Robust Language Identification Using Convolutional Neural Network Features

Sriram Ganapathy\(^1\), Kyu Han\(^1\), Samuel Thomas\(^1\), Mohamed Omar\(^1\), Maarten Van Segbroeck\(^2\), Shrikanth S. Narayanan\(^2\)

\(^1\)IBM T.J. Watson Research Center, Yorktown Heights, NY, USA.
\(^2\)Signal Analysis and Interpretation Laboratory, University of Southern California, Los Angeles, USA.

\{ganapath,kjhan,sthomas,mkomar\}@us.ibm.com,\{maarten,shri\}@sipi.usc.edu

Abstract

The language identification (LID) task in the Robust Automatic Transcription of Speech (RATS) program is challenging due to the noisy nature of the audio data collected over highly degraded radio communication channels as well as the use of short duration speech segments for testing. In this paper, we report the recent advances made in the RATS LID task by using bottleneck features from a convolutional neural network (CNN). The CNN, which is trained with labelled data from one of target languages, generates bottleneck features which are used in a Gaussian mixture model (GMM)-i\textbf{v}ector LID system. The CNN bottleneck features provide substantial complimentary information to the conventional acoustic features even on languages not seen in its training. Using these bottleneck features in conjunction with acoustic features, we obtain significant improvements (average relative improvements of 25\% in terms of equal error rate (EER) compared to the corresponding acoustic system) for the LID task. Furthermore, these improvements are consistent for various choices of acoustic features as well as speech segment durations.

Index Terms: Convolutional Neural Networks, Bottleneck Features, Language Identification.

1. Introduction

The DARPA Robust Automatic Transcription of Speech (RATS) [1] program targets the development of speech systems operating on highly distorted speech recorded over “degraded” radio channels. The data used here consists of recordings obtained from retransmitting a clean signal over eight different radio channel types, where each channel introduces a unique degradation mode specific to the device and modulation characteristics [1]. For the language identification (LID) task, the performance is degraded due to the short segment duration of the speech recordings in addition to the significant amount of channel noise. In this paper, we discuss the techniques developed to improve the LID system performance over the previous submission [2].

Traditionally, phoneme recognition followed by language modeling (PRLM) was one of the popular methods for automatic LID task [3, 4]. This approach uses a multilingual phoneme recognizer to generate phoneme sequences which are converted to language model (n-gram) features for the LID classifier. The success of this approach is dependent on the performance of the phoneme decoder. For relatively clean data with good phoneme recognition accuracies, the PRLM method provides good performance comparable to acoustic systems [5]. However, the performance of phoneme decoders and speech recognition systems is significantly degraded for the highly noisy data in the RATS corpus [6]. In the recent past, the use of multi-layer-perceptron (MLP) based posterior features were attempted for LID [7]. The Tandem features have shown promising results [8]. Motivated by this effort, we explore the use of convolutional neural network (CNN) based features for LID.

CNNs are variants of MLPs containing one or more convolutional layers and max pooling layers [9]. A convolutional layer consists of a set of weights which process a portion of the input signal. These weights are shared along the entire input space. The max pooling layer generates a lower resolution version of convolutional filter outputs by computing the maximum value of filter activations within a specified window. Recently, CNNs have shown promising results for various phoneme recognition and keyword spotting (KWS) tasks [10, 11].

In this paper, we develop a LID system using a CNN based phoneme recognizer trained on one of the target languages. The CNN is trained with log-mel spectrogram and contains a bottleneck (BN) layer before the output layer. For LID, the output of the BN layer from the trained CNN is used as feature representations for a Gaussian mixture model (GMM). The Gaussian mean supervector is converted to an i\textbf{v}ector representa-
The main advantage of CNNs for noisy and channel degraded speech comes from the use of local filters, which focus on relatively cleaner parts of the spectrum. In such a case, the assumption is that the local filters focus on relatively cleaner parts of the spectrum can still extract speech characteristics well enough to overcome any ambiguity arising from the noisy parts. The weight sharing and max pooling improve the robustness of the CNN to small frequency shifts. This is important because, for example, the deformations for the same phoneme may appear on slightly different frequencies for different speakers or even for the same speaker due to linear frequency transpositions caused by the channel [1]. Furthermore, weight sharing of the filters helps in avoiding the issues with over-fitting and improves generalization due to the reduced number of trainable parameters.

4. Analyzing CNN-BN features for LID

In this section, we explore the usefulness of CNN-BN features for the LID system. We use a CNN trained on Arabic Levantine (ALV). The spectrographic representation of a portion of Arabic recording is shown in the left panel of Fig. 3. The posterioriogram representation, which is the two dimensional plot of phoneme posteriors stacked along time, for this recording is shown in the bottom panel of Fig. 3. The similar plots for a Pashto (PUS) recording is shown in the right panel. Typically, a posterioriogram
with sharp activations indicates a good knowledge of the underlying phonetic content which could be useful for any application based on the posterior features. The posterior representation of ALV data is sharper and less noisy as the CNN is trained with ALV phonemes. Although the posteriogram for PUS data is noisy, there exists regions of the signal which generate sharp posteriors particularly for voiced regions. As seen in this figure, the information provided by the spectrogram and posteriogram streams are quite complimentary. The BN features used for LID experiments are a linear transformation of the posterior outputs except for a softmax operation.

For LID experiments, we concatenate the acoustic features with BN features and train the GMM based ivector model. In Fig. 4, we plot the first two principal components of the MLP hidden layer representation obtained using 30s recordings. The left panel shows a scatter plot for the LID system which uses acoustic features alone (in this case, power normalized cepstral coefficients [14]) and the right panel shows the same plot where the system was trained using a concatenation of acoustic and BN features. The scatter plot of the two significant PCA dimensions reveals that the fusion of acoustic and CNN-BN features improves the separation between the language classes considerably. This is desirable for improving the LID performance as the reduced overlap among language classes would result in a smaller number of false alarms for any given threshold.

5. Experiments

The development and test data for the LID experiments use the LDC releases of the Phase-I RATS LID development [1]. This consists of speech recordings from previous NIST-LRE clean recordings as well as other RATS clean recordings passed through eight (A-H) noisy communication channels. The training data contains about 270 hours of audio recorded over each radio channel. The five target languages are Arabic, Farsi, Dari, Pashto and Urdu. In addition to this, the database consists of several other impostor languages. In our experiments, the GMM-UBM is trained using 43,607 recordings from the eight channels. The utterance level GMM statistics are used to train a factor analysis based ivector projection [12]. This model is trained with 33,672 recordings of 120sec duration. The i-vectors are used in a backend consisting of MLP hidden layer projection followed by a SVM training with the 12th order polynomial kernel (Sec. 3). We use 250k recordings of all durations for the MLP training and 82,398 recordings for SVM training. The test data consists of two subsets - 52,789 recordings from the eight noisy channels and four durations (120s, 30s, 10s and 3s) called the EVAL set as well as 9,899 recordings from the DEV set.

In the initial set of experiments reported in Table 1, we use acoustic features based on power normalized cepstral coefficients (PNCC) [14]. The PNCC features are used to train the LID system (Sec. 3) with 250 dimensional (optimized for
Figure 5: Performance of various acoustic features with and without BN features for various speech segment durations of DEV set.

Table 1: Performance (EER %) of the LID system on the EVAL test set (DEV set in parentheses) for PNCC features with CNN features from ALV, FAS.

<table>
<thead>
<tr>
<th>Feat.</th>
<th>120s</th>
<th>30s</th>
<th>10s</th>
<th>3s</th>
</tr>
</thead>
<tbody>
<tr>
<td>PNCC</td>
<td>1.3</td>
<td>2.9</td>
<td>6.7</td>
<td>14.2</td>
</tr>
<tr>
<td>BN-ALV</td>
<td>1.3</td>
<td>2.3</td>
<td>5.9</td>
<td>15.3</td>
</tr>
<tr>
<td>BN-FAS</td>
<td>1.1</td>
<td>2.3</td>
<td>6.0</td>
<td>15.0</td>
</tr>
<tr>
<td>PNCC + BN-ALV</td>
<td>0.8</td>
<td>2.0</td>
<td>4.9</td>
<td>12.2</td>
</tr>
<tr>
<td>PNCC + BN-FAS</td>
<td>0.8</td>
<td>2.4</td>
<td>5.4</td>
<td>12.7</td>
</tr>
</tbody>
</table>

Table 2: Performance (EER %) of the LID systems on the DEV set using PNCC features with CNN features from ALV, FAS fused at various levels - feature, ivector and score.

<table>
<thead>
<tr>
<th>Cond.</th>
<th>120s</th>
<th>30s</th>
<th>10s</th>
<th>3s</th>
</tr>
</thead>
<tbody>
<tr>
<td>PNCC + ALV-BN Fusion</td>
<td>2.7</td>
<td>4.3</td>
<td>6.6</td>
<td>11.7</td>
</tr>
<tr>
<td>ivec</td>
<td>2.3</td>
<td>3.7</td>
<td>6.8</td>
<td>13.4</td>
</tr>
<tr>
<td>Score</td>
<td>2.6</td>
<td>3.8</td>
<td>6.4</td>
<td>12.5</td>
</tr>
<tr>
<td>PNCC + FAS-BN Fusion</td>
<td>2.8</td>
<td>3.6</td>
<td>6.6</td>
<td>11.1</td>
</tr>
<tr>
<td>ivec</td>
<td>2.3</td>
<td>4.0</td>
<td>6.1</td>
<td>12.8</td>
</tr>
<tr>
<td>Score</td>
<td>2.4</td>
<td>3.7</td>
<td>6.1</td>
<td>12.3</td>
</tr>
</tbody>
</table>

best performance [2]) ivectors followed by the SVM classifier. We experiment with the addition of CNN BN features generated from ALV-CNN as well as FAS-CNN to train the LID system with 300 dimensional ivectors. We also experiment with the use of BN features alone without any acoustic features with 200 dimensional ivectors. As seen in Table 1, the performance of the BN-FAS features are moderately better than the performance of the PNCC features. The use of BN features in addition with PNCC features provides significant improvement in performance for LID system for various test segment durations as well as the choice of test set. The BN features provide about 21% relative improvement in the EVAL set and about 25% in the DEV set.

The impact of BN features for various acoustic features is shown in Fig. 5. Here, we use a variety of feature processing techniques like mel frequency cepstral coefficients (MFCC) [15], frequency domain linear prediction (FDLP) [16], Gammatone [17] and cortical [18] features. In these experiments, the ALV-CNN based BN features are used and the results are reported on the DEV set for different speech segment durations. As seen in Fig. 5, the performance of all these features are improved by the use of BN features. The relative improvements are consistent even for short speech segment durations. These results illustrate that the bottleneck features based on CNN are both informative as well as complimentary to any choice of acoustic features for the LID task.

The results presented till now use the BN features in concatenation with the acoustic features. The final set of experiments, reported in Table 2, investigate the other methods of fusing the two streams, namely ivector fusion, where the ivectors from the two systems are used to jointly train the back-end classifier as well as score fusion, where the scores from the two LID systems (acoustic and BN) are linearly combined with equal weighting. The feature fusion provides the best results although the ivector fusion provides good results for the 120s duration.

6. Summary

We have presented the application of convolutional neural network based phoneme recognition features for the LID task on the highly distorted radio channel data. The CNN BN features provide robust representations which are quite useful for the LID task by themselves. When the BN features are used in conjunction with acoustic features, significant improvements are obtained. These results are consistent for a variety of acoustic feature representations as well as the use of different target languages in CNN training. These experiments encourage us to pursue the use of multi-lingual CNNs in the future.
7. References


