DISCRIMINATIVE MLPS IN HMM-BASED RECOGNITION OF SPEECH IN CELLULAR TELEPHONY

Sunil Sivadas¹, Pratibha Jain¹ and Hynek Hermansky¹²

¹Oregon Graduate Institute of Science and Technology, Portland, Oregon, USA.
²International Computer Science Institute, Berkeley, California, USA.
email: {sunil,pratibha,hynek}@ece.ogi.edu

ABSTRACT

Deviating from the conventional Hidden Markov Model-Multi-Layer Perceptron (HMM-MLP) hybrid paradigm of using MLP for classification, the proposed discriminative MLP technique uses MLP as a mapping module for feature extraction for conventional HMM-based systems. The MLP is discriminatively trained on the phonetically labeled training data to generate the phoneme posterior probabilities. We achieved a relative word error rate reduction of 15-35% on AURORA Phase 2 continuous digit recognition task defined by ETSI.

1. INTRODUCTION

Features play a critical role in the performance of speech recognizer. As the dimensionality of the raw speech signal is too high to be modeled directly, a dimensionality reduction technique that suppresses the undesired variability with minimal loss of discrimination forms the core of the feature extraction methods. As shown in Figure 1 majority of the feature extraction schemes consist of two parts.

Speech Signal → Time-Frequency Representation → Linear/Nonlinear Transformation → Features

Physical Evidence (Critical Bands) → Phonetic Knowledge (Task-Independent/Dependent Database)

Figure 1: Generic feature extraction.

In the first part the signal is transformed into a spectral vector based on some knowledge about our auditory system, for example critical band filters [1]. The second part, motivated by statistical pattern recognition methods, is a linear/nonlinear transformation which further reduces the dimensionality. The most popular transform is the Discrete Cosine Transform (DCT). Although this works reasonably well it is very sensitive to the variations in the environment. Data driven linear transformations such as Linear Discriminant Analysis (LDA) [2, 3] have been shown to outperform DCT in speech recognition tasks [4, 5].

Standard speech recognition systems train an acoustic model to classify these feature vectors to basic speech units. Mostly the probability distribution functions are modeled as mixtures of Gaussian distributions (GMMs). The parameters of the GMM systems are usually estimated to maximize the likelihood of the training observations, rather than maximizing the discriminability of the classes. Hidden Markov Model/Multi-Layer Perceptron (HMM-MLP) hybrid approach [6] combines the discriminative power of neural network with the sequential decoding capabilities of HMM. It replaces the HMM’s GMMs with a MLP that is discriminatively trained to output each HMM state’s posterior probability given the feature vector. Recently MLP has successfully been used to estimate feature transformation suitable for continuous HMM systems [7]. In this method the neural net is trained to maximize the mutual information between features and the HMM state emission probabilities. Hermansky et al. [8, 9] use MLP to learn the mapping of any feature set to the posterior probabilities of context-independent phonemes. The posteriors when used as input features for the GMMs have achieved an average 35% relative word error rate (WER) reduction over mel-cepstral features under matched test conditions on the AURORA Phase 1 continuous digit recognition task [10]. In this paper we have studied the performance of this non-linear feature extraction technique on AURORA Phase 2 [11] which consists of both matched and mismatched test conditions.

The next section describes this non-linear feature extraction technique in detail. Section 3 describes the AURORA task, section 4 explains various experimental results and the last section summarizes the results.

2. DISCRIMINANT NONLINEAR FEATURE EXTRACTION

The basic architecture of the system is illustrated in Figure 2. The initial feature representation is a conventional front-end, for example PLP [12]. These features are fed into a 3 layer MLP with 9 frames of context at the rate of 100Hz. The MLP is trained by backpropagation with minimum cross entropy criterion. The targets are obtained by forced alignment of AURORA training data using a hybrid (HMM/MLP) recognition system trained on the OGI Numbers task [13] and realigned using embedded training on AURORA training data. The outputs of the trained MLP represent estimates of the phoneme posterior probabilities. To circumvent the skewed distribution of the posterior probabilities we remove the final softmax nonlinearity in the output layer [9]. The pre-nonlinearity outputs are close to the logarithm of the posterior probabilities. The
log-posteriors are whitened using global Karhunen-Loeve (KL) transform before being modeled by GMMs.

The MLP functions as a nonlinear feature transformation that takes multiple feature vectors as input to generate an improved discriminant feature vector that is fed into the GMM system. This can be interpreted as a dimensionality reduction transformation that utilizes individual scatter matrices, unlike LDA that collapses all class information into only two scatter matrices. From the individual scatter matrices of the phonemes present in the training data, MLP learns the boundaries between them. However, this comes at the cost of generalization and task-independence as the variability of the phonemes in the training data is significantly affected by their context and the additive noise present, in the case of noisy training data. This can be overcome to a certain extent by using a cross-validation dataset representative of the unseen data, but is restricted by the amount of development data available which is none in the case of AURORA. Another option is to train the net on a task-independent dataset that manifests all the variabilities in the target database.

By adding a MLP in our feature extraction module we have increased the number of parameters and computation in the system compared to the conventional feature extraction. But this is insignificant compared to the complexity using a large feature vector of past and future frames in the estimation of HMM parameters. By training the MLP on a task independent database we can reuse the net, thereby reducing the computation.

3. AURORA TASK

The AURORA task [10, 11] is initiated by European Telecommunications Standards Institute (ETSI) to standardize a robust feature extraction technique for distributed speech recognition. In comparison to the AURORA Phase 1, that used modified TIDIGITS downsampled from 20kHz to 8kHz, energy normalized and matched testing condition, Phase 2 consists of utterances that are not energy normalized and testing on unseen noise. Noise is artificially added to the TIDIGITS at various Signal to Noise Ratios (SNR) ranging from clean, 20dB to 0dB in steps of 5dB. Noise signals represent the most probable application scenarios for cellular phones. They have been recorded at the following places: (i) Suburban train (ii) Crowd of people (babble) (iii) Car (iv) Exhibition hall (v) Restaurant (vi) Street (vii) Airport and (viii) Train station. The long term spectra of all noises are shown in Figure 3. Training is performed on both clean and noisy data. It consists of 8440 utterances equally split into 20 subsets with 422 utterances in each subset. The 20 subsets represent 4 different noise scenarios at 5 different SNRs. The 4 noises are: suburban train, babble, car and exhibition hall at SNRs of 20dB, 15dB, 10dB, 5dB and clean condition. The speech and noise are filtered with the G.172 characteristic.

Figure 2: Discriminative MLPs in HMM-based recognition of speech.

Figure 3: Long-term spectra of all noise signals.
Three different test sets are used to assess 4004 utterances divided into 4 sets of 1001 utterances each. One noise signal is added to each subset at SNRs of clean, 20dB to 0dB in steps of 5dB. Thus, each set consists of (4x6x1001) 24024 utterances. In the first test set (Test A) the four noises are same as the ones used in the training set, resulting in matched testing condition. The second set (Test B) has four different noises, viz., restaurant, street, airport and train station, thus creates a mismatch between train and test data. The third test set (Test C) contains 2 of the 4 subsets. Here, the speech and noise are filtered with a MIRS characteristic to create a channel mismatch. The noise types used are suburban train and street.

The reference recognizer is a GMM system based on HTK software package with 11 whole word digit models, with 16 states and 3 Gaussians per state and two silence models with 3 states and 1 state with 6 Gaussians per state for inter-word pauses and intra-word pauses respectively.

### Table 1: Relative word error rate (%) reduction (20-0dB) for various features. PLP-MLP and LDA-MLP are the non-linearly transformed PLP and LDA features respectively.

<table>
<thead>
<tr>
<th>FEATURES</th>
<th>Test A</th>
<th>Test B</th>
<th>Test C</th>
</tr>
</thead>
<tbody>
<tr>
<td>PLP</td>
<td>2</td>
<td>6</td>
<td>11</td>
</tr>
<tr>
<td>PLP-MLP</td>
<td>21</td>
<td>-18</td>
<td>-1</td>
</tr>
<tr>
<td>LDA</td>
<td>19</td>
<td>28</td>
<td>38</td>
</tr>
<tr>
<td>LDA-MLP</td>
<td>28</td>
<td>27</td>
<td>30</td>
</tr>
<tr>
<td>TRAPs</td>
<td>16</td>
<td>19</td>
<td>37</td>
</tr>
</tbody>
</table>

4. EXPERIMENTS

We have used both spectral and temporal features as our initial feature representation [14]. The spectral features are derived from speech frames of 25ms duration at the rate of 100Hz. The baseline system as defined by ETSI uses 20 features consisting of 12 MFCC+log energy along with their delta and acceleration coefficients. Other features used are PLP (39 features: 13 cepstral coefficients + delta+delta) [12], temporal LDA based RASTA-like features (39 features) [15, 16] and the TempoRAI Patterns (TRAP) [17]. PLP and LDA features are fed into a 3 layer MLP with 9 frames of context resulting in 351 inputs connected to a hidden layer of 500 units and an output layer of 24 units, one for each monophone class in the dictionary. For TRAP features, each critical band MLP has 101 point input, 300 hidden units and 24 outputs. The combiner MLP has 24 x 15 = 360 inputs, 300 hidden units and 24 output units. Table 1 shows the performance of these features with respect to the baseline under different test conditions.

From Table 1 we see that although MLP improves performance over the baseline in matched test condition (Test A) its performance compared to the direct HMM method becomes non-existent in the mismatched test conditions (Test B & C). To study the relative WER reduction for various features at moderate and severe SNR conditions we divided the range of SNRs into two, 20dB-10dB and 5dB-0dB. Table 2 gives this comparison. It can be seen that the advantage provided by the MLP is absent in the 5-0dB range compared to the high SNR case, the only exception being the matched test condition. This can be attributed to the fact that MLP is not trained under 0dB SNR condition. The influence of various noises can be found from Table 3. Babble noise and Street noise are the most detrimental ones as they contain audible background speech. Another interesting observation that can be made is that MLP is not robust to convolutive distortions. This can be found by comparing the results for PLP in Subway noise from Test C (22%) and Test A (27%) conditions. In Test C additive noise is contributing more to the mismatch than the channel differences. This is evident from the difference between the results in Subway noise (seen) and Street noise (unseen). The robustness of LDA based features and TRAP to convolutive noise comes from the temporal filtering.

### Table 2: Comparison of relative word error rate (%) reduction for various features at moderate and severe SNR conditions. Each column gives the average WER reduction in the given SNR range.

<table>
<thead>
<tr>
<th>Features</th>
<th>Test A (dB)</th>
<th>Test B (dB)</th>
<th>Test C (dB)</th>
</tr>
</thead>
<tbody>
<tr>
<td>PLP</td>
<td>9</td>
<td>1</td>
<td>11</td>
</tr>
<tr>
<td>PLP-MLP</td>
<td>30</td>
<td>20</td>
<td>-18</td>
</tr>
<tr>
<td>LDA</td>
<td>-16</td>
<td>19</td>
<td>36</td>
</tr>
<tr>
<td>LDA-MLP</td>
<td>33</td>
<td>28</td>
<td>-28</td>
</tr>
<tr>
<td>TRAPs</td>
<td>3</td>
<td>19</td>
<td>24</td>
</tr>
</tbody>
</table>

4.1. Feature combination

The three types of features extract different properties of the classes. So it is natural to expect some complementary information from the three feature streams [1]. We combined the feature streams by averaging their log posterior probabilities followed by whitening before modeling them using GMMs. The results for the combination of various streams are given in Table 4.

These results are consistent with our earlier observation that spectral and temporal features provide complementary information. By combining them the performance of the combined system is better than the performance of the individual systems. We see that by combining two temporal features the performance is comparable to the best of the two. In all the cases dramatic improvement in the Test C condition is due to the fact that temporal features are robust to channel variabilities.

5. SUMMARY AND CONCLUSIONS

From the results we see that the discriminant nonlinear feature extraction based on MLP needs improvement in many areas. Since the MLP is trained on the AURORA database to discriminate the phonemes occurring in the training data it is task specific and language dependent. Its performance on a cross-language task is yet to be seen. Although MLP gives dramatic gains in the matched test condition, its contribution is not prominent in the mismatched condition. To investigate the possibility of overtraining we will evaluate the training of MLP on a development database with same vocabulary and all target SNR conditions.
Table 3: Comparison of relative word error rate (%) reduction for various features in different noise conditions.

<table>
<thead>
<tr>
<th>Features</th>
<th>Test A Noises</th>
<th>Test B Noises</th>
<th>Test C Noises</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Sub Bab Car Exh</td>
<td>Res Str Air Trn</td>
<td>Sub Str</td>
</tr>
<tr>
<td>PLP</td>
<td>2 4 -5 8</td>
<td>6 1 7 8</td>
<td>11 10</td>
</tr>
<tr>
<td>PLP-MLP</td>
<td>33 11 23 30</td>
<td>1 -35 -33 4</td>
<td>35 -40</td>
</tr>
<tr>
<td>LDA</td>
<td>29 14 17 16</td>
<td>36 18 29 30</td>
<td>50 23</td>
</tr>
<tr>
<td>LDA-MLP</td>
<td>43 20 26 20</td>
<td>33 18 22 33</td>
<td>45 15</td>
</tr>
<tr>
<td>TRAPS</td>
<td>17 3 26 18</td>
<td>20 19 11 26</td>
<td>41 33</td>
</tr>
</tbody>
</table>

Table 4: Relative word error rate (%) reduction (20-0dB) for various feature combinations.

<table>
<thead>
<tr>
<th>Features</th>
<th>Test A</th>
<th>Test B</th>
<th>Test C</th>
</tr>
</thead>
<tbody>
<tr>
<td>TRAPS + PLP-MLP</td>
<td>29</td>
<td>26</td>
<td>37</td>
</tr>
<tr>
<td>TRAPS + LDA-MLP</td>
<td>28</td>
<td>24</td>
<td>37</td>
</tr>
<tr>
<td>TRAPS + PLP-MLP</td>
<td>34</td>
<td>27</td>
<td>40</td>
</tr>
</tbody>
</table>

6. ACKNOWLEDGMENTS

The research was supported by NSA/DoD under MDA904-06-C-2089, NSF under IRI-0712579 and by an industrial grant from Intel Corporation.

7. REFERENCES