DISTRIBUTED SPEECH RECOGNITION USING NOISE-ROBUST MFCC AND TRAPS-ESTIMATED MANNER FEATURES

Pratibha Jain¹, Hynek Hermansky¹,² and Brian Kingsbury³

¹OGI School of Science and Engineering, OHSU, Portland, Oregon, USA.
²International Computer Science Institute, Berkeley, California, USA.
³IBM T. J. Watson Research Center, Yorktown Heights, New York, USA
email: {pratibha, hynek}@ece.ogi.edu, bedk@us.ibm.com

ABSTRACT

In this paper, we investigate the use of TemPoRal PatternS (TRAPS) classifiers for estimating manner of articulation features on the small-vocabulary Aurora-2002 database. By combining a stream of TRAPS-estimated manner features with a stream of noise-robust MFCC features (earlier proposed in the Aurora-2002 evaluation by OGI, ICSI and Qualcomm), we obtain an average absolute improvement of 0.4% to 1.0% in word recognition accuracy over noise-robust MFCC baseline features on Aurora tasks. This yields an average relative improvement of 54% over the reference end-pointed MFCC baseline. Estimation of the manner features can be performed on the server without increasing the terminal-side computational complexity in a distributed speech recognition (DSR) system.

1. INTRODUCTION

A TempoRAI PatternS (TRAPS) acoustic model focuses on the class-dependent temporal patterns in individual critical bands using a two-stage classification strategy [1]. The first stage of a TRAPS model learns the temporal energy patterns of the acoustic units independently using either a simple linear discriminant function or a multi-layer perceptron (MLP) in each individual critical band. The second stage then combines decisions from the individual critical bands to produce a final classification. In prior work, a temporal context of ~ 1 s., which covers around 2–3 syllables, has been used for learning the frequency-localized temporal patterns. A MLP has been used to combine the decisions from the individual critical bands. Prior work on TRAPS models shows that TRAPS-estimated features provide around 16–20% complementary information to conventional cepstral features [1]. It also shows that combination of TRAPS-estimated features and cepstral features improves speech recognition performance [3, 4]. Motivated from the above studies, we propose an optimized TRAPS-based system to estimate manner of articulation features at the server-end in DSR systems. The TRAPS-based system is optimized to meet the minimum computational complexity, memory and delay requirements suggested in the Aurora tasks. The optimized TRAPS-based system has an algorithmic latency of about 90 ms and it uses around 19000 parameters. Previously proposed TRAPS systems typically had latencies of 500 ms and 72000 parameters or more [1]. We also present a motivation behind estimating manner of articulation features by showing gain in joint mutual information between phone classes and the features on using these features along with conventional cepstral based features.

2. TERMINAL-SIDE PROCESSING

The terminal-side signal processing in our proposed system is identical to that submitted in the Aurora-2002 evaluation by OGI, ICSI and Qualcomm [9]. First, Wiener filtering is applied to suppress noise in the power spectra. The denoised power spectra are then binned into fifteen Mel-spectral bins, compressed with a log nonlinearity, and filtered by an FIR filter estimated using linear discriminant analysis to reduce channel effects in the log critical band energies. MFCCs are computed and then downsampled by a factor of two. Voice activity detection (VAD) is also performed by the terminal, so the final bit-stream that is transmitted to the server contains both compressed MFCC coefficients and VAD flags.

Fig. 1. Overall system description using TRAPS
3. SERVER-SIDE PROCESSING

At the server, the bit-stream is uncompressed and upsampled by a factor of two. Two streams of features are then generated from these MFCC features. The first stream is produced by processing the MFCC features through an on-line mean and variance normalization stage, and then appending delta and delta-delta features to form a 45-dimensional feature vector. These features were earlier proposed in Aurora-2002 evaluation jointly by OGI, ICSI and Qualcomm [9]. The second stream consists of six manner of articulation features that are estimated using a TRAPS model. These two feature streams are concatenated and whitened. This forms the final 51-dimensional feature vector that is given to the back-end, HMM-based speech recognizer. Only frames marked as speech by the terminal-side voice activity detector are passed to the speech recognizer [9]. The Overall system with TRAPS is shown in Fig. 1.

4. TRAPS ESTIMATION OF MANNER FEATURES

4.1. Input to the TRAPS models

The fifteen Mel critical band energies are reconstructed by applying Inverse Discrete Cosine Transform (IDCT) to upsampled quantized MFCC features that are sent from the terminal side [9]. While earlier TRAPS models used ~ 1s. of context (±50 samples around the current sample), we used 190 ms of context (±9 samples around the current sample) in our experiments. The reduction in the size of the input context also satisfies the limit on the total algorithmic latency. The 19-sample input window in each reconstructed critical band is normalized to have zero mean and unit variance, and is then reduced to 10 samples using principal component analysis (PCA). The PCA basis is estimated from reconstructed energy trajectories in the fifth Mel band on an independent training set of noisy TIMIT utterances. The 10 PCA basis covers 99% of the total variability present in the original 19-point long feature space. The PCA basis vectors are similar to the cosine basis as shown in Figure 2. In effect, 10 point modulation spectrum, calculated over 19-point long context window, is used by band-classifiers. Given that the original input to the TRAPS system has been upsampled by a factor of two, it is not surprising that a 50% reduction in the dimensionality of the 19 point long original TRAPS input vector through PCA is possible. The current processing is not equivalent to a simple linear transform, however, due to the variance normalization that takes place between the upsampling of the MFCC features and the PCA.

4.2. MLP band-classifiers

For each reconstructed Mel critical band, a feed-forward multi-layer perceptron is trained to classify input vectors into six broad categories based on manner of articulation. The classes are “vocalic/approximant,” “nasal,” “fricative,” “flap,” “stop,” and “silence” [5]. Each MLP has 10 input units, a single layer of 25 hidden units having sigmoidal activation functions, and 6 output units with a softmax activation function. The MLPs are trained using error backpropagation with a cross-entropy error criterion. Performance on a held-out cross-validation set determines when training is stopped. The MLPs are trained on noisy-TIMIT database that has noises present in the multi-style training dataset of Aurora-2 task. We had two reasons for using the manner features. First, we expected the manner features to generalize well across different tasks and noise conditions because each manner-based class is composed of multiple phones so there is more data available to train the TRAPS-based acoustic models. Second, in an earlier study on SPINE (Speech In Noisy Environments [7]) data, we observed that the joint mutual information between acoustic features and phone labels was higher by 0.12 bits for a feature vector computed from manner features and PLP-cepstra than for a feature vector computed from PLP-cepstra alone [6] (Table 1). Both feature vectors had the same number of dimensions. In this study, the mutual information between features and phone labels was approximated by using a Gaussian mixture model to vector quantize (VQ) the speech features and then computing the mutual information between the VQ labels and the phone labels.

4.3. Merging the information from individual critical bands

Like previous TRAPS systems [1], we use a feed-forward MLP to merge the band-local decisions. Unlike previous systems, we use a PCA transform to reduce the size of the input from 90 dimensions (15 Mel-band classifiers × 6 manner log-posteriors, i.e. linear outputs from MLP band-classifiers) to 60 dimensions. The PCA transform accounts for 99% of the variability in the original 90-dim. feature space. The merging MLP has a single hidden layer of 200 sigmoidal units and 6 output units with a softmax non-linearity. The merging MLP is trained using the same (noisy TIMIT) database and training algorithm as used for the MLP band-classifiers. The estimated log-posterior probabilities (linear outputs) from the merging MLP comprise the second feature stream used in recognition.

4.4. TRAPS latency and complexity

The optimized TRAPS-based system using PCAs for parameter reduction has around 19000 parameters, while pre-
vious (unoptimized) TRAPS-based [1] systems use 72000 or more parameters. The TRAPS-based system has algorithmic latency of 90 ms and its computational complexity is around 1 MOPS.

5. EXPERIMENTS

As part of the advanced front-end standardization, ETSI 1 defined development databases that contain digit strings in six different languages: Aurora-2 (English) and Aurora-3 (Italian, Finnish, Spanish, German and Danish) [8, 10]. Each task covers three conditions: well-matched, medium-mismatched, and high-mismatched conditions. The moderately mismatched condition has only additive noise mismatch, while the highly mismatched condition has both additive noise and channel mismatches between the training and testing datasets [8]. The recognizer is ETSI-specified HTK based whole-word models with 16 states per HMM and 3 diagonal-covariance components per state. For more detail refer [10].

6. RESULTS AND DISCUSSION

We compare the performance (measured in terms of word accuracy) using the noise-robust MFCC features [9] alone with the merged robust MFCC and manner features on the Aurora-2 and Aurora-3 tasks. On Aurora-2 data we observe an improvement of 1.0–1.5% absolute with multi-style training and around 0.9% absolute with clean training data when the manner features are included (Table 2). On Aurora-3 data, we observe an absolute improvement in average recognition performance of 0.4% in the well-matched condition, an improvement of 0.8% in the moderately mismatched condition, and a degradation of 0.4% in the highly mismatched condition when the manner features are included (Table 3). Since the TRAPS classifier was trained on TIMIT which covers only noises present in the multi-style training database of the Aurora-2 task, we believe that the parameters of the TRAPS system do not generalize well for the Aurora-3 mismatched conditions. An overall relative improvement of 53% and 55% over the reference endpointed, unquantized MFCC is achieved in Aurora-2 and Aurora-3 tasks on incorporating TRAPS estimated features, respectively (Tables 4, 5).

7. CONCLUSION

We see consistent performance improvements on the Aurora-2 data when we include TRAPS-estimated manner of articulation features in the recognizer. We see similar

---

1 The European Telecommunication Standard Institute

<table>
<thead>
<tr>
<th>features</th>
<th>I(class; features)</th>
</tr>
</thead>
<tbody>
<tr>
<td>PLP-cepstra</td>
<td>4.41 bits</td>
</tr>
<tr>
<td>PLP-cepstra &amp; manner</td>
<td>4.53 bits</td>
</tr>
</tbody>
</table>

Table 1. Estimated mutual information between manner labels and features. In both cases the feature vector has 39 dimensions.

<table>
<thead>
<tr>
<th>multi-style</th>
<th>clean</th>
</tr>
</thead>
<tbody>
<tr>
<td>robust MFCC</td>
<td>with TRAPS</td>
</tr>
<tr>
<td>Set A</td>
<td>91.06</td>
</tr>
<tr>
<td>Set B</td>
<td>90.61</td>
</tr>
<tr>
<td>Set C</td>
<td>90.10</td>
</tr>
<tr>
<td>overall</td>
<td>90.69</td>
</tr>
</tbody>
</table>

Table 2. Word accuracies (%) on Aurora-2 data, by training data (multi-style vs. clean) and by feature set (noise-robust MFCCs alone, or robust MFCCs augmented with TRAPS-estimated manner features).
Table 3. Word accuracies (%) on Aurora-3 data, by feature set, condition and language. The conditions are well-matched (WM), moderately mismatched (MM), and highly mismatched (HM). The languages are Italian (IT), Finnish (FI), Spanish (SP), German (DE), and Danish (DK).

<table>
<thead>
<tr>
<th></th>
<th>Aurora-3 % Word Error Rate</th>
<th></th>
<th>Aurora-3 % Relative Improvement</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>FI</td>
<td>SP</td>
<td>GE</td>
</tr>
<tr>
<td></td>
<td>Well(x40%)</td>
<td>3.17</td>
<td>2.63</td>
</tr>
<tr>
<td></td>
<td>Mid(x35%)</td>
<td>10.40</td>
<td>5.35</td>
</tr>
<tr>
<td></td>
<td>High(x25%)</td>
<td>14.42</td>
<td>10.38</td>
</tr>
<tr>
<td></td>
<td>Overall</td>
<td>8.51</td>
<td>5.52</td>
</tr>
</tbody>
</table>

Table 4. Relative improvements over the reference end-pointed, unquantized plain MFCC

Table 5. Relative improvements over the reference end-pointed, unquantized plain MFCC


8. *ETSI ES 202 050 v.0.1.1 Speech Processing, Transmission and Quality Aspects (STQ); Distributed speech recognition; Advanced feature extraction algorithm*, April 2002.
