Large-margin Conditional Random Fields for Single-microphone Speech Separation

Yu Ting Yeung¹, Tan Lee¹, Cheung-Chi Leung²

¹Department of Electronic Engineering, The Chinese University of Hong Kong, Hong Kong SAR, China
²Institute for Infocomm Research, A*STAR, Singapore

{ytyeung,tanlee}@ee.cuhk.edu.hk, ccleung@i2r.a-star.edu.sg

Abstract

Conditional random field (CRF) formulations for single-microphone speech separation are improved by large-margin parameter estimation. Speech sources are represented by acoustic state sequences from speaker-dependent acoustic models. The large-margin technique improves the classification accuracy of acoustic states by reducing generalization error in the training phase. Non-linear mappings inspired from the mixture-maximization (MIXMAX) model are applied to speech mixture observations. Compared with a factorial hidden Markov model baseline, the improved CRF formulations achieve better separation performance with significantly fewer training data. The separation performance is evaluated in terms of objective quality measures and speech recognition accuracy on the reconstructed sources. Compared with the CRF formulations without large-margin parameter estimation, the improved formulations achieve better performance without modifying the statistical inference procedures, especially when the sources are modeled with increased number of acoustic states.

Index Terms: single-microphone speech separation, conditional random fields, large-margin parameter estimation

1. Introduction

Single-microphone speech separation attempts to reconstruct at least two sources from only one speech mixture. A unique solution is unlikely to be achieved due to the problem’s extremely under-determined nature. Model-based methods provide an alternative approach to estimating the most probable source observations given statistical information of the sources, typically modeled in hidden Markov model (HMM) based acoustic models. Single-microphone speech separation can be formulated as a problem of classifying underlying source acoustic states, given a speech mixture. The source observations are subsequently reconstructed by statistical filtering process.

The performance of statistical model-based methods has been progressively improved with graphical models such as factorial HMM [1,2]. Factorial HMM is a generative model, which achieves good performance when model mis-specification is insignificant [3,4]. However, model mis-specification is common in real-world data such as speech. A discriminative model like conditional random field (CRF) [5] is less sensitive to model mis-specification [6,7]. Moreover, CRF does not require conditional independence assumption on observations, which facilitates observation integration for further improving classification accuracy. CRF is a promising technique for speech applications [8], including phone recognition [9,10] and single-microphone speech separation [11–13].

A discriminative model like CRF is still prone to over-fitting when the amount of training data is limited. Large-margin parameter estimation refines the decision boundaries such that different classes are separated with maximized distance during the training phase. This improves model generalization [14,15]. Large-margin techniques on Gaussian models were proposed to various pattern recognition tasks [16–19]. A large-margin CRF formulation was also derived in [20]. In this paper, the CRF formulations for single-microphone speech separation [13] are improved with large-margin parameter estimation. Another improvement is the application of non-linear mappings on speech mixture observations. The non-linear mappings are inspired from the mixture-maximization (MIXMAX) model [21,22], which has been extensively used in generative modeling for single-microphone speech separation.

In Section 2 there is a review of generative modeling and minimum mean square error (MMSE) source estimation. Large-margin CRF formulation using non-linear mappings on speech mixture observations is presented in Section 3. The experimental setup and results are presented in Section 4. The separation performance is evaluated in terms of objective quality measures and speech recognition accuracy of the reconstructed sources. The paper is concluded in Section 5.

The notations applied in this paper are defined as follows: $\mathbf{y} = \{y_t\}_{t=1}^T$, $\mathbf{x}_k = \{x_{k,t}\}_{t=1}^T$ and $\mathbf{s}_k = \{s_{k,t}\}_{t=1}^T$ denote the sequences of observations of the speech mixture, observations of source $k$, and acoustic states from the acoustic model of source $k$ respectively. For brevity, the set of acoustic state sequences $\{s_{k,t}\}_{t=1}^T$ is parenthesized. The component $f$ of the observations of source $k$ and the speech mixture are defined as $x_{k,f}$ and $y_{f}$ respectively. A training data set $\mathcal{D} = \{\mathbf{y}^{(r)}, \{s_{k,t}^{(r)}\}_{t=1}^T\}_{r=1}^R$ is also defined, with $R$ instances of speech mixtures and the reference source state sequence sets.

2. Factorial HMM for speech separation

2.1. Minimum mean square error source estimation

Under a soft-decision scheme which minimizes the effect of mis-classification of the source acoustic states, the statistical filtering process is equivalent to minimum mean square error (MMSE) source estimation. The estimation of source $k$ at frame $t$, $\mathbb{E}(x_{k,t}|y)$ is given by

$$
\mathbb{E}(x_{k,t}|y) = \sum_{\{s_{k,t}\}} p(\{s_{k,t}\}|y) \times \mathbb{E}(x_{k,t}|y, \{s_{k,t}\}). \quad (1)
$$
The computation of $\mathbb{E}(x_{t,k} \mid y, \{s_{k,(t)}\})$ and $p(\{s_{k,(t)}\} \mid y)$ is the key problem in speech separation. The generative modeling approach starts from the joint probability $p(s_k, y)$, which can be represented by factorial HMM [23].

$$p(\{s_k\}, y) = \prod_t p(y_t \mid [s_{k,(t)}]) \times \prod_k \prod_t p(s_{k,t} \mid s_{k,t-1}) .$$  \hspace{1cm} (2)

For each source $k$, the state transition probabilities $p(s_{k,t} \mid s_{k,t-1})$ and the prior probabilities $p(s_{k,t} \mid s_{k,0}) = p(s_k)$ are obtained from the acoustic models. The state interaction $p(y_t \mid \{s_{k,(t)}\})$ and the conditional mean $\mathbb{E}(x_{k,t} \mid y, \{s_{k,(t)}\})$ are obtained from an interaction model $p(y_t | x_{1:t}, x_{2:t}, \ldots, x_{K:t})$. Exact interaction model is computationally intractable. The MIXMAX model approximates $y_t = \max(x_{1:t}, x_{2:t}, \ldots, x_{K:t})$ in log-spectrum [21, 22]. At each frequency component, the model assumes that the log-spectrum of speech mixture is dominated by only one source. The MIXMAX model is a non-linear MMSE estimator of speech mixture given the sources [25]. We adopt the model for $\mathbb{E}(x_{k,t} \mid y, \{s_{k,(t)}\})$ throughout this paper.

The state-level joint probability $p(\{s_{k,(t)}\}, y)$ is computed by marginalizing $p(s_k, y)$ with exact or approximated inference algorithms such as junction tree algorithm [26] or loopy belief propagation [27]. The posterior probability $p(\{s_{k,(t)}\} \mid y)$ is obtained from $p(\{s_{k,(t)}\}, y)$ by the Bayes’ theorem.

2.2. Data-driven state interaction

The state interaction $p(y_t \mid \{s_{k,(t)}\})$ can also be obtained by data-driven approach from the training data set $\mathcal{D}$. For instance, by modeling $p(y_t \mid \{s_{k,(t)}\})$ with multivariate Gaussian distribution, the distribution of the speech mixture observations $p(y_t)$ subsequently follows a Gaussian mixture model (GMM),

$$p(y_t) = \sum_{\{s_{k,(t)}\}} p(\{s_{k,(t)}\} \mid y_t) \cdot \pi(\{s_{k,(t)}\}) .$$  \hspace{1cm} (3)

where the prior probabilities $p(\{s_{k,(t)}\}) = \prod_k p(s_k)$ are regarded as the weights on the Gaussian components.

3. Large-margin parameter estimation

3.1. Conditional random fields

In a CRF formulation, $p(\{s_{k,(t)}\} \mid y)$ is obtained from a maximum entropy probability distribution $p(\{s_k\} \mid y)$,

$$p(\{s_k\} \mid y) = \frac{1}{Z(y)} \exp \left( \Phi(\{s_k\} \mid y) \right)$$  \hspace{1cm} (4)

where $Z(y) = \sum_{\{s_k\}} \exp \left( \Phi(\{s_k\} \mid y) \right)$ is the partition function, $\Phi(\{s_k\} \mid y) = \Lambda \cdot F$ is a log-potential function, and $\Lambda$ and $F$ are the vectors of canonical parameters and feature functions respectively. The canonical parameters $\Lambda$ can be estimated by minimizing the conditional negative log-likelihood $\mathcal{L}(\Lambda) = -\sum_t \Phi(\{s_k\} \mid y_t) + \log Z(y_t)$ on the training data set $\mathcal{D}$.

The feature functions are elaborated into state feature functions $f_{a}(\cdot)$ which associate the source states and the observations, and edge feature functions $f_{b}(\cdot)$ which describe co-occurrence of the states. By defining the corresponding canonical parameters $\lambda_{a}$ and $\lambda_{b}$, we have

$$\Phi(\{s_k\} \mid y) = \sum_k \sum_t \sum_{a \in \mathcal{A}} \lambda_{a} f_{a}(\cdot) + \sum_{(a,b) \in \mathcal{E}} \sum_{\beta} \lambda_{b} f_{b}(\cdot) \hspace{1cm} (5)$$

Figure 1: Conditional random fields for single-microphone speech separation with two sources. Examples of a state feature function $f_{a}(\cdot)$, an in-chain edge feature function $f_{b1}(\cdot)$ and a cross-chain edge feature function $f_{b2}(\cdot)$ are highlighted.

According to a graph $(\mathcal{V}, \mathcal{E})$ as shown in Figure 1. This leads to dynamic conditional random fields [28]. Since the conditional independence of $\{s_{k,(t)}\}$ given $y_t$ is generally invalid, there are edges connecting the nodes of different sources for modeling the potential dependency.

3.2. Large-margin parameter estimation

Large-margin parameter estimation aims at finding canonical parameters $\Lambda$ such that $\Phi(\{s_k\} \mid y^r)$ of the correct state sequence set $\{s_k\}^{(r)}$ is greater than $\Phi(\{s_k\} \mid y^r)$ of any incorrect state sequence set $\{s_k\} 
eq \{s_k\}^{(r)}$, by a metric $\mathcal{H}(\{s_k\}^{(r)}, \{s_k\}^{(r)}) = 0$ and $\mathcal{H}(\{s_k\}, \{s_k\}^{(r)}) > 0$ such as Hamming distance [29]. This constraint is written as,

$$\Phi(\{s_k\}^{(r)} \mid y^r) - \max_{\{s_k\}} \left( \Phi(\{s_k\} \mid y^r) + \mathcal{H}(\{s_k\}, \{s_k\}^{(r)}) \right) = 0, \forall r.$$  \hspace{1cm} (6)

Since $\max\{\cdot\}$ is non-differentiable, the differentiable soft-maximum $\log\text{softmax}\{\cdot\}$ is more preferable in gradient descent. Slack variables $\zeta_r \geq 0, \forall r$ are introduced to constrain $\max\{\cdot\} = \log\text{softmax}\{\cdot\} - \zeta$ as $\max\{\cdot\} \leq \log\text{softmax}\{\cdot\}$. We should minimize the slack variables $\zeta$, such that $\log\text{softmax}\{\cdot\}$ is as close to $\max\{\cdot\}$ as possible. A small $\zeta$ also implies that $\Phi(\{s_k\} \mid y^r)$ is much larger than each $\Phi(\{s_k\} \mid y^r) + \mathcal{H}(\{s_k\}, \{s_k\}^{(r)})$ term when (6) is satisfied. When (6) is not satisfied, the slack variables allow the violation of the constraint in a soft-margin formulation of support vector machine (SVM) [14]. The violation should be minimized in hope that $\Phi(\{s_k\} \mid y^r)$ is still greater than $\Phi(\{s_k\} \mid y^r)$. Moreover, $\lambda$ can be scaled to produce arbitrary large margins, a regularization term $c||\Lambda||^2$ is applied to limit the “size” of the canonical parameters [16]. The optimization problem is now formulated as

$$\min_{\Lambda} \sum_{r} \zeta_r + c||\Lambda||^2$$  \hspace{1cm} (7)

s.t. $\Phi(\{s_k\} \mid y^r) - \log\text{softmax}(\{s_k\}) + \mathcal{H}(\{s_k\}, \{s_k\}^{(r)}) = -\zeta_r, \forall r$

$\zeta_r \geq 0, \forall r$.  \hspace{1cm} (7)

The slack variable $\zeta$ is the conditional negative log-likelihood of the correct sequence set $\{s_k\}^{(r)}$ of $r^{th}$ training data. Hence,
(7) is transformed into an unconstrained minimization problem with objective function

\[
\mathcal{L}(\Lambda) = \sum_{s} \left(-\Phi\left(\{s_k\}\right) | y^{(r)} \right) + \log \sum_{\{s_k\}} e^{\Phi\left(\{s_k\}\right)} + \mathcal{H}(\{s_k\}) \right) + c|\Lambda|^{2}.
\]

An identical objective function was derived in [20], but our derivation emphasizes the bounding property of soft-maximum. The posterior probability \(p(\{s_k\}|y^{(r)})\) is derived as

\[
p(\{s_k\}|y^{(r)}) = \frac{\exp\{\Phi(\{s_k\})\} - \mathcal{H}(\{s_k\})}{\sum_{s} e^{\Phi(\{s_k\})}}.
\]

When \(\mathcal{H}(\{s_k\})\) is removed, (9) is equivalent to (4). Since \(\mathcal{H}(\{s_k\})\) is independent of \(\Lambda\), the objective functions of (9) and (4) have the same gradients. There are no other changes in the parameter estimation processes except the slight difference of the objective functions. During speech separation, \(p(\{s_k\}|y)\) is computed from \(p(\{s_k\}|y)\) by (4) as if conventional CRF.

3.3. Feature functions for CRF

3.3.1. State feature functions with non-linear mappings

A state feature function \(f_{s}(\cdot)\) can be defined according to sufficient statistics of the observations, such as first and second moments labeled as \((M1)\) and \((M2)\) respectively.

\[
f^{(M1)}_{x_{a,t},y_{i,t}}(s_{k,t},y_{i,t}) = \delta(s_{k,t} = i)y_{i,t}
\]

\[
f^{(M2)}_{x_{a,t},y_{i,t}}(s_{k,t},y_{i,t}) = \delta(s_{k,t} = i)(y_{i,t})^2
\]

where \(i\) denotes the specific state in the acoustic model, \(\delta(s_{k,t} = i)\) is an indicator function which equals to 1 when the condition \((s_{k,t} = i)\) is satisfied or zero otherwise.

CRF allows a state feature function to be defined with an arbitrary function. Let \(K(u, m, \sigma) = \frac{1}{\sigma \sqrt{2\pi}} e^{-\frac{(u-m)^2}{2\sigma^2}}\) be a Gaussian function and \(G(u, m, \sigma) = \int_{-\infty}^{\infty} K(v, m, \sigma)dv\) be a sigmoid function. When the sources are modeled in log-spectral domain and the emission probability of an acoustic state \(i\) follows multivariate Gaussian distribution with mean \(\mu_{i,f}\) and variance \(\sigma_{i,f}^2\) for each component \(f\), \(K(y_{i,f}, \mu_{i,f}, \sigma_{i,f})\) and \(G(y_{i,f}, \mu_{i,f}, \sigma_{i,f})\) are equivalent to \(p(x_{a,t,f} = y_{i,f}|s_{k,t} = i)\) and \(p(x_{a,t,f} < y_{i,f}|s_{k,t} = i)\) respectively which carry significant physical meanings according to the MIXMAX model [21, 22].

Given the acoustic state \(i\) of source \(k\), \(p(x_{a,t,f} = y_{i,f}|s_{k,t} = i)\) is the probability that source \(k\) dominates the speech mixture, \(p(x_{a,t,f} < y_{i,f}|s_{k,t} = i)\) is the probability that the source is masked by other sources. They can be incorporated into \(f_{s}(\cdot)\) as non-linear mappings on speech mixture observations,

\[
f^{(K)}_{x_{a,t},y_{i,t}}(s_{k,t},y_{i,t}) = \delta(s_{k,t} = i)K(y_{i,f}, \mu_{i,f}, \sigma_{i,f})
\]

\[
f^{(G)}_{x_{a,t},y_{i,t}}(s_{k,t},y_{i,t}) = \delta(s_{k,t} = i)G(y_{i,f}, \mu_{i,f}, \sigma_{i,f}).
\]

The non-linear mappings transform the speech mixture observations from real space \(\mathbb{R}\) to probability space \([0,1]\). We consider the application of these mappings as the revival of the idea of integrating initial separation results from factorial HMM in CRF formulations [12].

### Table 1: The settings of CRF formulations for single-speech source

<table>
<thead>
<tr>
<th>CRF</th>
<th>Non-linear mappings</th>
<th>Observations</th>
<th>Large-margin</th>
</tr>
</thead>
<tbody>
<tr>
<td>CRFMOMENT</td>
<td>No, (M1, M2)</td>
<td>LOG</td>
<td>No</td>
</tr>
<tr>
<td>CRFNLMP</td>
<td>Yes, (K, G)</td>
<td>LOG</td>
<td>No</td>
</tr>
<tr>
<td>LM-CRFNLMP</td>
<td>Yes, (K, G)</td>
<td>LOG</td>
<td>Yes</td>
</tr>
<tr>
<td>LM-CRFFC</td>
<td>Yes, (K, G)</td>
<td>LOG, MFCC</td>
<td>Yes</td>
</tr>
</tbody>
</table>

3.3.2. Edge feature functions

A count-based indicator is defined to collect the statistics of the state pair connected by an edge \((a, b)\),

\[f_{\delta}(s_{a}, s_{b}) = \delta(s_{a} = i)\delta(s_{b} = j)\]

where \(i, j\) denote the corresponding acoustic states. When a state pair \((s_{a}, s_{b})\) is within the same source with frames \(a\) and \(b\) adjacent to each other, the in-chain edge feature function models the transition between the states along the time axis for the temporal inference. A cross-chain edge feature function corresponds to the occurrence of a state pair from different sources but at the same time instant, as appears in Figure 1.

### 4. Experiments

#### 4.1. Experimental setup

Two tasks are designed for performance evaluation. The first task is to reconstruct individual speech sources from a speech mixture of two different speakers. The log-spectrum of the sources are recovered by MMSE estimation. The waveforms are reconstructed using the mixture phase spectrum by the overlap-add method. The second task is speech recognition on reconstructed speech sources. Speech data of 3 male and 3 female speakers are extracted from the GRID Corpus [30]. Three types of speech mixtures, namely Male+Male, Male+Female and Female+Female, are prepared at power ratio of 0 dB. Short-time speech analysis with Hamming window of 32 ms and frame shift of 10 ms is applied. For each speaker, 450 clean utterances are used as training data, and 50 unseen utterances are used for evaluation. Each set of mixture data consists of 2,500 speech mixtures for evaluation, and over 200,000 speech mixtures are available for model training.

For the speech separation task, speaker-dependent HMM acoustic models with 128 and 512 states are developed from clean training utterances. The acoustic models are based on 257-dimensional log-magnitude spectrum (from 512-point fast Fourier transform), with a multivariate Gaussian distribution emission probability for each state. Due to the high feature dimension, each frame can be well-characterized by single multivariate Gaussian distribution [31]. The state sequences of clean training utterances are obtained during the training process. These state sequences will be used for CRF parameter estimation and data-driven approach in factorial HMM.

The configurations of CRF formulations are listed in Table 1. CRF with non-linear mappings (CRFNLMP) serves as the CRF baseline for large-margin formulation (LM-CRFNLMP). As shown in our experimental results, CRFNLMP performs significantly better than CRFMOMENT which uses sufficient statistics of speech mixture observations as state feature functions. To further evaluate the benefits of
observation integration, 39-dimensional Mel-frequency cepstral coefficients (MFCC, 12 coefficients + log-energy + \( \Delta \) and \( \Delta \Delta \) coefficients) of speech mixtures are included in addition to 257-dimensional log-magnitude observations for LM-CRF\textsuperscript{FMCC}. The mean and the diagonal covariance of MFCC vectors for each state are obtained for performing non-linear mappings. Only 2,000 training speech mixtures for each speaker pair are applied for CRF parameter estimation. The parameters are updated with averaged stochastic gradient descent [32]. Loopy belief propagation [33] is applied to approximate the gradients and the log-partition function for parameter estimation, and the required posterior probabilities during speech separation. We refer readers to [34] for the details of statistical inference.

Factorial HMM by data-driven approach (FHMM\textsuperscript{DATA}) is considered as a better baseline for CRF formulations since both involve training speech mixtures for parameter estimation. The entire 200,000\textsuperscript{1} training mixtures of each speaker pair are applied for modeling the \textit{state interaction} as described in section 2.2. Factorial HMM experiments with the MIXMAX model are also performed (FHMM\textsuperscript{MIXMAX}). Junction tree algorithm is applied for the required posterior probabilities during speech separation. This setup favors factorial HMM with 100 times more training data and an exact statistical inference algorithm.

The speech recognition system is developed with HTK [35]. Word-based GMM-HMM speaker-independent acoustic models are trained with clean speech using standard MFCC feature vectors. Each word model consists of 4 to 8 states with 32 Gaussian components per state. A grammar network specifying the command recognition task of GRID Corpus is applied for speech recognition. The word error rate for clean speech is less than 1%.

### 4.2. Results and discussion

The overall separation results and the breakdown results are listed in Table 2 and 3 respectively. Perceptual evaluation

### Table 2: Overall speech separation results in PESQ, SDR (dB) (higher is better) and WER (%) (lower is better)

<table>
<thead>
<tr>
<th></th>
<th>PESQ</th>
<th>SDR</th>
<th>WER</th>
<th>PESQ</th>
<th>SDR</th>
<th>WER</th>
</tr>
</thead>
<tbody>
<tr>
<td>FHMM\textsuperscript{MIXMAX}</td>
<td>2.37</td>
<td>7.56</td>
<td>7.50</td>
<td>2.53</td>
<td>8.29</td>
<td>5.14</td>
</tr>
<tr>
<td>FHMM\textsuperscript{DATA}</td>
<td>2.45</td>
<td>7.76</td>
<td>6.82</td>
<td>2.61</td>
<td>8.48</td>
<td>4.80</td>
</tr>
<tr>
<td>CRF\textsuperscript{MOMENT}</td>
<td>2.54</td>
<td>8.26</td>
<td>7.29</td>
<td>2.64</td>
<td>8.77</td>
<td>6.09</td>
</tr>
<tr>
<td>CRF\textsuperscript{NLMAP}</td>
<td>2.59</td>
<td>8.27</td>
<td>5.46</td>
<td>2.68</td>
<td>8.69</td>
<td>5.09</td>
</tr>
<tr>
<td>LM-CRF\textsuperscript{NLMAP}</td>
<td>2.60</td>
<td>8.31</td>
<td>5.38</td>
<td>2.70</td>
<td>8.78</td>
<td>4.59</td>
</tr>
<tr>
<td>LM-CRF\textsuperscript{FMCC}</td>
<td>2.61</td>
<td>8.37</td>
<td>5.12</td>
<td>2.73</td>
<td>8.83</td>
<td>4.48</td>
</tr>
</tbody>
</table>

\(^1\)We have tried to train FHMM\textsuperscript{DATA} with only 2000 training speech mixtures, but it is a failure with significantly lower PESQ and double of WER than FHMM\textsuperscript{MIXMAX} in the case with 512 acoustic states.

### Table 3: The breakdown results of Table 2. The IDs in the brackets correspond to the speaker ID in the GRID Corpus.

<table>
<thead>
<tr>
<th></th>
<th>PESQ</th>
<th>SDR</th>
<th>WER</th>
<th>PESQ</th>
<th>SDR</th>
<th>WER</th>
</tr>
</thead>
<tbody>
<tr>
<td>FHMM\textsuperscript{MIXMAX}</td>
<td>2.39</td>
<td>6.20</td>
<td>7.88</td>
<td>2.65</td>
<td>9.75</td>
<td>3.74</td>
</tr>
<tr>
<td>FHMM\textsuperscript{DATA}</td>
<td>2.50</td>
<td>6.32</td>
<td>7.30</td>
<td>2.70</td>
<td>9.94</td>
<td>3.41</td>
</tr>
<tr>
<td>CRF\textsuperscript{MOMENT}</td>
<td>2.49</td>
<td>6.52</td>
<td>10.23</td>
<td>2.76</td>
<td>10.33</td>
<td>3.51</td>
</tr>
<tr>
<td>CRF\textsuperscript{NLMAP}</td>
<td>2.54</td>
<td>6.49</td>
<td>8.17</td>
<td>2.80</td>
<td>10.18</td>
<td>3.32</td>
</tr>
<tr>
<td>LM-CRF\textsuperscript{NLMAP}</td>
<td>2.57</td>
<td>6.58</td>
<td>7.38</td>
<td>2.82</td>
<td>10.25</td>
<td>3.14</td>
</tr>
<tr>
<td>LM-CRF\textsuperscript{FMCC}</td>
<td>2.59</td>
<td>6.63</td>
<td>7.35</td>
<td>2.86</td>
<td>10.29</td>
<td>2.90</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>PESQ</th>
<th>SDR</th>
<th>WER</th>
<th>PESQ</th>
<th>SDR</th>
<th>WER</th>
</tr>
</thead>
<tbody>
<tr>
<td>FHMM\textsuperscript{MIXMAX}</td>
<td>2.24</td>
<td>5.64</td>
<td>11.18</td>
<td>2.52</td>
<td>9.03</td>
<td>5.05</td>
</tr>
<tr>
<td>FHMM\textsuperscript{DATA}</td>
<td>2.34</td>
<td>5.76</td>
<td>10.29</td>
<td>2.57</td>
<td>9.21</td>
<td>4.70</td>
</tr>
<tr>
<td>CRF\textsuperscript{MOMENT}</td>
<td>2.43</td>
<td>6.21</td>
<td>11.34</td>
<td>2.67</td>
<td>9.77</td>
<td>4.41</td>
</tr>
<tr>
<td>CRF\textsuperscript{NLMAP}</td>
<td>2.51</td>
<td>6.29</td>
<td>8.10</td>
<td>2.71</td>
<td>9.69</td>
<td>3.91</td>
</tr>
<tr>
<td>LM-CRF\textsuperscript{NLMAP}</td>
<td>2.52</td>
<td>6.33</td>
<td>7.88</td>
<td>2.72</td>
<td>9.74</td>
<td>3.89</td>
</tr>
<tr>
<td>LM-CRF\textsuperscript{FMCC}</td>
<td>2.55</td>
<td>6.39</td>
<td>7.66</td>
<td>2.73</td>
<td>9.76</td>
<td>3.83</td>
</tr>
</tbody>
</table>

Speech recognition performance is measured in word error rate (WER). Large-margin parameter estimation has brought substantial WER reduction in nearly all test cases, especially for the cases with 512 states. LM-CRF\textsuperscript{NLMAP} outperforms factorial HMM baseline FHMM\textsuperscript{DATA} with 100 times fewer training speech mixtures. We attribute the improvement to the ability of large-margin technique for avoiding over-fitting. Without large-margin technique, over-fitting is observed in \textit{Male + Male} set. WER with 512 states is higher than that of with 128 states in CRF\textsuperscript{NLMAP} as shown in Table 3. The situation is rectified with large-margin technique in LM-CRF\textsuperscript{NLMAP}. By further integrating MFCC observations, the best results are now dominated by LM-CRF\textsuperscript{FMCC}.

The results also show that applying non-linear mappings leads to significant WER reduction in speech recognition task. Without non-linear mappings, CRF\textsuperscript{MOMENT} obtains lower speech recognition accuracy than that of FHMM\textsuperscript{DATA}. After applying non-linear mappings, CRF\textsuperscript{NLMAP} achieves over 16 \% relative WER reduction from CRF\textsuperscript{MOMENT}, although it does not outperform FHMM\textsuperscript{DATA} with 512 states until large-margin technique is applied. The results reveal that the improved performance in terms of objective quality measures does not necessarily reflect the improvement of speech recognition accuracy.

The results also agree with the argument that the performance of a generative model is comparable to a discriminative model when the model mis-specification becomes insignificant [3,7]. The data-driven approach for \textit{state interaction} with a sufficient amount of training speech mixtures, and the increased number of acoustic states for modeling the sources have minimized model mis-specification in factorial HMM formulations. The minimized model mis-specification leads to a reduction of performance gap between FHMM\textsuperscript{DATA} and LM-CRF\textsuperscript{NLMAP}.

### 5. Conclusions

Large-margin conditional random fields (CRF) with non-linear mappings on speech mixture observations are investigated for single-microphone speech separation. Experimental results show that the improved CRF formulations achieve competitive performance to factorial HMM with significantly fewer training data. The large-margin technique is especially effective with the increased number of acoustic states, where over-fitting may occur with conventional discriminative modeling approaches. Other applications of non-linear mappings also deserve further investigation. As the non-linear mappings are related to the MIXMAX model, they may be useful to adapt the CRF formulations to speech mixtures with power ratios different from the training condition.

### 6. Acknowledgements

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 410410 and 413811) from the Hong Kong Research Grants Council.

Audio samples: [http://www.ee.cuhk.edu.hk/~ytyeung/is2014.htm](http://www.ee.cuhk.edu.hk/~ytyeung/is2014.htm)
7. References


