INTERSPEECH 2009 BRIGHTON

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.

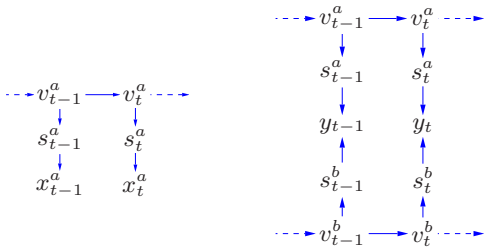
Index Terms: Speech separation, variational inference, loopy belief propagation, factorial hidden Markov models, ASR, Iroquois, Max model.

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.



(a) Speaker Feature Model (b) Mixed Feature Model

Figure 1: a) Generative model (GM) for the features, x^a, of single source: an HMM with grammar states, 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 $\mathbf{x}^k$ of source k given the discrete acoustic state s^k is modeled as a diagonal covariance Gaussian, $p(\mathbf{x}^a|s^a) = \prod_f \mathcal{N}(x^k_f; \mu_{f,s^k}, \sigma^2_{f,s^k})$, 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 per speaker unless otherwise noted.

Grammars: The task grammar is represented by a sparse matrix of state transition probabilities, $p(v^k_t|v^k_{t-1})$. 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.

3. Interaction Model

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

$$y(t) = \sum_k x_k(t). \quad (1)$$

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

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

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, $E\big(|Y|^2|\{X^k\}\big) = \sum_k |X^k|^2$. 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 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,

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

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 factorial hidden Markov model (HMM) of the sources. Recently [7] showed that in fact $E_\theta(\mathsf{y}|x^a, x^b) = \max(x^a, x^b)$ 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:

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

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, $p(y|\{s^{\tilde{k}}\})$, and the conditional expectations of the features of speaker k, $E(x^k|\{s^{\tilde{k}}\})$, 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 $p_{\mathsf{x}^k}(y|s^k) \triangleq p(\mathsf{x}^k = y|s^k)$ for random variable x^k, and $\Phi_{\mathsf{x}^k}(y|s^k) \triangleq p(\mathsf{x}^k \leq y|s^k) = \int_{-\infty}^{y} p(x^k)$ be the cumulative distribution of x^k evaluated at y. Following [5, 4]:

$$\begin{aligned} p(\mathsf{y} \leq y|\{s^{\tilde{k}}\}) &= p(\max_k \mathsf{x}^k \leq y|\{s^{\tilde{k}}\}), \\ &= \prod_k \Phi_{\mathsf{x}^k}(y|s^k), \end{aligned} \tag{5}$$

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

$$p(y|\{s^{\tilde{k}}\}) = \sum_k p_{\mathsf{x}^k}(y|s^k)\prod_{j\neq k}\Phi_{\mathsf{x}^j}(y|s^j). \tag{6}$$

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

$$\pi_k \triangleq p(\mathsf{x}^k = y|\mathsf{y} = y, \{s^{\tilde{k}}\}) = \left(\sum_j \frac{p_{\mathsf{x}^j}(y|s^j)}{\Phi_{\mathsf{x}^j}(y|s^j)}\right)^{-1} \frac{p_{\mathsf{x}^k}(y|s^k)}{\Phi_{\mathsf{x}^k}(y|s^k)},$$

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

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

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

$$p(\mathsf{x}^k) = \mathcal{N}(\mathsf{x}^k = y|\mu_{s^k}, \sigma^2_{s^k}), \tag{7}$$

$$\Phi_{\mathsf{x}^k}(y|s^k) = \int_{-\infty}^{y} \mathcal{N}(\mathsf{x}^k = y|\mu_{s^k}, \sigma^2_{s^k})\mathrm{d}y,$$

$$E(\mathsf{x}^k|\mathsf{x}^k < y, \{s^{\tilde{k}}\}) = \mu_{s^k} - \frac{\sigma^2_{s^k} p_{\mathsf{x}^k}(y|s^k)}{\Phi_{\mathsf{x}^k}(y|s^k)}. \tag{8}$$

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

$$\hat{p}(\mathbf{y}|s^k) = \sum_{\{s^j : j\neq k\}} \prod_{j\neq k} \hat{p}(s^j) \prod_f p(y_f|\{s^{\tilde{k}}\}) \tag{9}$$

be iteratively computed for each source. 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 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 $\mathbf{y} = [y_1, ..y_f, .., y_F]^T$ under the max model is:

$$\begin{aligned} \log p(\mathbf{y}) &= \log \sum_{\{s^{\tilde{k}}\}} \prod_{k'} \hat{p}(s^{k'}) \prod_f p(y_f|\{s^{\tilde{k}}\}), \\ &= \log \sum_{\{s^{\tilde{k}}\}} \prod_{k'} \hat{p}(s^{k'}) \prod_f (\sum_k p_{\mathsf{x}^k}(y_f|s^k) \prod_{j\neq k} \Phi_{\mathsf{x}^j}(y_f|s^j)). \end{aligned} \tag{10}$$

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

The tightness of the bound $\mathcal{L}'$ depends on the dependency structure and parameters of $q(k|f, \{s^{\tilde{k}}\})$ and $q(\{s^{\tilde{k}}\})$. The bound is *tight* if $q(\{s^{\tilde{k}}\}) = p(\{s^{\tilde{k}}\}|\mathbf{y})$ and $q(k|f, \{s^{\tilde{k}}\}) = p(\mathsf{x}^k_f = y_f|\mathsf{y}_f = y_f, \{s^{\tilde{k}}\})$, since $p(\mathsf{x}^k_f = y_f, \mathsf{y}_f = y_f|\{s^{\tilde{k}}\})$ / $p(\mathsf{x}^k_f = y_f|\mathsf{y}_f = y_f, \{s^{\tilde{k}}\}) = p(\mathsf{y}_f = y_f, \{s^{\tilde{k}}\})$, which is independent of variable k. Thus $q(k|f, \{s^{\tilde{k}}\})$ 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 $\{s^{\tilde{k}}\}$.

Without constraints on the variational parameters, inference would be exponentially complex in the number of sources. To

$$\log p(\mathbf{y}) = \log \sum_{\{s^{\tilde{k}}\}} \frac{q(\{s^{\tilde{k}}\})}{q(\{s^{\tilde{k}}\})} \prod_{k'} \hat{p}(s^{k'}) \prod_f \sum_k p_{\mathsf{x}^k}(y_f|s^k) \prod_{j\neq k} \Phi_{\mathsf{x}^j}(y_f|s^j), \tag{11}$$

$$\geq \mathcal{L} = -D(q(\{s^{\tilde{k}}\})||\hat{p}(\{s^{\tilde{k}}\})) + \sum_{\{s^{\tilde{k}}\}} q(\{s^{\tilde{k}}\}) \sum_f \log \sum_k p_{\mathsf{x}^k}(y_f|s^k) \prod_{j\neq k} \Phi_{\mathsf{x}^j}(y_f|s^j). \tag{12}$$

$$\mathcal{L} = -D(q(\{s^{\tilde{k}}\})||\hat{p}(\{s^{\tilde{k}}\})) + \sum_{\{s^{\tilde{k}}\},f} q(\{s^{\tilde{k}}\}) \log \left(\sum_k \frac{q(k|f,\{s^{\tilde{k}}\})}{q(k|f,\{s^{\tilde{k}}\})} p_{\mathsf{x}^k}(y_f|s^k) \prod_{j\neq k} \Phi_{\mathsf{x}^j}(y_f|s^j) \right), \tag{13}$$

$$\geq \mathcal{L}' = -D(q(\{s^{\tilde{k}}\})||\hat{p}(\{s^{\tilde{k}}\})) + \sum_f E_{q(k,\{s^{\tilde{k}}\}|f)} \Big(\log p_{\mathsf{x}^k}(y_f|s^k) + \sum_{j\neq k} \log \Phi_{\mathsf{x}^j}(y_f|s^j) \Big) + \sum_f H(q(k|f,\{s^{\tilde{k}}\})). \tag{14}$$

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 $q(\{s^{\tilde{k}}\}) = p(\{s^{\tilde{k}}\}|\mathbf{y})$ and $q(k|f,\{s^{\tilde{k}}\}) = p(\mathsf{x}_f^k = y_f|\mathsf{y}_f = y_f, \{s^{\tilde{k}}\})$, the bound $\mathcal{L}'$ is *tight*.

make inference tractable we constrain the variational parameters to $q(\{s^{\tilde{k}}\}) = \prod_k q(s^k)$ and $q(k|f,\{s^{\tilde{k}}\}) = q(k|f,s^i)$ where i is the index of the source receiving a message from the other sources during loopy belief propagation. This makes the message computation scale *linearly* with acoustic model size. Here we compute a new bound each time a marginal likelihood is needed, which makes the algorithm quadratic in N.

Optimizing $\mathcal{L}'$ w.r.t. to the variational parameters leads to iterative updates for the components of q, for messages sent to source i, as shown in Box 2. Surveying the updates for q, we can see that combinations of source states are never considered. The optimization of q scales *linearly* with acoustic model size and the number of sources. The source features, furthermore, can be reconstructed in time linear in the number of source by replacing π_k with $q(k|f,s^i)$ in (7).

6. Loopy Belief Propagation

Inference using belief propagation (BP) [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 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 sources, but inference is $O(\max(D_s, D_v)^{N+1})$ [3].

To avoid the combinatorial explosion of exact inference, we iteratively estimate the configurations of the speakers by doing loopy BP (LBP) on the factorial structure of the model. This decouples the direct dependencies between the grammar chains of the sources, reducing the complexity of temporal inference, (i.e., inference along the temporal chains of grammar states) to $O(TND_v^2)$ [4]. It does not decouple the dependencies between acoustic states of the sources given the observation. In this paper we use the variational approximation presented in the previous section, which reduces the complexity of the acoustic state to acoustic state messages from $O(D_s^N)$ to $O(IND_s)$, where I is the number of iterations needed to optimize the variational message. The Bayes net for our model (Figure 1(b)) has loops so there is no guarantee of convergence.

A natural approach to approximate inference is to iteratively decode each source given the current estimates of the acoustic state likelihoods of the source, which are influenced by the current 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 [2]. Each source in turn receives messages (i.e., probability estimates) from the other sources. Initially all messages are initialized to be uniform, and $\hat{p}(v_1^k)$ is initialized to the prior for v_1^k for all k. The messages require $\hat{p}(\mathbf{y}|s_t^i)$ (9), which is computed exactly for MSP, and using (15) in Box 2, for Variational MSP (VMSP). For a given source i the incoming messages are computed in three steps:

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

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

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

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

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

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

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

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.

7. Experiments

Tables 1 and 2 summarize the error rate and complexity of our multi-talker speech recognition system on the SSC task [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 [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

$$\log \hat{p}(\mathbf{y}|s^i) = \sum_f \Big[q(i|f,s^i) \log p_{\mathbf{x}_f^i}(y_f|s^i) + (1-q(i|f,s^i)) \log \Phi_{\mathbf{x}_f^i}(y_f|s^i) + \sum_{j \neq i} \big[q(j|f,s^i) E_{q(s^j)} \log p_{\mathbf{x}_j^k}(y_f|s^j) + (1-q(j|f,s^i)) E_{q(s^j)} \log \Phi_{\mathbf{x}_f^j}(y_f|s^j) \big] - \sum_k q(k|f,s^i) \log q(k|f,s^i) \Big] \quad (15)$$

$$\log q(s^i) = c_1 + \log p(s^i) + \log \hat{p}(\mathbf{y}|s^i) \quad (16)$$

$$\log q(s^j) = c_2 + \log p(s^j) + \sum_f \Big[q(j|f) \log p_{\mathbf{x}_f^j}(y_f|s^j) + (1-q(j|f)) \log \Phi_{\mathbf{x}_f^j}(y_f|s^j) \Big] \quad \text{for } j \neq i \quad (17)$$

$$\log q(i|f,s^i) = c_3 + \sum_f \Big[\log p_{\mathbf{x}_f^i}(y_f|s^i) + \sum_{j \neq i} E_{q(s^j)} \log \Phi_{\mathbf{x}_f^j}(y_f|s^j) \Big] \quad (18)$$

$$\log q(j|f,s^i) = c_4 + \sum_f \Big[E_{q(s^j)} \log p_{\mathbf{x}_f^j}(y_f|s^j) + \sum_{k \notin \{i,j\}} E_{q(s^k)} \log \Phi_{\mathbf{x}_f^k}(y_f|s^k) + \log \Phi_{\mathbf{x}_f^i}(y_f|s^i) \Big] \quad \text{for } j \neq i, \quad (19)$$

where $q(k|f) = \sum_{s^i} q(s^i) q(k|f,s^i)$, and c_1 through c_4 are log normalization constants.

Box 2: Iterative variational updates for messages sent to source i. Each update increases the lower bound on the likelihood.

Case	Humans	Viterbi	MSP	VMSP $D_s = 256$ (1024)
ST	34.0	36.4	38.6	53.3 (51.3)
SG	19.5	**14.0**	**14.4**	**17.7 (16.3)**
DG	11.9	**10.8**	**10.8**	14.2 (12.2)
Overall	22.3	**21.2**	**22.1**	29.7 (27.8)

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 are bolded. In all cases oracle speaker identities and gains were used.

Algorithm	Viterbi	Viterbi	MSP	VMSP
Beam size	20000	400	Full	Full
Error Rate	21.2	22.2	22.1	29.7
Temporal	$O(BD_v^N)$	$O(BD_v^N)$	$O(ND_v^2)$	$O(ND_v^2)$
Acoustic	$O(D_s^N)$	$O(D_s^N)$	$O(D_s^N)$	$O(N^2 D_s)$

Table 2: Task error rate (for D_s =256, D_v = 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.

algorithm described in [2], were utilized by VMSP. The gains of the speakers were further optimized w.r.t. the variational bound each time a source likelihood was computed. VMSP scores 3.5% absolute better than the next-best linear-time algorithm. Table 4 summarizes the performance of VMSP as a function of number of sources on a small dataset derived from utterances extracted from the SSC test data. The results demonstrate that the algorithm can separate more than two sources using a single channel. The results are exciting because inference scales *linearly* with both language *and* acoustic model size, making the algorithm applicable to much more complex problems.

8. References

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

[2] 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.

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

Author	Hershey †	Rennie †	Virtanen	Barker †
Method	Viterbi	VMSP	Iterative Viterbi	Fragment Decoding
Inference	Exponential	Linear		
TER	21.6	30.7	34.2	35.2

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 $D_s = 256$ and estimated speaker identities and gains [2] were used. For references consult [1]. † denotes "et al."

Number of Speakers	Target Speaker (F)	Masker			Overall
		1 (M)	2 (F)	3 (M)	
2	27	17	-	-	22
3	40	28	37	-	35
4	47	58	51	51	51

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. Demixed utterances from the SSC test set were mixed directly on top of each another to construct the mixtures. Here $D_s = 1024$.

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

[5] 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.

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

[7] 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.

[8] F. Kschischang, B. Frey, and H. Loeliger, "Factor graphs and the sum-product algorithm," *IEEE Trans. on Info. Theory*, vol. 47, no. 2, pp. 498–519, 2001.

[9] Y. Weiss and W. Freeman, "On the optimality of solutions of the max-product belief-propagation algorithm in arbitrary graphs," *IEEE Trans.on Info. Theory*, vol. 47, no. 2, pp. 736–744, 2001.