training data using inferred acoustic and grammar state sequences.

3. Interaction Model

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

\[ y(t) = \sum_k x_k(t). \]  

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

\[ |Y|^2 = \sum_k |X|^2 + \sum_{j \neq k} |X|^2 |X|^2 \cos(\theta_j - \theta_k). \]
where $\theta_k$ 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 $x^s \triangleq \log |X^s|^2$ and $y \triangleq \log |Y|^2$. Assuming uniformly distributed source phases, $E(\log |X^s|^2|X^k) = \sum_k |X^k|^2$. 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 is equivalent to:

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

and historically motivated the max approximation to $y$,

$$y \approx \max_k x^k.$$  \hspace{1cm} (3)

The max approximation was first used in [5] for noise adaptation. In [6], the max approximation was used to compute joint state likelihoods of speech and noise and find their optimal state sequence under a factorially hidden Markov model (HMM) of the sources. Recently [7] showed that in fact $E_p(y|x^s, x^k) = \max(x^s, x^k)$ 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:

$$p(y|x^s) = \delta(y - \max_k x^k),$$  \hspace{1cm} (4)

where $\delta()$ is the Dirac delta function.

### 4. Exact Inference in the Max Model

In this section we review how the joint acoustic state likelihoods of the speakers, $p(y|x^s)$, and the conditional expectations of the features of speaker $k$, $E(x^k|\{x^s\})$, are computed at each frequency. These quantities form the basis of any exact inference strategy. As in the previous section, frequency subscripts are omitted wherever possible for simplicity.

Let $p_k(y|x^s) \equiv p(x^s = y|x^s)$ for random variable $x^s$, and $\Phi_y(y|x^s) \equiv p(x^s \leq y|x^s) = \int_0^y p(x^s)$ be the cumulative distribution of $x^s$ evaluated at $y$. Following [5, 4],

$$p(y \leq y|x^s) = p(\max_k x^k \leq y|x^s),$$  \hspace{1cm} (5)

since the sources generate their features independently. The state likelihoods given $y$ are then obtained by differentiating:

$$p(y|\{x^s\}) = \sum_k p_k(y|x^s) \prod_{j \neq k} \Phi_y(y|x^j).$$  \hspace{1cm} (6)

From this we readily see that the individual terms in the above sum correspond to $p(y = y, x^s = y|x^s) = \delta(x^s = y|\{x^s\})$, which is the conditional probability that source $k$ is maximum then:

$$p_k = p(x^k = y|y, \{x^s\}) = \left( \sum_j p_k(y|x^s) \right)^{-1} \frac{p_k(y|x^s)}{\Phi_y(y|x^s)}.$$  

and the expected value of $x^k$ given $\{\tilde{x^s}\}$ is

$$E(x^k|y, \{\tilde{x^s}\}) = \pi_k y + (1 - \pi_k) E(x^k|x^k < y, \{\tilde{x^s}\}),$$

The utility of the max model hinges upon how readily $p_k(y|x^s)$, $\Phi_y(y|x^s)$, and $E(x^k|x^k < y, \{\tilde{x^s}\})$ can be computed. In this paper we assume that the sources, conditioned on their states, are gaussian-distributed at each frequency:

$$p(x^k) = N(x^k = y|\mu_k, \sigma^2_k),$$  \hspace{1cm} (7)

$$\Phi_y(y|x^s) = \int_0^y N(x^k = y|\mu_k, \sigma^2_k)dy,$$

$$E(x^k|x^k < y, \{\tilde{x^s}\}) = \mu_k - \frac{\sigma^2_k}{\Phi_y(y|x^s)}.$$  \hspace{1cm} (8)

### 5. Variational Inference in the Max Model

The loopy belief propagation algorithm presented in [4] and extended in this paper requires that the marginal likelihoods

$$p(y|x^s) = \sum_{\{f\} \neq k} \prod_f \tilde{p}(s^f) \prod_f p(y_f|\{s^f\})$$  \hspace{1cm} (9)

be iteratively computed for each source. In general this computation requires at least $O(D^2 s)$ operations per source, where $D_s$ is the number of acoustic states per source, because all possible combinations of source states 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 (6) 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. The log-probability of a mixed feature $y = [y_1, \ldots, y_T]$ under the max model is:

$$\log p(y) = \log \sum_{\{s^f\}} \prod_f \tilde{p}(s^f) \prod_f p(y_f|\{s^f\}),$$  \hspace{1cm} (10)

Using Jensen's inequality, a lower bound on $\log p(y)$, $L$, is formed by introducing the variational distribution $q(\{s^f\})$, as shown in box 1, equation (12). A further bound on $\log p(y)$, $L'$, is obtained by introducing the variational distribution $q(k|f, \{s^f\})$ to take the sum over $k$ outside the log in $L$, as shown in box 1, equation (14). This bound differs from those derived using standard variational inference methods in that the variational distribution $q(k|f, \{s^f\})$ is defined over variable $k$, which is not in the generative model for the data.

The tightness of the bound $L'$ depends on the dependency structure and parameters of $q(k|f, \{s^f\})$ and $q(\{s^f\})$. The bound is tight if $q(\{s^f\}) = p(\{s^f\}|y)$ and $q(k|f, \{s^f\}) = p(x_f = y_f|y_f = y_f, \{s^f\})$, since $p(x_f = y_f|y_f = y_f, \{s^f\}) / p(x_f = y_f|y_f = y_f, \{s^f\}) = p(y_f = y_f|y_f = y_f, \{s^f\})$, which is independent of variable $k$. Thus $q(k|f, \{s^f\})$ can be interpreted as a probabilistic mask, representing the a posteriori probability that feature bin $f$ is dominated by source $k$, given a set of source states $\{s^f\}$.

Without constraints on the variational parameters, inference would be exponentially complex in the number of sources. To
Box 1: Variational bounds on the log probability of the data under the max model. Here $D$ denotes Kullback-Leibler divergence and $H$ denotes entropy. When $q(\{s^k\}) = p(\{s^k\}|y)$ and $q(k|f, \{s^k\}) = p(k = f|y_f = y_f, \{s^k\})$, the bound $\mathcal{L}'$ is tight.

6. Loopy Belief Propagation

Inference using belief propagation (BP) \cite{8, 9} consists of passing messages between connected variables of the model according to a message passing schedule. If the model is tree-structured, and no messages are approximated, the max-product variant of BP can be used to recover the exact MAP configuration of the variables. In this regard it performs the same function as the Viterbi algorithm for HMMs. Similarly the sum-product variant of BP can recover the exact marginals of all variables in an efficient manner. For factorial HMMs it reduces to the forward-backward algorithm. We can do exact inference on our factorial HMM, using either max-product or sum-product BP by combining the acoustic and grammar states, respectively, across the different decoding results of the other sources. Combining the max-product and sum-product algorithms to implement this approach leads to the following message-passing schedule, called the max-sum product (MSP) algorithm \cite{2}. Each source in turn receives messages (i.e., probability estimates) from the other sources. Initially all messages are initialized to be uniform, and $\hat{p}(v_i^f)$ is initialized to the prior for $v_i^f$ for all $k$. The messages require $p(y|s^k_i)$ (9), which is computed exactly for MSP, and using (15) in Box 2, for Variational MSP (VMS). For a given source $i$ the incoming messages are computed in three steps:

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

   \[
p(\tilde{x}_t^i|v_i^f) = \sum_{s^k_i} p(s^k_i|v_i^f)\hat{p}(x_t^i|s^k_i)
   \]

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

   \[
   \hat{p}_{bw}(v_i^f) = \max_{v_i^{f-1}} p(v_i^f|v_i^{f-1}, s^k_i)\hat{p}_{bw}(v_i^{f-1})
   \]

   \[
   \hat{p}_{bw}(v_i^f) = \max_{v_i^{f+1}} p(v_i^f|v_i^{f+1}, s^k_i)\hat{p}_{bw}(v_i^{f+1})
   \]

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

   \[
   \hat{p}(s_i^k) = \sum_{v_i^f} p(s_i^k|v_i^f)\hat{p}(s_i^k|v_i^f)
   \]

The arguments of the maximization in $\hat{p}_{bw}(v_i^f)$ 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.

7. Experiments

Tables 1 and 2 summarize the error rate and complexity of our multi-talk speech recognition system on the SSC task \cite{1}, as a function of separation algorithm. Recognition was done on the reconstructed target signal using a conventional single-talker speech recognition system that does speaker-dependent labeling \cite{3}. For all iterative algorithms, the message passing schedule was executed for 10 iterations. After inferring the grammar state sequences, conditional minimum mean squared error (MMSE) estimates of the sources were reconstructed. Table 3 summarizes the overall task error rate of the three best published results on the SSC task by algorithms that scale linearly with speaker model size, and the top performing system for reference. Here estimated speaker identities and gains, output by the
Table 1: SSC task error rate as a function of separation algorithm and test case. Conditions are: *same talker* (ST), *same gender* (SG), *different gender* (DG). In all cases the max model was used to approximate the acoustic likelihoods of the sources. Results exceeding human performance arebolded. In all cases oracle speaker identities and gains were used.

<table>
<thead>
<tr>
<th>Beam size</th>
<th>Viterbi</th>
<th>Beam size</th>
<th>Viterbi</th>
<th>Viterbi</th>
<th>Viterbi</th>
<th>Viterbi</th>
</tr>
</thead>
<tbody>
<tr>
<td>ST</td>
<td>34.0</td>
<td>36.4</td>
<td>38.6</td>
<td>53.3</td>
<td>51(3.3)</td>
<td></td>
</tr>
<tr>
<td>SG</td>
<td>19.5</td>
<td>14.0</td>
<td>14.4</td>
<td>17.7</td>
<td>16(3.3)</td>
<td></td>
</tr>
<tr>
<td>DG</td>
<td>11.9</td>
<td>10.8</td>
<td>10.8</td>
<td>14.2</td>
<td>12(2.2)</td>
<td></td>
</tr>
<tr>
<td>Overall</td>
<td>22.5</td>
<td>21.2</td>
<td>22.1</td>
<td>29.7</td>
<td>27(2.8)</td>
<td></td>
</tr>
</tbody>
</table>

Table 2: Task error rate (for $D_v = 256$, $D_m = 506$, $N = 2$) and complexity of temporal and acoustic inference as a function of algorithm and beam size ($B$). In all cases oracle speaker identities and gains were used.

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Viterbi</th>
<th>Viterbi</th>
<th>Viterbi</th>
<th>Viterbi</th>
</tr>
</thead>
<tbody>
<tr>
<td>Beam size</td>
<td>20000</td>
<td>400</td>
<td>Full</td>
<td>Full</td>
</tr>
<tr>
<td>Error Rate</td>
<td>21.2</td>
<td>22.2</td>
<td>22.1</td>
<td>29.7</td>
</tr>
<tr>
<td>Acoustic</td>
<td>$O(BD_vN)$</td>
<td>$O(BD_vN)$</td>
<td>$O(ND_v)$</td>
<td>$O(ND_v)$</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 3: Task error rate as a function of the top performing algorithms on the SSC task. The system presented by Hershey et al. (Algonquin-based result depicted) performs the best on the task but scales exponentially with speaker model size. VMSP is the best-scoring algorithm that scales linearly with model size. Here $D_v = 256$ and estimated speaker identities and gains [2] were used. For references consult [1]. 1 denotes “et al.”

<table>
<thead>
<tr>
<th>Method</th>
<th>Viterbi</th>
<th>VMSP</th>
<th>Iterative Viterbi</th>
<th>Fragment Decoding</th>
</tr>
</thead>
<tbody>
<tr>
<td>Inference</td>
<td>Exponential</td>
<td>Linear</td>
<td>Linear</td>
<td></td>
</tr>
<tr>
<td>TER</td>
<td>21.6</td>
<td>30.7</td>
<td>34.2</td>
<td>35.2</td>
</tr>
</tbody>
</table>

Table 4: WER (letter and digit) as a function of number of sources for synthetic mixtures (100 utterances) obtained using the VMSP separation algorithm as a multi-talker decoder. The SNR of the target speaker is 0 dB. The average SNR of the masking speaker(s) is 0 dB, -4.8 dB, and -7 dB for the 2, 3, and 4 source mixing scenarios, respectively. 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. Here $D_v = 1024$.

<table>
<thead>
<tr>
<th>Number of Speakers</th>
<th>Target Speaker (F)</th>
<th>Masker 1 (M)</th>
<th>Masker 2 (F)</th>
<th>Masker 3 (M)</th>
<th>Overall</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>27</td>
<td>17</td>
<td>-</td>
<td>-</td>
<td>22</td>
</tr>
<tr>
<td>3</td>
<td>40</td>
<td>28</td>
<td>-</td>
<td>-</td>
<td>35</td>
</tr>
<tr>
<td>4</td>
<td>47</td>
<td>58</td>
<td>51</td>
<td>51</td>
<td>51</td>
</tr>
</tbody>
</table>

8. References

