USING DYNAMIC CONDITIONAL RANDOM FIELD ON SINGLE-MICROPHONE SPEECH SEPARATION

Yu Ting Yeung, Tan Lee

Department of Electronic Engineering
The Chinese University of Hong Kong
Hong Kong SAR, China
{ytyeung,tanlee}@ee.cuhk.edu.hk

Cheung-Chi Leung

Institute for Infocomm Research
A*STAR
Singapore
ccleung@i2r.a-star.edu.sg

ABSTRACT

The use of dynamic conditional random field (DCRF) for model-based single-microphone speech separation is investigated. The speech sources are represented by acoustic state sequences from speaker-dependent acoustic models. The posterior probabilities of the source acoustic states given a speech mixture are inferred with a maximum entropy probability distribution which is represented by DCRF. The posterior probabilities are needed for minimum mean-square error estimation of the speech sources. Loopy belief propagation is applied for the inference. Averaged stochastic gradient descent and limited-memory BFGS are compared for parameter estimation. With the log-magnitude spectrum of the speech mixture as input observation, the proposed method achieves better separation performance in terms of Blind Source Separation Metrics (SDR, SAR, SIR) and PESQ than a factorial hidden Markov model baseline system in our experiments.

Index Terms— single-microphone speech separation, dynamic conditional random field

1. INTRODUCTION

Single-microphone speech separation is a problem of reconstructing two or more speech sources from only one speech mixture. The problem has many potential applications in speech processing, for example, robust speech recognition in adverse environments and audio information retrieval from live recordings. It is a challenging problem that represents a special case of speech enhancement under non-stationary interference. It is also an extreme case of underdetermined source separation, which is unlikely to have a unique source reconstruction.

Statistical model-based approach is among many approaches to the problem [1]. Its aim is to estimate speech sources given a speech mixture and the acoustic models of the corresponding sources as prior information. When the speech sources are modeled as being generated from the states in the acoustic models, the posterior probabilities of source states given the observations of the speech mixture are computed accordingly, e.g., by factorial hidden Markov model (HMM) [2][3]. Factorial HMM models the generation process of the speech mixture from the speech sources. With the likelihood of the speech mixture given the source states and the prior distribution of the source states from the acoustic models, the posterior probabilities of the underlying source states are computed by Bayes' Theorem. The speech sources are then reconstructed by approaches such as minimum mean-square error (MMSE) estimation [4].

The factorial HMM method requires the likelihood of the speech mixture given the source states. The likelihood can be derived from an interaction model, which is the probability distribution of the observations of the speech mixture given the sources. The exact interaction model is computationally intractable [5]. Approximations such as the MIXMAX model are required [6], and the accuracy of the approximations can significantly affect the separation performance [7]. It is also possible to infer the likelihood from training data, but conditional independence on observations is usually assumed to maintain computational feasibility. This restricts the integration of different types of observations for inference.

Conditional random field (CRF) is a Markov random field (MRF) conditioned on observations [8]. A method based on linear-chain CRFs was proposed in [9] for single-microphone speech separation. Different types of observations were integrated in linear-chain CRFs and maximum *a posteriori* (MAP) estimation of the most probable source acoustic state sequences was performed. This method requires the initial separation results from factorial HMM to achieve the improved separation performance. Dynamic conditional random field (DCRF) is a generalization of CRF on an arbitrary undirected graphical structure [10]. In this paper, we propose to apply DCRF instead of factorial HMM to compute the required posterior probabilities of the source states for MMSE estimation of speech sources. The proposed method does not rely on the initial separation results from factorial HMM for improving the separation performance. As a generalization of CRF, different types of observations can be integrated into the graphical model without assuming conditional independence. Loopy belief propagation (LBP) [11] is applied for parameter estimation and computing the posterior probabilities in DCRF. Due to the non-convexity of LBP approximation, we have also evaluated the separation performance with parameters estimated by two numerical optimization algorithms, namely limited-memory Broyden–Fletcher–Goldfarb–Shanno (BFGS) method [12] and averaged stochastic gradient descent (ASGD) [13].

The paper is organized as follows. Section 2 provides the background of MMSE source estimation in single-microphone speech separation. The formulation and the inference of DCRF are presented in Section 3. The experimental setup is presented in Section 4. The speech sources are reconstructed by MMSE estimation. The separation results are discussed in Section 5. The paper is concluded in Section 6.

This research is partially supported by the Project Grant from the Shun Hing Institute of Advanced Engineering, the Chinese University of Hong Kong, the General Research Fund (Ref: CUHK 414010 and 413811) from the Hong Kong Research Grants Council.

978-1-4799-0356-6/13/$31.00 ©2013 IEEE

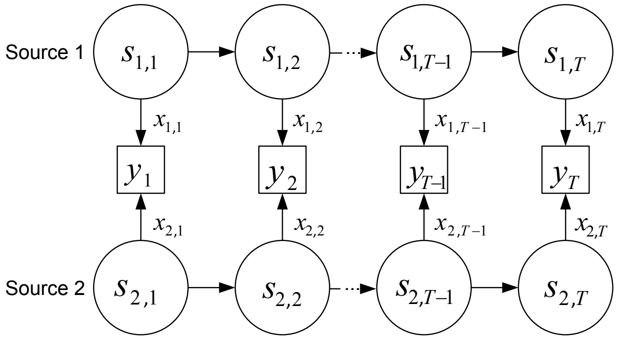

Fig. 1. An example of factorial HMM for single-microphone speech separation with two sources. $x_{m,t}$ are the observations of the speech sources which are hidden variables in the formulation.

2. MMSE ESTIMATION OF SPEECH SOURCES

Let us consider a single-microphone speech separation problem with M sources. The speech mixture is modeled as instantaneous addition of speech sources in time domain. Throughout this paper, we denote $m \in \{1, 2, \ldots, M\}$ as an index of the sources, and use bold font-face to indicate a frame sequence. After short-time analysis of the speech signals, we obtain $\mathbf{y} =(y_1, y_2, \ldots, y_T)$ and $\mathbf{x_m} =(x_{m,1}, x_{m,2}, \ldots, x_{m,T})$ as the feature sequences of length T for the observed speech mixture and the m^{th} speech source, respectively. We also denote $\mathbf{s_m} =(s_{m,1}, s_{m,2}, \ldots, s_{m,T})$ as the acoustic state sequence of source m, which generates $\mathbf{x_m}$. There is no unique solution for source reconstruction. A reasonable estimation of source $\mathbf{x_m}$ is the MMSE estimator given the mixture $\mathbf{y}$. The MMSE estimator of each feature dimension d of $x_{m,t}$ at time index t, i.e., $\hat{x}^d_{m,t} = \mathbb{E}(x^d_{m,t}|\mathbf{y})$, is expressed as

$$\mathbb{E}(x^d_{m,t}|\mathbf{y}) = \sum_{\{s_{m,(t)}\}} p(\{s_{m,(t)}\}|\mathbf{y})\mathbb{E}(x^d_{m,t}|\mathbf{y},\{s_{m,(t)}\}) \ , \quad (1)$$

where $\{s_{m,(t)}\} = \{s_{1,t}, s_{2,t}, \ldots, s_{M,t}\}$ is a set of acoustic states for all M speech sources $\{x_{m,(t)}\} = \{x_{1,t}, x_{2,t}, \ldots, x_{M,t}\}$ at time index t. The index t is parenthesized in the set notations to indicate that it is a constant index. The expectation $\mathbb{E}(x^d_{m,t}|\mathbf{y},\{s_{m,(t)}\})$ is the MMSE estimator of $x_{m,t}$ at each feature dimension d, given the acoustic states $\{s_{m,(t)}\}$. The acoustic states are used to derive the parameters for statistical filtering.

The posterior probability $p(\{s_{m,(t)}\}|\mathbf{y})$ of the source states given the speech mixture can be considered as a weight on the filter output $\mathbb{E}(x^d_{m,t}|\mathbf{y},\{s_{m,(t)}\})$. Let $\{\mathbf{s_m}\} = \{\mathbf{s_1}, \mathbf{s_2}, \ldots, \mathbf{s_M}\}$ be a set of acoustic state sequences of all M speech sources, $p(\{s_{m,(t)}\}|\mathbf{y})$ can be computed by marginalizing the joint density $p(\{\mathbf{s_m}\}, \mathbf{y})$ and conditioning the probabilities on $\mathbf{y}$. In a generative modeling approach such as factorial HMM, $p(\{\mathbf{s_m}\}, \mathbf{y})$ is expressed as

$$p(\{\mathbf{s_m}\}, \mathbf{y}) = \prod_t p(y_t|\{s_{m,(t)}\}) \times \prod_m \prod_t p(s_{m,t}|s_{m,t-1}) \ , \quad (2)$$

where $p(s_{m,1}|s_{m,0}) = p(s_{m,1})$ is the prior probability of the given state. A graphical model illustration of factorial HMM for single-microphone speech separation with two sources is given as in Figure 1. The likelihood $p(y_t|\{s_{m,(t)}\})$ is derived from an interaction model $p(y_t|\{x_{m,(t)}\})$, such as the MIXMAX model which assumes $y^d_t \approx \max(\{x^d_{m,(t)}\})$ in log-spectral domain.

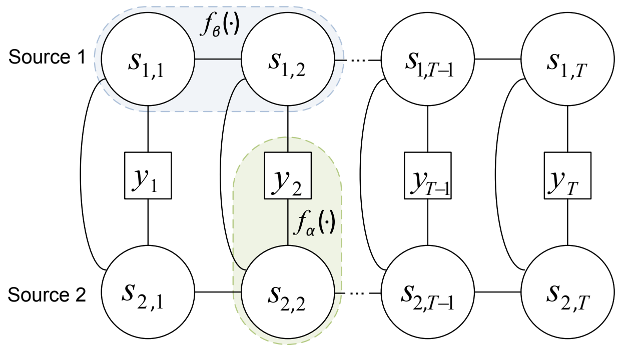

Fig. 2. An example of DCRF for single-microphone speech separation with two sources. DCRF is defined according to an undirected graphical model. Examples of a state feature function $f_\alpha(\cdot)$ and an edge feature function $f_\beta(\cdot)$ are highlighted.

3. DYNAMIC CONDITIONAL RANDOM FIELD FOR SPEECH SEPARATION

3.1. Formulation of dynamic conditional random field

An alternative approach to obtaining $p(\{s_{m,(t)}\}|\mathbf{y})$ for MMSE estimation is to infer $p(\{\mathbf{s_m}\}|\mathbf{y})$ directly from training data. For single-microphone speech separation, the training data consist of mixture $\mathbf{y}$ and the corresponding source state sequences $\{\mathbf{s_m}\}$. The difficulty of directly modeling $p(\{\mathbf{s_m}\}|\mathbf{y})$ is on the determination of a suitable distribution. The distribution must be consistent to the statistics associated with $\mathbf{y}$ and $\{\mathbf{s_m}\}$ in the training data. This problem can be formulated as an entropy maximization problem [14]. The feasible solution of $p(\{\mathbf{s_m}\}|\mathbf{y})$ is known as the maximum entropy probability distribution [15],

$$p(\{\mathbf{s_m}\}|\mathbf{y}) = \frac{\exp \sum_t \sum_k \lambda_k f_k(\{\mathbf{s_m}\}, \mathbf{y},t)}{Z(\mathbf{y})} \quad (3)$$

where $Z(\mathbf{y}) = \sum_{\{\mathbf{s_m}\}} \exp \sum_t \sum_k \lambda_k f_k(\{\mathbf{s_m}\}, \mathbf{y},t)$ is known as partition function. The function $f_k(\{\mathbf{s_m}\}, \mathbf{y},t)$ is the k^{th} feature function or sufficient statistic associated with $\mathbf{y}$ and $\{\mathbf{s_m}\}$. λ_k is the corresponding canonical parameter. The conditional probability $p(\{\mathbf{s_m}\}|\mathbf{y})$ follows a log-linear model. Conditional independence is not assumed among the feature functions. Let $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ be an undirected graph with a vertex set $\mathcal{V}$ and an edge set $\mathcal{E}$. By defining a set of feature functions $\{f_k\} = \{f_\alpha\} \cup \{f_\beta\}$, where $f_\alpha(\cdot)$ is a state feature function and $f_\beta(\cdot)$ is an edge feature function, and $\{\lambda_k\} = \{\lambda_\alpha\} \cup \{\lambda_\beta\}$ as the corresponding canonical parameters, Equation 3 can be rewritten as

$$\begin{aligned} p(\{\mathbf{s_m}\}|\mathbf{y}) = & \frac{1}{Z(\mathbf{y})} \exp\left(\sum_m \sum_t \sum_\alpha \lambda_\alpha f_\alpha(s_{m,t}, y_t)\right) \\ & \times \exp\left(\sum_{(a,b)\in\mathcal{E}} \sum_\beta \lambda_\beta f_\beta(s_a, s_b)\right) , \end{aligned} \quad (4)$$

where $s_{m,t}$ is a source state variable associated with node (m, t) which represents a frame at time index t of source m. The state variables s_a and s_b are associated with nodes a and b which are connected by an edge (a, b) in the edge set $\mathcal{E}$. In this study, $f_\alpha(\cdot)$

and $f_\beta(\cdot)$ are defined as,

$$f_\alpha(s_{m,t}, y_t) = g(y_t, d) \tag{5}$$

$$f_\beta(s_a, s_b) = \begin{cases} 1 & \text{, if } s_a = i \text{ and } s_b = j \\ 0 & \text{, otherwise} \end{cases} \tag{6}$$

where i and j denote the specific states of the corresponding acoustic models. Two types of state feature functions, $g_1(y_t, d) = y_t^d$ and $g_2(y_t, d) = (y_t^d)^2$ are defined. This choice corresponds to the sufficient statistics for the first and second moments of feature dimension d of an observed speech mixture y_t and the source state at $s_{m,t}$. An edge feature function corresponds to a count of a state pair connected by edge (a, b).

The undirected graphical model defined by Equation 4 can be represented with a dynamic conditional random field (DCRF) [10]. Figure 2 illustrates the DCRF applied in this study for single-microphone separation with two sources. This DCRF is a special case. In terms of graphical structure, it is a moral graph of the factorial HMM shown as in Figure 1 [16]. This indicates that $p(\{\mathbf{s_m}\}|\mathbf{y})$ from both formulations should be the same in an ideal case given the exact parameters, the same type of observation and an exact inference algorithm. Since the conditional independence of $\{s_{m,(t)}\}$ given y_t is generally invalid, there are edges connecting the nodes of different sources at the same time instant to model the dependence. Note that more arbitrary graphical structures can also be defined from Equation 4, but we do not investigate them in this study.

3.2. Parameter estimation and inference

The canonical parameters can be estimated by minimizing the negative conditional log-likelihood on R training samples, i.e.,

$$\begin{aligned} \mathcal{L}(\lambda) = -\sum_{r=1}^{R} \Bigg[& \sum_m \sum_t \sum_\alpha \lambda_\alpha f_\alpha(s^r_{m,t}, y^r_{t,}) \\ & + \sum_{(a,b)\in\mathcal{E}} \sum_\beta \lambda_\beta f_\beta(s^r_a, s^r_b) - \log Z(\mathbf{y^r}) \Bigg] + c||\lambda||_2^2 \end{aligned} \tag{7}$$

where c is a regularization factor and λ is the vector containing all the canonical parameters. The minimization of $\mathcal{L}$ does not have a closed-form solution due to the regularization term $||\lambda||_2^2$, and thus is performed by numerical optimization techniques such as gradient descent. In a generic graphical model, exact computation of $\log Z(\mathbf{y^r})$ of the r^{th} training sample is a combinatorial problem. The computational complexity is exponential to the number of sources. In gradient descent, $\log Z(\mathbf{y^r})$ and its gradient $\nabla_{\boldsymbol{\lambda}} \log Z(\mathbf{y^r})$ are updated at each iteration. It is more preferable to compute $\log Z(\mathbf{y^r})$ and its gradient approximately to reduce the computation.

Approximating $\log Z(\mathbf{y^r})$ with loopy belief propagation (LBP) has been successful in solving many graphical model problems [11]. LBP is a message-passing algorithm which ignores the loops in the graphical structure and computes the messages as in a tree-structured graphical model. If the LBP algorithm converges, the fixed point is a zero-gradient point of Bethe free energy [17]. If LBP is applied to a tree-structured graphical model, it reduces to forward-backward algorithm and computes $\log Z(\mathbf{y^r})$ exactly. However, for a generic graphical model, the approximated $\log Z(\mathbf{y^r})$ is neither a upper-bound nor a lower-bound of the exact solution. Due to the loss of convexity of the original problem, suitable numerical optimization algorithms should be chosen carefully.

LBP also computes the marginal distributions $\mathcal{B}_a$ and $\mathcal{B}_{ab}$ over the source state variables s_a, s_b at each node a and edge (a, b) respectively. The marginals $\mathcal{B}_a(s_a)$ and $\mathcal{B}_{ab}(s_a, s_b)$ are subject to normalization constraints, i.e., $\sum_{s_a} \mathcal{B}_a(s_a) = 1$ and $\sum_{s_a} \mathcal{B}_{ab}(s_a, s_b) = \mathcal{B}_b(s_b)$. The marginals are essential to approximate the gradient $\nabla_\lambda \log Z(\mathbf{y^r})$ in parameter estimation [15]. They are also needed for computing $p(\{s_{m,(t)}\}|\mathbf{y})$ during speech separation. The pairwise marginal $\mathcal{B}_{ab}(s_a, s_b)$ can be interpreted as the joint probability of source states s_a and s_b, either within the same source with frames a and b adjacent to each other or from different sources but at the same time instant. For a two-source case, let $a = (1, t)$ and $b = (2, t)$ be the nodes representing a frame of source 1 and a frame of source 2 respectively, at the same time index t. The posterior probability $p(s_{1,t}, s_{2,t}|\mathbf{y})$ can be approximated as $\mathcal{B}_{(1,t)(2,t)}(s_{1,t}, s_{2,t})$. When there are $M > 2$ sources, as the marginals computed by LBP are only approximations due to the loops in the graphical structure, $p(\{s_{m,(t)}\}|\mathbf{y})$ is approximated as

$$p(\{s_{m,(t)}\}|\mathbf{y}) \approx \prod_m \mathcal{B}_{(m,t)}(s_{m,t}) \; . \tag{8}$$

4. EXPERIMENTAL SETUP

Experiments on single-microphone speech separation with two sources ($M = 2$) are carried out. Speech materials of 3 male and 3 female speakers are extracted from the GRID Corpus [18]. The speech materials are re-sampled into 16 kHz. The experimental data for each speaker consists of 500 utterances. For each speaker, 450 randomly selected utterances are used as training source utterances. The remaining 50 utterances are treated as the evaluation source utterances. The utterances are mixed into 3 sets of speech mixtures, namely *Male+Male*, *Male+Female* and *Female+Female*, at power ratio of 0 dB. Each set of speech mixtures consists of a training set and an evaluation set. The training set contains around 2000 speech mixtures generated from training source utterances of the speaker pair. The evaluation set consists of another 2500 speech mixtures generated from the evaluation source utterances of each speaker pair. Short-time speech analysis is applied with Hamming window of 32 ms and frame shift of 10 ms.

Speaker-dependent Gaussian mixture models (GMM) with 128 and 512 components are trained from 257-dimensional log-magnitude spectra from the training source utterances. HMM acoustic models with 128 and 512 acoustic states are refined from the speaker-dependent GMMs by ignoring the component weights. The emission probability density of each state is a multivariate Gaussian distribution. Instead of the uniform transition probability densities between acoustic states, the transition probability densities are updated by iteratively decoding the source state sequences from the training source utterances by Viterbi algorithm to improve the separation performance in factorial HMM. This process also provides the source state sequences as the training data of DCRF. Since the feature dimension of the acoustic models is high, usually there is only one acoustic state being dominant at each frame [19].

Factorial HMM with the MIXMAX interaction model is used as the baseline in our experiments. The observed speech mixture y_t is based on 257-dimension log-magnitude spectrum. Sum-product loopy belief propagation (LBP) [20] is applied to infer $p(s_{1,t}, s_{2,t}|\mathbf{y})$ with the aforementioned acoustic models.

In DCRF, the training data are composed of 257-dimension log-magnitude spectrum as the observation of speech mixture y_t, and the corresponding source state sequences for each set of speech mixtures. For a fair comparison between factorial HMM and DCRF

Table 1. Speech separation results of factorial HMM (FHMM) and DCRF in SDR (dB), SAR (dB), SIR (dB) and PESQ. The IDs in the brackets correspond to the speaker ID in the GRID Corpus.

		Male (1) + Male (2)				**Male (17) + Female (18)**				**Female (24)+ Female (25)**				**Overall**			
		SDR	**SAR**	**SIR**	**PESQ**	**SDR**	**SAR**	**SIR**	**PESQ**	**SDR**	**SAR**	**SIR**	**PESQ**	**SDR**	**SAR**	**SIR**	**PESQ**
512	**FHMM**	5.75	8.98	***9.41***	2.33	8.86	11.21	13.76	2.57	8.27	10.88	12.41	2.46	7.63	10.36	11.86	2.45
	DCRF (L-BFGS)	6.19	9.83	9.37	2.41	9.58	11.96	14.22	2.67	8.81	11.57	12.74	2.58	8.19	11.12	12.11	2.56
	DCRF (ASGD)	***6.32***	***10.07***	9.35	***2.48***	***9.78***	***12.19***	***14.28***	***2.72***	***8.99***	***11.74***	***12.91***	***2.63***	***8.36***	***11.33***	***12.18***	***2.61***
128	**FHMM**	5.11	8.34	8.81	2.18	7.92	10.36	13.13	2.42	7.29	9.93	11.55	2.26	6.77	9.54	11.16	2.28
	DCRF (L-BFGS)	5.78	9.29	***9.14***	2.37	8.86	11.23	13.82	2.58	8.11	10.75	12.25	2.43	7.58	10.43	11.74	2.46
	DCRF (ASGD)	***5.87***	***9.49***	9.10	***2.41***	***9.05***	***11.42***	***13.98***	***2.62***	***8.27***	***10.96***	***12.35***	***2.46***	***7.73***	***10.62***	***11.81***	***2.50***

in computing $p(s_{1,t}, s_{2,t}|\mathbf{y})$, we only investigate the log-magnitude spectrum observation of speech mixture as in factorial HMM. We adopt and compare limited-memory BFGS (L-BFGS) [12] and averaged stochastic gradient descent (ASGD) [13] in parameter estimation. L-BFGS is a quasi-Newton method for finding a stationary point of the objective function. The stationary point corresponds to a local extrema. ASGD is an online learning algorithm. It has been found successful in parameter estimation of several statistical models, including CRF [21]. It updates the parameters with each training sample, which effectively makes use of the redundancies in training data. It is also expected that ASGD helps to escape away from some local extrema for better optimal points due to its stochastic nature. In the evaluation of speech separation, $p(s_{1,t}, s_{2,t}|\mathbf{y})$ is also approximated by sum-product LBP.

For both factorial HMM and DCRF methods, MMSE estimation of the sources based on Equation 1 is performed with the same speaker-dependent acoustic models. The expectation $\mathbb{E}(x^d_{m,t}|\mathbf{y}, s_{1,t}, s_{2,t})$ is implemented as soft-mask filtering as proposed in [22]. The time-domain source signals are reconstructed using the phase spectrum of speech mixtures by the overlap-add method.

5. RESULTS AND DISCUSSION

The reconstructed source signals are compared with the reference source signals in the corpus. Blind Source Separation Evaluation Metrics [23] and Perceptual Evaluation of Speech Quality (PESQ) [24] are adopted as the metrics to signal quality. Source-to-distortion ratio (SDR) measures the overall distortion of the output signal. Source-to-artifacts ratio (SAR) measures the artifact introduced by the separation algorithm. Source-to-interferences ratio (SIR) measures the amount of remaining interfering sources. PESQ is an objective metric to predict human perceptual quality. Separation results with 128 and 512 acoustic states are listed in Table 1. The results are averaged over 2500 speech mixtures and two speakers in each evaluation set.

In both cases with 128 and 512 acoustic states, DCRF consistently achieves better overall separation results than factorial HMM. DCRF generally achieves higher SDR, SAR and SIR for the reconstructed speech sources, except for the *Male + Male* set in which factorial HMM achieves slightly higher SIR for 512 acoustic states. The improvement in speech quality is also evidenced by the consistently higher average PESQ for DCRF. The separation results with DCRF trained by ASGD tend to be slightly better than those trained by L-BFGS. This supports our claim that a suitable numerical optimization algorithm is required for better DCRF training.

DCRF tends to introduce fewer artifacts to the reconstructed sources. It is evidenced by the significant improvement of SAR. The overall improvement is nearly 1 dB. Moreover, by utilizing the log-linear model for the feature functions, frequency components of log-magnitude spectrum of a speech mixture are not assumed statistically independent. This probably contributes to the performance improvement.

6. CONCLUSION

The use of dynamic conditional random field (DCRF) for single-microphone speech separation is investigated in this paper. DCRF is applied to replace factorial HMM in computing the posterior probabilities of the source states given a speech mixture. The posterior probabilities are required for MMSE estimation of speech sources. Experimental results show that when compared with factorial HMM with the MIXMAX model baseline, DCRF tends to achieve better signal quality in terms of SAR, SAR, SIR and PESQ for the reconstructed speech sources. The experiments are based on the same speaker-dependent acoustic models, log-magnitude spectrum as the observation of speech mixture, loopy belief propagation for inference and MMSE estimation for reconstruction of speech sources in both DCRF and factorial HMM. We have evaluated the separation performance with parameters estimated by averaged stochastic gradient descent (ASGD) and limited-memory BFGS (L-BFGS). We opt for ASGD for parameter estimation of DCRF, as separation performance is slightly better than that of by L-BFGS. Several aspects of speech separation with DCRF deserve further investigation. They include unsupervised learning of the source state sequences for training data, discovery and integration of more effective observations for better inference and the use of more arbitrary graphical structures for speech separation problem.

Audio samples are available: www.ee.cuhk.edu.hk/~ytyeung/dmmse.htm .

7. REFERENCES

[1] M. Cooke, J. R. Hershey, and S. J. Rennie, “Monaural speech separation and recognition challenge,” *Computer Speech & Language*, vol. 24, no. 1, pp. 1–15, 2010.

[2] S. T. Roweis, “One microphone source separation,” in *Advances in Neural Information Processing Systems*. MIT Press, 2001, pp. 793–799.

[3] S. J. Rennie, J. R. Hershey, and P. A. Olsen, “Single-channel multitalker speech recognition,” *Signal Processing Magazine, IEEE*, vol. 27, no. 6, pp. 66–80, 2010.

[4] Y. Ephraim, “A Bayesian estimation approach for speech enhancement using hidden Markov models,” *Signal Processing, IEEE Transactions on*, vol. 40, no. 4, pp. 725–735, 1992.

[5] J. R. Hershey, P. A. Olsen, and S. J. Rennie, “Signal interaction and the devil function,” in *Eleventh Annual Conference of the International Speech Communication Association, Interspeech 2010*, 2010, pp. 334–337.

[6] A. Nadas, D. Nahamoo, and M. A. Picheny, “Speech recognition using noise-adaptive prototypes,” *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 37, no. 10, pp. 1495–1503, 1989.

[7] S. J. Rennie, J. R. Hershey, and P. A. Olsen, “Variational loopy belief propagation for multi-talker speech recognition,” in *Tenth Annual Conference of the International Speech Communication Association, Interspeech 2009*, 2009, pp. 1331–1334.

[8] J. D. Lafferty, A. McCallum, and F. C. N. Pereira, “Conditional random fields: Probabilistic models for segmenting and labeling sequence data,” in *Proceedings of the Eighteenth International Conference on Machine Learning*, 2001, pp. 282–289.

[9] Y. T. Yeung, T. Lee, and C.-C. Leung, “Integrating multiple observations for model-based single-microphone speech separation with conditional random fields,” in *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*, 2012, pp. 257–260.

[10] C. Sutton, A. Mccallum, and K. Rohanimanesh, “Dynamic conditional random fields : Factorized probabilistic models for labeling and segmenting sequence data,” *Journal of Machine Learning Research*, vol. 8, pp. 693–723, 2007.

[11] K. P. Murphy, Y. Weiss, and M. I. Jordan, “Loopy belief propagation for approximate inference: An empirical study,” in *Proceedings of Uncertainty in AI*, vol. 9, 1999, pp. 467–475.

[12] J. Nocedal, “Updating quasi-Newton matrices with limited storage,” *Mathematics of Computation*, vol. 35, no. 151, pp. 773–782, 1980.

[13] B. T. Polyak and A. B. Juditsky, “Acceleration of stochastic approximation by averaging,” *SIAM Journal on Control and Optimization*, vol. 30, no. 4, pp. 838–855, 1992.

[14] E. T. Jaynes, “On the rationale of maximum-entropy methods,” *Proceedings of the IEEE*, vol. 70, no. 9, pp. 939–952, 1982.

[15] M. J. Wainwright and M. I. Jordan, “Graphical models, exponential families, and variational inference,” *Foundations and Trends in Machine Learning*, vol. 1, no. 1–2, pp. 1–305, 2008.

[16] F. V. Jensen, *An introduction to Bayesian networks*. London: UCL Press, 1996.

[17] J. S. Yedidia, W. T. Freeman, and Y. Weiss, “Generalized belief propagation,” in *Advances in Neural Information Processing Systems*. MIT Press, 2001, pp. 689–695.

[18] M. Cooke, J. Barker, S. Cunningham, and X. Shao, “An audiovisual corpus for speech perception and automatic speech recognition,” *The Journal of the Acoustical Society of America*, vol. 120, no. 5, pp. 2421–2424, 2006.

[19] N. Merhav and Y. Ephraim, “Maximum likelihood hidden Markov modeling using a dominant sequence,” *Signal Processing, IEEE Transactions on*, vol. 39, no. 9, pp. 2111–2115, 1991.

[20] S. J. Rennie, J. R. Hershey, and P. A. Olsen, “Single-channel speech separation and recognition using loopy belief propagation,” in *Acoustics, Speech and Signal Processing (ICASSP), 2009 IEEE International Conference on*, 2009, pp. 3845–3848.

[21] S. V. N. Vishwanathan, N. N. Schraudolph, M. W. Schmidt, and K. P. Murphy, “Accelerated training of conditional random fields with stochastic gradient methods,” in *Proceedings of the 23rd International Conference on Machine Learning*, 2006, pp. 969–976.

[22] M. H. Radfar and R. M. Dansereau, “Single-channel speech separation using soft mask filtering,” *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 15, no. 8, pp. 2299–2310, 2007.

[23] E. Vincent, R. Gribonval, and C. Fevotte, “Performance measurement in blind audio source separation,” *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 14, no. 4, pp. 1462–1469, 2006.

[24] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, “Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs,” in *Acoustics, Speech, and Signal Processing, 2001. Proceedings. (ICASSP ’01). IEEE International Conference on*, vol. 2, 2001, pp. 749–752.