Minimum Classification Error Training of Hidden Conditional Random Fields for Speech and Speaker Recognition

WEI-TYNG HONG

Department of Communications Engineering
Yuan Ze University
Chungli, 320 Taiwan

Hidden conditional random fields (HCRFs) are derived from the theory of conditional random fields with hidden-state probabilistic framework. It directly models the conditional probability of a label sequence given observations. Compared to hidden Markov models, HCRFs provide a number of benefits in the acoustic modeling of speech signals. Prior works for training on HCRFs were accomplished with gradient descent based algorithms by conditional maximum likelihood criterion. In this paper, we extend that methodology by applying minimum classification error criterion-based training technique on HCRFs. Specifically, we adopt generalized probabilistic descent (GPD)-based training algorithm with HCRF framework to improve the discrimination capabilities of acoustic models for speech and speaker recognition. Two tasks including a speaker identification and a Mandarin continuous syllable recognition are applied to evaluate the proposed approach. We present the results on the MAT2000 database and these results confirm that the HCRF/GPD approach has good capabilities for speech recognition and speaker identification regardless of the length of the test and training speech or the presence of noise. We note that the HCRF/GPD enjoys its potential for development in acoustic modeling.

Keywords: speech recognition, speaker recognition, hidden conditional random field, discriminative training algorithm, Mandarin syllable recognition

1. INTRODUCTION

Automatic speaker identification and speech recognition, the tasks of recognizing the speaker identity and determining what was said for the input speech using trained acoustic models, have been two interesting research fields for many years. The problem of speaker identification and speech recognition can be treated as sequence classification problems in which the most important issue is how to build efficiently an accurate acoustic models by training speech based on commonly used speech features (e.g., cepstral coefficients). In the past, many methods have been proposed for acoustic modeling. These approaches can be divided into three categories: the neural network (NN)-based approach, the statistical approach, and the hybrid based approach. The NN-based approach for identification, for example [1, 2], uses neural networks as pattern classifiers to discriminate each input speech among speaker classes. There are two main advantages to this approach. One is that an NN classifier can be trained directly from a training data set without making strong assumptions about the distributions of candidate classes. The other is that an NN classifier can possess high discrimination capability obtained via competitive training. However, a large amount of training data is usually required in the NN-based approach. Another issue compared with the statistical approach is that the NN

Received February 3, 2012; accepted June 18, 2012.
Communicated by Vincent Shin-Mu Tseng.
approach lacks closed-form expressions to directly perform acoustic model adaptation (e.g., speaker adaptation).

The statistical approach usually adopts modeling techniques with the maximum a posteriori (MAP) or maximum likelihood (ML) criterion to solve classification problems. This approach assumes that the features of each candidate class are clustered together and governed by a given probability distribution. A set of training data is needed to estimate these probability distributions. Most present day speech recognition and speaker identification systems employ the generative models for acoustic modeling, e.g., the popular hidden Markov models (HMMs) and the Gaussian Mixture Model (GMMs). In attempting to improve the discrimination capabilities, the discriminative training techniques, for example the maximum mutual information (MMI) criterion-based [3] and minimum classification error (MCE) criterion-based techniques [4, 5], have been shown to give significant improvement on generative models. Besides the popular methods using GMM-based modeling, the methods of super vector machine are adopted to accomplish speaker verification [6] and speaker identification [1]. Although the related research on statistical approaches for speaker identification has a long record of history, its potential for development has not been exhausted.

Apart from using the conventional generative models for acoustic modeling, the Conditional Random Fields (CRFs) which belong to a class of direct models have shown advantages for labeling sequential data [7, 8]. The CRFs are derived from Markov random fields that attempt to classify observations by maximizing the conditional probability of the labels given observations. To increase the flexibility and more accurately model the temporal structure of sequential data, Gunawardana et al. [9] and Quattoni et al. [10] augmented CRFs with intermediate hidden variables. The hidden-state probabilistic CRF framework is called hidden conditional random fields (HCRFs). In [11], conditional augmented models are proposed to overcome the drawback of training in CRF framework. Subsequent studies have shown that HCRF outperforms HMM in speech recognition [12-14]. The HCRFs used in these studies were trained with gradient based algorithms by conditional maximum likelihood (CML) criterion.

In this paper, we extends that methodology by applying MCE criterion-based training technique on HCRFs. Specifically, we adopt generalized probabilistic descent (GPD)-based training algorithm [4] with HCRF framework to improve the discrimination capabilities of acoustic models for speech and speaker recognition. The novel algorithm allows for obtaining more improvements in recognition accuracy than CML-based HCRF. Although the experiments in this study are conducted on Mandarin speech and speaker recognition, there is no inherent limitations in the proposed HCRF/GPD framework that prevents the use of it from recognition for other languages. This study also investigates the performance of acoustic models built with various models using the same discriminative training procedures. The purpose of this study is to identify the advantages of using MCE criterion on HCRF with GPD approach, in which the models are trained directly match the goal of recognition system.

The remainder of the paper is organized as follows. In section 2, the principle of the HCRF-based acoustic modeling and the discriminative training algorithm on HCRF are described. In section 3, the performance of the proposed method is examined by using a speaker identification task and a continuous Mandarin syllable recognition task. The discussion of our findings and some conclusions are given in the last section.
2. ACOUSTIC MODELING WITH HIDDEN CONDITIONAL RANDOM FIELDS

In this section, we firstly review the basic concept of HCRF. Secondly, the proposed training method on HCRF is demonstrated in detail.

2.1 Hidden Conditional Random Fields

Consider the task of predicting a speech or speaker class based on the observation sequence \( \mathbf{o} = (o_1, o_2, \ldots, o_T) \) associated with a hidden state sequence \( \mathbf{s} = (s_1, s_2, \ldots, s_T) \) from an utterance with \( T \) frames. Note that the state sequence \( \mathbf{s} \) is not observed and observation \( o_t \) is the common speech feature vector (e.g., cepstral coefficients) at \( t \)th frame. Furthermore, we denote the parameter vector of a HCRF by \( \theta \) and express the joint conditional probability of class label \( c \) and state sequence \( \mathbf{s} \) given observations by

\[
p(c, s | \mathbf{o}; \theta) = \frac{1}{Z(\mathbf{o}; \theta)} \cdot e^{\phi(c, s, \mathbf{o}; \theta)}
\]  

(1)

where \( Z(.) \) is a normalization function defined as follows:

\[
Z(\mathbf{o}; \theta) = \sum_{c, s \in \mathbf{c}} e^{\phi(c, s, \mathbf{o}; \theta)}
\]  

(2)

which ensures that the summation over all models of conditional probability obeys the probability axioms. The \( \phi(.) \) is a real-valued potential function which is usually characterized in CRF framework by \( \phi(.) = \langle \theta, f \rangle \), i.e., the inner-product between \( \theta \) and the vectored-value feature function \( f \). The inner-product gives the measure of the suitability between the feature function of observations and its corresponding CRF parameter vector. Further, the CRF framework allows flexibility in assigning the potential function. In this paper, \( \phi(.) \) is written as the following form for dealing with speech signal:

\[
\phi(c, s, \mathbf{o}, \theta) = \sum_{t=1}^{T} \sum_{\alpha} \left( \theta^{(\alpha)}_{c, s, t} \cdot f^{(\alpha)}(c, s, \mathbf{o}) \right)
\]  

(3)

where \( f^{(\alpha)} \) refers to the \( \alpha \)th vectored-value feature function which depends on the class label \( c \), observation sequence \( \mathbf{o} \), and the hidden state sequence \( s \). The term \( \theta^{(\alpha)} \) is the HCRF parameter vector associated with the feature function \( f^{(\alpha)} \).

Given these definitions, the posteriori probability of class label \( c \) can be calculated with marginalization of the joint conditional probability of class label and state sequence as follows:

\[
p(c | \mathbf{o}; \theta) = \sum_{s \in \mathbf{s}} p(c, s | \mathbf{o}; \theta).
\]  

(4)

To model the temporal structure of speech signals and allow comparison with HMM, this study adopts the linear chain structure in Fig. 1 for HCRF. Under restrictions relevant to
HCRF, the potential function $\phi(\cdot)$ can be rewritten in terms of expressions that can be calculated using dynamic programming analogous to the framework of HMM. Accordingly, the $\phi(\cdot)$ is taken as the following compact form:

$$
\phi(c,s,o;\theta) = \sum_{t=1}^{T} \left( \langle \theta_{c,s,d}^{(0)} \rangle + \langle \theta_{c,s,d}^{(1)} \rangle (c,s,o) + \langle \theta_{c,s,d}^{(2)} \rangle (c,s,o) \right)
$$

$$
= \sum_{t=1}^{T} \left( \langle \theta_{c,s,d}^{(0)} \rangle + \langle \theta_{c,s,d}^{(1)} \rangle (c,s,o) + \langle \theta_{c,s,d}^{(2)} \rangle (c,s,o) \right)
$$

(5)

where $\xi(x)$ produces a vector whose entries are the main diagonal of matrix $x$, i.e., $\xi(x) = [x_{11}, x_{22}, ..., x_{DD}]^T$; $u$ is a vector with the same length of $\theta^{(0)}$ and all entries of $u$ are 1.

Aligning the linear chain structure of HCRF with that used in HMM makes it possible to obtain the associated $d$th component $\theta_{c,s,d}$ in the HCRF parameter vector through the following equations [13]:

$$
\theta_{c,s,d}^{(0)} = -\frac{1}{2} \left( \log 2\pi \sigma_{c,s,d}^2 + r_{c,s,d} \cdot \mu_{c,s,d}^2 \right)
$$

$$
\theta_{c,s,d}^{(1)} = r_{c,s,d} \cdot \mu_{c,s,d}
$$

$$
\theta_{c,s,d}^{(2)} = -\frac{1}{2} r_{c,s,d}
$$

(6)

where $\mu_{c,s,d}$ is the $d$th component of HMM emission mean at state $s$ of class $c$ and $\sigma_{c,s,d}$ denotes the $d$th diagonal component of the HMM emission standard deviation for state $s$ of class $c$. $r_{c,s,d}$ is the reciprocal of the variance defined as $1/\sigma_{c,s,d}^2$. Note that we omit the mixture index for each state in above expressions for conciseness.

2.2 MCE Training on HCRF

Consider a set of discriminant functions $\{g_c(o; \theta)\}$. The discriminant function associated with class $c$ in HCRF framework is defined as

Fig. 1. Linear chain structure of HCRF.
\begin{align}
g_c(o; \theta) &= \log[p(c, s^*_c | o; \theta)] \\
&= \phi(c, s^*_c, o; \theta) - \log(Z(o; \theta)) 
\end{align}

where \( s^*_c \) is the maximum conditional likelihood state sequence that satisfy

\[ s^*_c = \arg \max_s \log[p(c, s | o; \theta)]. \]

The dynamic programming technique can be applied to find the maximum conditional likelihood state sequence. We adopt the segmental GPD (generalized probabilistic descent) procedure [4] to perform the MCE-based discriminative training. Let \( o \) come by class label \( c \) from \( M \) classes, the misclassification measure of the GPD training is defined as

\[ d_c(o) = -g_c(o; \theta) + g_{c'}(o; \theta) \]

where \( c' = \arg \max g_j(o; \theta) \)

i.e., the misclassification measure is constructed from the discriminant values between the correct hypothesis and the most competitive hypothesis. Following the HCRF framework, \( d_c(o) \) can be expressed by

\[ d_c(o) = \phi(c, s^*_c, o; \theta) - \phi(c, s^*_c, o; \theta). \]

Then the classification error can be approximated with a zero-one sigmoid loss function:

\[ \ell_c(o; \theta) = \frac{1}{1 + \exp(-\gamma d_c(o) + \delta)}. \]

According to the gradient of the loss function, the HCRF parameter at state \( s \) of class \( c \) belonging to the \( \alpha \)th feature function can be re-estimated as follows:

\[ \hat{\theta}^{(\alpha)}_{c,s} = \theta^{(\alpha)}_{c,s} - \eta \nabla \ell_c(o; \theta) \big|_{\theta = \hat{\theta}^{(\alpha)}_{c,s}} \]

where \( \eta \) is the learning rate for the steepest descent of the loss function. The parameter updating equations then be derived as follows:

\[ \hat{\theta}^{(\alpha)}_{c,s} = \theta^{(\alpha)}_{c,s} - \eta \times \nabla \ell_c(o; \theta) \big|_{\theta = \hat{\theta}^{(\alpha)}_{c,s}} \sum_{t=1}^{T} [-f_t^{(\alpha)}(c, s, o)] \]
\[
\tilde{\theta}_{e',s}^{(a)} = \theta_{e',s}^{(a)} - \eta \times \gamma \times \ell_{e'}(d_c(o)) \left[ 1 - \ell_{e'}(d_c(o)) \right] \sum_{t=1}^{T} f_t^{(a)}(c', s, o) 
\]  

(15)

The biggest difference between the maximum likelihood (ML)-based and MCE-based classification schemes is that ML-based classification only updates the models with correct labels. Although the overall likelihood increases with the quantity of training data, this results in patterns being distributed more widely in the feature space, with greater overlap and thus greater degradation of the discrimination capabilities of the trained models. MCE-based classification, on the other hand, modifies the models for both correct (i.e., class \( c \)) and the erroneous examples (i.e., class \( c' \)); therefore, as training quantity increases, the number of errors decreases, leading to better discrimination capabilities.

3. EVALUATION

In this section, we apply two tasks to evaluate the HCRF-based acoustic modeling approach. We firstly conduct a 200-speaker identification task. Secondly, the experiments with a Mandarin continuous syllable recognition task with HCRF models are demonstrated.

3.1 Performance Evaluation I

All speech signals were first pre-processed for each of 20-ms Hamming-windowed frame with 10-ms shift. A set of 26 recognition features was computed for each frame: 12 MFCCs, 12 delta MFCCs, and a delta log-energy and a delta-delta log-energy. To investigate the influence of different enrollment lengths on speaker modeling, the speaker models were trained with 5, 7, 10, or 20 utterances per speaker. Another 8 utterances per speaker were used for testing. The training and testing speech data in this study were selected from the MAT2000 database [15].

Three different speaker models including GMM/GPD, HMM/GPD and HCRF/GPD were evaluated. Fig. 2 illustrates the training procedures for the three models: (1) GMM/GPD: the baseline GMMs are trained with ML-based training, and GPD-based training is then applied to the baseline GMMs to obtain GMM/GPD speaker models. The number of mixture components in each GMM was empirically set to be 64, all with diagonal covariance matrices to model the likelihood probabilities for the speakers to be classified. (2) HMM/GPD: Each speaker model in this scheme consists of one initial HMM and one final HMM. Each initial and final HMM has 2 and 4 states, respectively. The baseline HMMs are trained with ML-based segmental K-means procedure and then refined by GPD-based training to generate HMM/GPD speaker models. The number of mixture components in each state was set to be 8 and accordingly the total number of parameters in HMMs was roughly the same as the number used in GMM/GPD. (3) HCRF/GPD: the baseline HCRFs are obtained by transformation through Eq. (6). The proposed training method was then performed to update the HCRF models to obtain the HCRF/GPD speaker models.
Fig. 2. The three training schemes for speaker identification.

Fig. 3. An example of the empirical loss reduction for each iteration using the proposed MCE-based training algorithm on HCRF-based speaker models.

Fig. 3 shows the learning curve of the proposed MCE-based training algorithm on HCRF speaker models. This figure shows that the average loss decreases monotonically with respect to number of iterations. This empirically presents the convergence of the proposed training algorithm. Then we investigated an ideal case of speaker identification.
In this case, all non-speech frames in the testing speech, including silence segments and short pauses, were removed perfectly by Viterbi segmentation. This is because the non-speech frames usually do not contain the information for speaker modeling. Furthermore, non-speech frames are adverse for speaker identification because they are deeply influenced by channel characters and noise effects. Fig. 4 shows the error rates of speaker identification for MATDB-4 testing speech in this ideal case. The experiments were performed by the three types of speaker models with various amounts of enrollment utterances. As expected, the error rates decreased rapidly as the enrollment speech increased. The lowest error rate of the three speaker model is 5.2%, which is obtained by the HCRF/GPD models with 20 enrollment utterances/speaker. The results of the ideal case serve as a reference for the following experiments.

Fig. 4. The error rates for MATDB-4 testing speech for GMM/GPD, HMM/GPD, and HCRF/GPD. All non-speech frame are pre-removed by Viterbi segmentation.

In practical situations, non-speech frames (silence or short pauses) cannot be completely removed from testing speech. The following discussion demonstrates the performance of speaker identification on testing speech without non-speech frame removal. Figs. 5 and 6 present the error rates of the MATDB-3 and MATDB-4 testing speech for different speaker models with various amounts of enrollment utterances. We also display the results of the GMM with 32 mixture components in these two figures. They are denoted as GMM*/GPBD. The results show that the GMM with 64 mixture components does not suffer from insufficient training data except in the case of MATDB-3 (isolated syllable) testing. In such case it consists of only 5 enrollment utterances per speaker. This extreme condition for speaker identification leads to unstable results. The length of the testing speech greatly influences speaker identification performance. The identification of MATDB-3 testing speech, which consists of only one isolated syllable per utterance, achieved relatively bad performance.

Figs. 5 and 6 also show that significant gaps exist between the error rates of HMM and HCRF and that of GMM. When one-stage dynamic programming algorithm [16] was conducted for speaker identification, GMM is considered as a single-state HMM. This accounts for the lack of transition ability between states, which creates errors due to inaccurate segmentation on testing speech. A detailed analysis on MATDB-4 testing speech between the baseline and GPBD schemes is shown in Table 1. Note that the baseline HCRF is generated from the transformation of the HMM/ML through Eq. (6). It leads to obtaining almost the same performances form the two schemes and thus we ignore the
Fig. 5. The error rates for MATDB-3 testing speech for GMM/GPD, HMM/GPD, and HCRF/GPD.

Fig. 6. The error rates for MATDB-4 testing speech for GMM/GPD, HMM/GPD, and HCRF/GPD.

Table 1. The error rates on MATDB-4 testing speech between baseline and GPD schemes.

<table>
<thead>
<tr>
<th>Number of Enrollment</th>
<th>Baseline</th>
<th>GMM</th>
<th>HMM</th>
<th>GPD</th>
<th>GMM</th>
<th>HMM</th>
<th>HCRF</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td></td>
<td>52.2</td>
<td>26.9</td>
<td>51.4</td>
<td>26.4</td>
<td>25.6</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td></td>
<td>31.7</td>
<td>18.1</td>
<td>31.0</td>
<td>17.2</td>
<td>51.6</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td></td>
<td>15.3</td>
<td>13.0</td>
<td>14.4</td>
<td>12.1</td>
<td>12.0</td>
<td></td>
</tr>
<tr>
<td>20</td>
<td></td>
<td>8.5</td>
<td>7.8</td>
<td>7.8</td>
<td>7.1</td>
<td>6.0</td>
<td></td>
</tr>
</tbody>
</table>

The best performance was achieved by the HCRF/GPD scheme for MATDB-4 testing speech with 20 enrollment utterances per speaker. This setup led to 23.0% and 15.4% decreases in error rate, respectively, compared with the results of the GMM/GPD and HMM/GPD. This indicates that the HCRF/GPD-based speaker models outperformed the GMM/GPD-based and HMM/GPD-based speaker models regardless of the length of the test or amount of enrollment speech.
This study also investigates the robust capability of the three schemes in dealing with noise effects. The noise used in this study is BABBLE noise of Signal Processing Information Base (SPIB) [17] which is generated by a hundred speakers talking in a canteen. The BABBLE noise makes a lot of damage to performances of speaker recognition because the noise spectrum seriously overlaps speech spectrum. In the experiment, the average SNR setting for noisy testing speech was 24 dB. Fig. 7 compares the error rates between the three schemes with different enrollment utterances for noisy MATDB-4 testing speech. Adding babble noise represents a simulation of many people talking with each other. The babble noise has therefore the characteristics of speech-like signal. Although the gain of noise is not heavy, the results indicate that noise greatly influences speaker identification performance, and dramatically degrades the performance. It is difficult to evaluate these three schemes in such poor error rates. To make a proper performance comparison between these schemes, we reduced the number of speakers to obtain moderate performances. A total of 50 speakers – 25 males and 25 females were used for the evaluation with noise effects. Fig. 8 shows the error rates for the 50-speaker identification task under noise effects. The HCRF/GPD scheme achieves the lowest error rate even with noise effects, with an approximately 16% error rate reduction compared with the counterpart HMM/GPD scheme with 20 enrollment utterances.

![Fig. 7. The error rates the 200-speaker identification task for noisy MATDB-4 testing speech under babble noise for GMM/GPD, HMM/GPD, and HCRF/GPD.](image1)

![Fig. 8. The error rates the 50-speaker identification task for noisy MATDB-4 testing speech between HMM/GPD and HCRF/GPD under babble noise.](image2)
3.2 Performance Evaluation II

A set of sub-syllable based HCRFs trained from MAT2000 database [15] was used for continuous Mandarin syllable recognition. We adopted 100 three-state right dependent *initial* models and 38 four-state context independent *final* models as the basic recognition units. The number of mixtures in each state varied and depended on the number of training samples, but the maximum number of speech models was set to 32, and 96 for the non-speech (or silence) model. The same recognition features as the previous speaker identification task were computed for each frame. The TEST500 set in MAT2000 is used for the testing speech in this task.

The following four schemes were used for the evaluations. Fig. 9 illustrates the relations between the schemes: (1) The HMM/ML scheme: the HMMs were trained by ML-based segmental $K$-means procedure. (2) The HMM/GPD scheme: the HMMs were trained with minimum classification error criterion by segmental GPD procedure. (3) The HCRF/SGA scheme: the HCRFs were trained with conditional ML criterion by the stochastic gradient ascent (SGA) [13] method. The bootstrapping HCRFs were obtained by the model transform from the models by HMM/ML scheme. (4) The HCRF/GPD scheme: the HCRFs were trained with minimum classification error criterion by proposed GPD procedure in this paper. The bootstrapping HCRFs were obtained by the model transform from the models by HMM/GPD scheme.

![Diagram of the four training schemes](image)

**Table 2. Syllable error rates (%) for the TEST500 testing speech.**

<table>
<thead>
<tr>
<th></th>
<th>HMM/ML</th>
<th>HCRF/SGA</th>
<th>HMM/GPD</th>
<th>HCRF/GPD</th>
</tr>
</thead>
<tbody>
<tr>
<td>HMM/ML</td>
<td>41.65</td>
<td>39.31</td>
<td>36.26</td>
<td>34.66</td>
</tr>
</tbody>
</table>

Table 2 shows the syllable error rates (%) of the four schemes tested on the TEST500 testing speech. The performance of the HMM/ML scheme was not good as expected because of their discrimination capabilities by ML criterion are fair. The continuous syllable recognition task is actually difficult because it comprises many highly confusable syllables. It can be also found from the results of the HMM/GPD and HCRF/GPD schemes, both adopted the MCE/GPD training approach, gave about 12.9% and 11.2% reduction respectively in error rates compared with the counterparts (i.e., HMM/
ML and HCRF/SGA) with ML-based schemes. This presents that the MCE/GPD-based models outperform the ML-based models. The lowest error rate of these schemes was 34.7%, obtained by the HCRF/GPD scheme. To sum up the results for clean testing speech on syllable recognition and speaker identification, we use the Table 3 to show that the improvement of HCRFs over HMMs on the two evaluations. A moderate error reduction is obtained from the HCRF/GPD scheme compared to the HMM/GPD scheme. Compared with the baseline scheme (HMM/ML), this setup reduced the syllable error rate and speaker identification error rate overall by about 16% and 23%, respectively.

Table 3. Error rates improvement (%) for clean testing speech on speech and speaker recognition.

<table>
<thead>
<tr>
<th></th>
<th>HMM/ML to HCRF/GPD</th>
<th>HMM/GPD to HCRF/GPD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Continuous syllable recognition</td>
<td>16.7</td>
<td>4.0</td>
</tr>
<tr>
<td>(20 enrollment utterances)</td>
<td>23.8</td>
<td>15.4</td>
</tr>
</tbody>
</table>

Table 4. Syllable error rates (%) for the noisy TEST500 testing speech.

<table>
<thead>
<tr>
<th>Noise/SNR</th>
<th>HMM/ML</th>
<th>HCRF/SGA</th>
<th>HMM/GPD</th>
<th>HCRF/GPD</th>
</tr>
</thead>
<tbody>
<tr>
<td>ROVER_2 / 12 dB</td>
<td>50.15</td>
<td>49.15</td>
<td>45.16</td>
<td>43.09</td>
</tr>
<tr>
<td>ROVER_2 / 9 dB</td>
<td>55.54</td>
<td>53.00</td>
<td>50.17</td>
<td>47.21</td>
</tr>
<tr>
<td>VOLVO / 12 dB</td>
<td>47.13</td>
<td>46.17</td>
<td>41.48</td>
<td>39.43</td>
</tr>
<tr>
<td>VOLVO / 9 dB</td>
<td>50.04</td>
<td>48.27</td>
<td>44.04</td>
<td>42.18</td>
</tr>
</tbody>
</table>

To generate the noisy testing speech, the ROVER_2 car noise from NTT-AT ambient noise database [18] and VOLVO car noise from SPIB [17] were added to the TEST500 testing speech with levels of 9 and 12 dB in SNR. Table 4 presents the syllable error rates of the noisy TEST500 testing speech. The best performance was still achieved by the HCRF/GPD scheme. This experimental result confirm that the MCE/GPD-based models still outperform the ML-based models in noisy speech testing speech. Figure 10 shows the average error rates (%) over noisy testing speech. The average error rates with HMM/GPD and HCRF/GPD outperform their counterparts (i.e., HMM/ML and HCRF/SGA) by 10.8% and 12.6%, respectively. The proposed HCRF/GPD approach outperform HMM/ML and HCRF/SGA by 15.3% and 12.6%, respectively. These results on noisy testing speech show consistency with previous experimental results.

Fig. 10. The average syllable error rates of the four training schemes for noisy TEST500 testing speech.
4. CONCLUSION

This paper proposes using generalized probabilistic descent-based training algorithm with HCRF framework to establish acoustic models. Two tasks were applied to evaluate the proposed approach. We firstly conduct a 200-speaker identification task in which the speakers were selected from the MAT2000 database. This study adopts the same discriminative training technique to train GMM, HMM, and HCRF speaker models, and investigates the speaker identification performance of the three speaker models using different amounts of training speech for clean and noisy testing speech. Experimental results of the task indicate that the HCRF model consistently achieved the lowest error rate among the three models regardless of the length of the test and training speech or the presence of noise. This setup led to 23.0% and 15.4% decreases in the error rate, respectively, compared with the results of the GMM and HMM schemes with 20 enrollment utterances per speaker. Secondly, a Mandarin continuous syllable recognition task with the MAT2000/TEST500 databases were demonstrated. These results show consistency with the previous task regardless of the experiments on clean and noisy TEST500 testing speech. The average error rates of noisy TEST500 testing speech with HCRF/GPD approach outperform HMM/ML and HCRF/SGA by 15.3% and 12.6%, respectively. These results confirm that the HCRF/GPD approach has good capabilities for speech recognition and speaker identification.

REFERENCES


18. NTT-AT, Ambient noise database for telephonometry, 1996.

Wei-Tyng Hong (洪維廷) received his B.S. degree in Communication Engineering from National Chiao Tung University, Taiwan, in 1991, and his M.S. and Ph.D. degrees in Communication Engineering from National Chiao Tung University, Taiwan, in 1993 and 1999, respectively. From 1999 to 2002, he was with the Industrial Technology Research Institute (ITRI), Taiwan, where he held position as a Researcher of Advanced Technology Center in the Computer and Communications Research Lab. Dr. Hong joined PenPower Technology Ltd., Taiwan as a Researcher in 2002 and as a Research Manager during 2003-2006. From 2003 to 2005, he had also been a PI of Leading Product Development Project, founded by MOEA, Taiwan. Since August 2006, he has been with the Department of Communication Engineering, Yuan-Ze University, Taiwan, where he is currently an Assistant Professor. His current research interests include speech signal processing, speech recognition, voice biometrics and soft computing techniques for pattern recognition.