A Penalized Logistic Regression Approach to Detection Based Phone Classification

Sabato Marco Siniscalchi¹, Torbjørn Svendsen¹, and Chin-Hui Lee²

¹Department of Electronics and Telecommunications 
Norwegian University of Science and Technology, Trondheim, Norway
²School of Electrical and Computer Engineering 
Georgia Institute of Technology, Atlanta, GA 30332 USA

{marco77, torbjorn}@iet.ntnu.no, chl@ece.gatech.edu

Abstract

Recently, we have proposed a detection-based speech recognizer which has two main components: a bank of phonetic feature detectors implemented with hidden Markov models (HMMs), and an event merger. Each detector generates a score that pertains to some phonetic features, e.g., voicing. The merger combines all these scores to generate phone labels. The parameters of the detectors and the merger can be optimized either separately or jointly, and we showed that penalized logistic regression machine (PLRM) is a convenient tool for joint optimization. We validated our approach on a rescoring scheme. In this work, we tackle the phone classification problem and show that high level phone accuracy can be achieved without a direct modeling of the phones when PLRM is used. We also show that better results can be obtained by increasing the number of phonetic features, and that our method outperforms phone classifiers trained either by maximum likelihood estimation, or maximum mutual information.

Index Terms: Detectors, discriminative training

1. Introduction

In recent years, there has been a renewed interest in integrating articulatory information into automatic speech recognition (ASR) systems, and several studies have been carried out on this topic (e.g., [1, 2, 3, 4, 5]). One of the main supporting argument of these studies is that ASR engines can be improved by using more linguistically motivated feature extraction and modeling techniques [6]. Indeed, it has been shown that articulatory features have several nice properties, such as robustness to noise and cross-speaker variation [1], portability across different languages [4], and explicit modeling of linguistic information which makes it easier to deal with non-native and hyper-articulated speech [2].

Several data-driven paradigms have been proposed to extract articulatory features along with different schemes to use these features with ASR systems. For example, in [1] a bank of artificial neural networks (ANNs) was built to classify phonological features. The posterior probabilities given by each ANN were concatenated and integrated by a higher-level ANN trained to generate phone posteriors. The higher-level ANN was then used in a hybrid ANN/HMM system for continuous number recognition. In [2], a stream architecture is proposed to augment acoustic models based on context-dependent sub-words with articulatory motivated acoustic models. For each articulatory feature, the articulatory-based acoustic model is implemented using a pair of Gaussian mixture models (GMMs) with

Figure 1: A detection-based system to generate phone posterior probabilities. The evidence merger is implemented with an ANN. complementary distribution. Conditional random fields were used in [5] for integrating outputs of several phonetic feature ANN-based classifiers.

We have built a set of binary value classifiers (detectors) to find the presence or absence of manner and place of articulation features directly from spectral features, such as Mel-frequency cepstrum coefficients (MFCCs). These detectors were built using ANNs in [3, 4], and their outputs were used as feature vectors to train a HMM/ANN phone recognizer. In [7], we instead used segment based speech detectors to learn the mapping from the spectral-based feature space to the phonetic feature space. These detectors were implemented with a pair of competing HMMs and were used as part of a word lattice rescoring algorithm. There, the outputs of the detectors were concatenated and passed to an higher-level ANN which provided phone posteriors.

Figure 1 shows the combined system made of detectors and the integrative ANN. The articulatory classifiers are used to indicate whether a particular feature exists in the frame (or segment) being analyzed. The ANN merges the articulatory classifiers’ outputs and generates evidence at the phone level. In this scheme, all of the classifiers (i.e., the speech detectors and the event merger) are treated as independent systems and trained separately. We have shown that that joint optimization of the speech detectors and the evidence merger using a global criterion function leads to better performance in a continuous phone recognition task [8]. In a different work, we have shown that by extending the set of detectors a better phone recognition performance can be achieved [3].

The goal of this work is to demonstrate that a high phone accuracy can be achieved by combining different phonetic features representations and jointly optimizing the parameters of
Table 1: Phonemes list in terms of phonetic features.

<table>
<thead>
<tr>
<th>Phonetic Classes</th>
<th>Phonemes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Vowel</td>
<td>iy ih eh ey ae aa aw ay ah oy ow uh uw er</td>
</tr>
<tr>
<td>Fricative</td>
<td>jh ch s sh z f th v dh hh</td>
</tr>
<tr>
<td>Nasal</td>
<td>m n ng</td>
</tr>
<tr>
<td>Stop</td>
<td>b d f g p t k dx</td>
</tr>
<tr>
<td>Approximant High</td>
<td>ch ih iy jh sh uh uw y</td>
</tr>
<tr>
<td>Coronal</td>
<td>d f x t s z dh th</td>
</tr>
<tr>
<td>Dental</td>
<td>hh</td>
</tr>
<tr>
<td>Glottal</td>
<td>b f m p v w</td>
</tr>
<tr>
<td>Labial</td>
<td>aa ae aw ay oy ah eh ey ow</td>
</tr>
<tr>
<td>Low</td>
<td>er r</td>
</tr>
<tr>
<td>Mid</td>
<td>k g ng</td>
</tr>
<tr>
<td>Retroflex Velar</td>
<td>aa ae ah ay ah b d dh dx eh</td>
</tr>
<tr>
<td>Voiced</td>
<td>er ey g ih iy jh l m n ng ow oy r uh uw v w y z</td>
</tr>
<tr>
<td>Round</td>
<td>aw ow uw uh v y oy r w</td>
</tr>
<tr>
<td>Tense</td>
<td>aa ae aw ey iy oy ow uw ch s sh f th t k hh</td>
</tr>
<tr>
<td>Anterior Back</td>
<td>b d dh dx f l m n p s t h v z w</td>
</tr>
<tr>
<td>Back</td>
<td>aa ah ay ow oy uh uw g k</td>
</tr>
<tr>
<td>Continuant</td>
<td>aa ae ah ay dh eh er r ey</td>
</tr>
<tr>
<td>Vocalic</td>
<td>1 f th iy oy ow s sh th uh uw v w y z aa ae ah aw ay eh er ey iy</td>
</tr>
<tr>
<td></td>
<td>1 ow oy r uh uw w y</td>
</tr>
</tbody>
</table>

Table 2: A list of phonetic features.

<table>
<thead>
<tr>
<th>Phonetic Classes</th>
<th>Approximant</th>
<th>Anterior</th>
<th>Back</th>
<th>Continuant</th>
<th>Dental</th>
<th>Fricative</th>
<th>Glottal</th>
<th>Labial</th>
<th>Nasal</th>
<th>Non-vowel</th>
<th>Vocalic</th>
</tr>
</thead>
<tbody>
<tr>
<td>Approximant High</td>
<td>ch ih iy jh sh uh uw y</td>
<td>d f x t s z dh th</td>
<td>hh</td>
<td>b f m p v w</td>
<td>aa ae aw ay oy ah eh ey ow</td>
<td>er r</td>
<td>k g ng</td>
<td>aa ae ah ay ah b d dh dx eh</td>
<td>er ey g ih iy jh l m n ng ow oy r uh uw v w y z</td>
<td>aw ow uw uh v y oy r w</td>
<td>aa ae aw ey iy oy ow uw ch s sh f th t k hh</td>
</tr>
<tr>
<td>Non-vowel</td>
<td>1 ow oy r uh uw w y</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Figure 2: The logistic regression model with the likelihood-ratio mapping.

The list of phonetic features showed in Table 2 is obtained by combining some of the Sound Pattern of English (SPE) features defined by Chomsky and Halle [10], and the manner and place of articulation classes plus voicing. This combination is quite empirical, but we believe it to be beneficial to the detection results because the speech detectors are not ideal. Therefore, some of the detection mistakes may be recovered by adding some redundancy. Moreover, different phonetic feature schemes provide a different organization of phonemes into phonetic classes, and that implies a different partitioning of the phoneme space. The use of different partitioning schemes may be beneficial when data-driven detectors are merged together. Experimental evidence (see Section 4) seems to confirm our intuition.

3. Joint Optimization of Detectors and Merger

The penalized logistic regression machine (PLRM) [11] is used for jointly training the parameters of the detectors and the merger. We have already shown the benefits of this choice in [8]. For the sake of completeness, we present again the criterion function that is to be minimized in order to estimate the model parameters. We also briefly explain the optimization algorithm used to minimize the criterion function. More details can be found in [12].

Assume that the sequence of $T$ feature vector $x = (x_1, \ldots, x_T)$ is extracted from a speech segment, with phone label $y \in \mathcal{Y}$, where $\mathcal{Y}$ is the set of $K$ phones (in the experiments

\begin{equation}
\frac{1}{T} \sum_{t=1}^{T} \log \frac{p(x_{1:t} | \lambda^m)}{p(x_{1:t} | \lambda^n)}
\end{equation}
we use a set of \(K = 39\) phones as explained later. PLRM employs a nonlinear logistic regression model of the form

\[
\hat{p}_k = \frac{e^{u_k^T \phi(x; \Lambda)}}{\sum_{l=1}^{K} e^{u_l^T \phi(x; \Lambda)}},
\]

for estimating the conditional probability of a phone label \(y\) given a segment \(x\). \(\phi\) is a nonlinear transformation that maps the feature sequence \(x\) into an \(M\)-dimensional vector \(\phi(x; \Lambda)\) parameterized by \(\Lambda\), and \(u_k\) is \(M\)-dimensional weight vectors. For convenience, the weight vectors are taken to be rows of a \(K \times M\)-dimensional parameter matrix \(W\). In this work, the set of \(M = 22\) speech detectors introduced in the previous section constitutes the mapping \(\phi\):

\[
x \mapsto \phi(x; \Lambda) = [1, \frac{1}{T} L \log \Lambda L, ..., \frac{1}{T} L \log \gamma L L M]^T.
\]

The set of all the HMM parameters of the detectors, i.e., \(\Lambda = \{\lambda_1, ..., \lambda_M, \lambda^2_1, ..., \lambda^2_M\}\) defines the above mapping.

We want to jointly optimize the detector parameters \(\Lambda\) and the parameters \(W\). We do this by estimating the pair \((W^*, \Lambda^*)\) that minimizes a global criterion function on a set of training data \(D = (x^{(i)}, y^{(i)})^T\), which consists of pairs of speech segments and their phone labels. We choose the criterion function [12]

\[
P^\delta_W(W; \Lambda; D) = -L \sum_{i=1}^{L} \log \hat{p}_{y(i)} + \frac{\delta}{2} \text{trace} \Gamma W \Sigma W^T,
\]

where the first term is the negative log of the logistic regression likelihood, and the second term is a penalty term weighted by a parameter \(\delta > 0\). The matrix \(\Gamma\) is a \(K \times K\) diagonal matrix whose \(k\)th diagonal element is the fraction of training samples with the \(k\)th class label, and \(\Sigma = (1/L) \Phi \Phi^T\), where \(\Phi\) is an \((M) \times L\) matrix with columns \(\phi(x^{(i)}, \Lambda)\). Although \(P^\delta_W(W; \Lambda; D)\) is convex with respect to \(W\), it is not guaranteed to be convex with respect to \(\Lambda\). A local minimum can be obtained by using a coordinate descent approach with coordinates \(W\) and \(\Lambda\). For the convex minimization with respect to \(W\), we use the method in [11]. As for the minimization with respect to \(\Lambda\), we use the Rprop method [13].

4. Experiments

In the following sections, we present the experimental setup, results, and discussion.

4.1. Experimental Setup

All experiments were conducted using the TIMIT corpus [14]. We mapped the original TIMIT phonetic labels into 39 phones [15] and ignored the glottal stop. MFCCs were used as acoustic parametrization of the speech signal. Spectral analysis was performed using a 40 channel Mel filter bank from 64Hz to 8kHz. The spectral-tilt was corrected using a pre-emphasis coefficients implemented in the HTK toolkit.

The number of inputs of the ANN is equal to 22 (i.e., equal to the number of phonetic features listed in Table 1). The set of log-likelihoods scores computed as in eq. 1 are used to feed the ANN. The number of outputs is 39, which corresponds to the number of phones listed in Table 1. The softmax activation function is used at the output of the ANN; whereas, each hidden node has a sigmoidal activation function. The standard back propagation algorithm was used as training method, and to avoid over-fitting the increment in classification error on the TIMIT development set during the training phase was adopted as stopping criterion. The same criterion was used to set the number of nodes in the hidden layer, and it was found that best performance on the development set is achieved with 800 hidden nodes. This configuration was used to compute the phone classification accuracy on the development set. The estimation of the HMM detector parameters was performed using MLE as implemented in the HTK toolkit.

The PLRM was implemented as in [12]. Only the mean values of the target HMMs were updated, while all the other target HMM parameters and all of the non-target HMM parameters were kept fixed. The initial weight matrix \(W\) was estimated using the penalized logistic regression machine (PLRM) with 10 Newton iterations. The RProp method with 50 iterations was used to update the HMM mean values. The update of \(W\) in each coordinate descent iteration was done using PLRM with 2 Newton iterations. The development set was used to select a coordinate descent iteration for which to stop the training algorithm. The value found was 6 coordinate descent iterations. The parameter \(\delta\) was set equal to 1000.

4.2. Results & Discussion

Table 3 summarizes the results obtained with the proposed phonetic features detector based approach when detector and merger (ANN) are trained separately. The first column of Table 3 shows the results using only 16 detectors (ANN\16-DET), that is manner and place of articulation detectors along with a detector for voicing. The second column, on the other end, shows the phone accuracy using all of the 22 detectors (ANN\22-DET). An absolute improvement of 1.7% is observed when more speech detectors are used, and this validates our initial intuition, that is, the proposed technique can benefit from the introduction of redundancy at the detection level.

| Table 2: Classification accuracy on the TIMIT development set when disjoint optimization of the parameters of the detectors and merger is performed. |
|-----------------|------------------|
| ANN\16-detector | ANN\22-detector  |
| 69.4%           | 71.1%            |

Table 4 shows that by jointly optimizing the parameters of all of the HMM-based phonetic features detectors within the
PLRM framework, a phone classification accuracy of 75.5% can be achieved. Further, the PLRM approach results in an absolute improvement of 4.4% if compared with the ANN[22]-DET technique. Table 4 also lists the phone accuracy that can be obtained using directly a set of HMM-based phone classifiers (PHN-CL) trained by either maximum likelihood estimation (MLE), or maximum mutual information (MMI) 2. Each phone model was implemented as a 3-state left-right HMM with 10 Gaussian mixture components per state. Also, the performance of the C-Aug method [9] is reported. The basic idea behind the C-Aug technique is to combine the strengths of generative models within a conditional model; a generative model readily handles sequential data and provides information about the observations. This information is then utilized by a discriminative exponential model. The C-Aug method tries to model long term dependency ignored by conventional HMM-based speech models. The interested reader is referred to [9] for more detail.

By comparing the results showed in Table 4 and Table 3, several conclusions can be drawn. The proposed detection-based approach without directly modeling phones outperforms HMM-based phone classifiers, even when a discriminative training technique is used to estimate the HMM-based phone model parameters. Further, our system can achieve virtually identical results to the recently proposed C-Aug method when comparable training and testing conditions are used. For the sake of completeness, it is worth to point out that C-Aug can be performed over MMI-trained phone models and a phone classification accuracy of 76.6% was reported. On the other end, we trained our set of detector using standard MLE, and we believe that a better performance can be reached by using MMI. Moreover, better phone classification results have been reported using hidden conditional random fields (HCRFs) [16], but the training and testing condition are different 2. In [17] penalized logistic regression over HMM-based phone models was used for phone classification with good results, but the signal parametrization is different.

5. Summary

In this paper, a detector based approach for phone classification was presented. A bank of HMM-based binary classifiers was designed for detecting several phonetic features which are then combined to generate phone labels. Two different combining schemes were presented. The first scheme is based on a feed-forward single layer ANN, and it was shown to achieve a phone classification accuracy comparable to a HMM-based phone classifier trained by MLE. A penalized logistic regression approach is used as a second scheme to jointly train all the phonetic feature detectors and directly generate phone labels.

With this second scheme performance comparable to discriminatively trained HMM-based phone model and to the recently proposed conditional augmented methods have been achieved.

6. References