University of Ljubljana System for Interspeech 2011 Speaker State Challenge

Rok Gajišek, Simon Dobrišek, France Mihelič
Faculty of Electrical Engineering, University of Ljubljana, Ljubljana, Slovenia
{rok.gajsek, simon.dobrisek, france.mihelic}@fe.uni-lj.si

Abstract
The paper presents our efforts in the Interspeech 2011 Speaker State Challenge. Both systems, for the Intoxication and the Sleepiness Sub-Challenge, are based on a Universal Background Model (UBM) in a form of a Hidden Markov Model (HMM), and the Maximum A Posteriori (MAP) adaptation. With the combination of our HMM-UBM-MAP derived supervectors and selected statistical functionals from the baseline feature set, we were able to surpass the baseline system in both sub-challenges. By employing majority voting fusion of best systems we were able to further improve the performance. In the Intoxication Sub-Challenge our best result on the test set is 67.46%, and in the Sleepiness Sub-Challenge 71.28%.

Index Terms: Intoxication, Sleepiness, HMM-UBM-MAP

1. Introduction
Analyzing speech in order to recognize gender, age, emotions, or other paralinguistic information has been gaining a lot of attention over the past years. The Interspeech 2011 Speaker State Challenge [1], being a third consecutive Interspeech competition focusing on analysis of paralinguistics, in itself exhibits a trend of increasing interest in this subject. Following The Interspeech 2009 Emotion Challenge [2], which focused on recognizing real-life spontaneous emotions from children’s speech, and The Interspeech 2010 Paralinguistic Challenge [3], which dealt with gender, age and affect recognition, organizers of this year’s challenge provide an excellent ground for assessment of different systems for the task of recognizing intoxication and sleepiness.

Focusing only on the acoustic features, there are two predominant ways of modeling, the supra-segmental and frame-based modeling [4]. The supra-segmental modeling calculates numerous statistical functionals for the given utterance’s frame-level features, combines them in a vector which represent’s a given utterance and forms an input for the classification task. On the other hand, the frame-based modeling tries to estimate the distribution of the frame level features either per each class, or first estimate a Universal Background Model (UBM) [5] from all the train data, and later assess the difference between each individual utterance and the UBM. A Gaussian Mixture Model (GMM) is usually used as the UBM and estimated by Expectation-Maximization (EM) algorithm. A Maximum A Posteriori (MAP) adaptation (where weights and covariances are fixed and only means are adapted) has been successfully applied as a estimation criteria for capturing the difference between the UBM and a particular utterance [5].

In this paper we present our approach to modeling acoustic features by using a Hidden Markov Model (HMM) as a UBM. By MAP adaptation (means only) of the UBM for each sample we derive a new HMM per sample, with the difference between them being in the means of Gaussian distributions. These are then concatenated into a vector and form an input to the SVM classification.

The following chapter describes our HMM-UBM-MAP modeling. In Sections 3 and 4 the Intoxication and the Sleepiness Sub-Challenges are presented, respectively. In the Section 5. results for the development and test sets are given, and in Section 6. we give our conclusions.

2. UBM–MAP modeling using HMMs
Modeling an acoustic feature space using an adaptation of a UBM was first introduced for the task of speaker verification [5] and has since been successfully applied to a number of other paralinguistic recognition tasks, such as emotion recognition [6], [7], age and gender detection [8], affect recognition [9]. The UBM is usually represented by a Gaussian Mixture Model (GMM), which is well known for its ability of representing arbitrary distributions [5]. Since the goal of the UBM is to represent a general acoustic space, the labels of the training samples are not considered. Instead, all training data is pooled together and by Expectation-Maximization (EM) estimation the final UBM is produced. Next, for each utterance in the database a new GMM model is estimated from the UBM by employing Maximum A Posteriori (MAP) adaptation. During the adaptation process, the weights and covariance matrices are fixed in order to capture the difference between a particular sample and the UBM only in the transformed means. For each sample, the corresponding adapted model’s means are concatenated into a vector, forming a set of training vectors for the final Support Vector Machines (SVM) classification. A popular substitution for the GMM is a Hidden Markov Model (HMM), renowned for being successfully applied to speech recognition and having the advantage of incorporating temporal information. However, the drawbacks lie in the computational complexity and selection of an appropriate model structure. Furthermore, we were unable to find a report in the literature stating that choosing a more complex form of likelihood function as a UBM gave an improvement over the standard GMM–UBM (the same observation is given in [5] for the task of text-independent speaker detection). As we show in the Results section, we were able to improve the recognition performance by using a HMM instead of a GMM in the standard UBM-MAP scheme.

In the Intoxication Sub-Challenge the phone level transcriptions were provided for all the data sets. Employing the training set and its corresponding transcriptions, a 3-state left-to-right HMM can be build for each allophone in the database (prior to building the HMM, we combined certain similar allophones as presented in Table 1). If transcriptions are not available for the database a speech recognition system can be employed in order to produce the transcriptions. Combining Gaussian distributions from all phoneme HMM’s states and discarding transitional probabilities, we can construct a GMM–UBM which can
be used with the standard MAP adaptation, described in [5]. This way, instead of having Gaussian distribution determined by some global statistical method such as K-means clustering or Linde-Buzo-Grey algorithm [10], our GMM-UBM relies on allophone segmentation. This increases the possibility that, during MAP adaptation of the UBM for each utterance, identical or at least similar allophones are modifying the same components of the GMM-UBM, and thus capturing the variability between different speaker states on an allophone level. Furthermore, the MAP adaptation for each utterance can be evaluated for the HMM as well [10], provided that transcriptions are available. Thus, our final system employed the MAP adaptation on the HMM, while fixing the covariance matrices, weights and transitions, and modifying only the means of Gaussian distributions in states. These transformed means from all HMM states are then concatenated into a vector, representing the corresponding utterance. By repeating the process of MAP adaptation and concatenation of means into a vector for all train and test samples, we produce a set of feature vectors, representing the train and test samples. The next step of classification is done using the SVMs. Prior to classification, the minority class is inflated using the SMOTE algorithm [11].

For the HMM training and the HMM-MAP adaptation the HTK toolkit was used [10], the frame level features were calculated using the openSMILE feature extractor [12], and Weka [13] was used for the SVM classification.

3. Intoxication Sub-Challenge System

Using our proposed HMM-UBM-MAP procedure a number of different configurations were evaluated using the train and development data sets. Based on these findings the final systems were selected for the evaluation on the test data. In the following section all different experiments are described and discussed.

3.1. Acoustic Features

The acoustic feature set in our HMM-UBM-MAP procedure consisted of 1-12 MFCCs and the RMS Energy, with their first order delta coefficients. We used the openSMILE feature extractor tool [12] to calculate the MFCCs and the RMS Energy for every 20 millisecond long frame (with a 50% overlap), following the configuration used for the baseline feature set [1], producing a 26 dimensional feature vector per frame.

3.2. Modeling and Classification

The availability of transcriptions for all sets of data in the Intoxication Sub-Challenge, enabled use to build a 3-state HMM for each phoneme, with one 26-dimensional Gaussian distribution in every state. Since the initial evaluation was done on the development set, only the train set was used during the training of the HMMs. After MAP adaptation of the HMM for each train and test utterance and the application of SMOTE oversampling on the minority class, the SVM classification was used to evaluate the performance of our HMM-UBM-MAP procedure. Prior to building the HMM, the number of base units (allophones) was reduced by combining similar phones together as presented in the Table 1.

With the merging of the allophones we reduced the list of base units to 47. Having a 26-dimensional Gaussian distribution per state, the length of the MAP adapted vector of means is $47 \times 3 \times 26 = 3666$. We also experimented with modifying the total number of Gaussian distributions. By tying some of the states we reduced the size of the MAP adapted vector of means, which is an input to the SVM classification, but we did not observe any improvement in the recognition performance on the development set. Similarly, by raising the number of Gaussians in the HMM and thus increasing the size of the vector of means, the recognition accuracy on the development set did not improve either.

In order to further improve the performance we considered using the F0 based features from the IS11 baseline low level descriptor set [1] as well. We compared our HMM-UBM-MAP modeling approach with the supra-segmental approach. To avoid the singularities in covariance matrices during EM training we added a Gaussian noise to the frame level features that were based on the voice related LLD. After which we followed the same procedure as described above for MFCCs. Using 47 HMMs with 3 states and a 12-dimensional Gaussian distribution, we produce, after the MAP adaptation, a 1692 dimensional vector.

During evaluation we used the train set for all the training procedures and the development set exclusively for testing. For predicting the labels on the test set, a new system was built following the same steps as before, except the train and the development sets were combined and used for training the final system.

We evaluated different combinations of HMM based UBM systems with different sets of the baseline’s features. For the final submission we fused our two best systems with the baseline on the decision level by using majority voting.

4. Sleepiness Sub-Challenge System

In the Sleepiness Sub-Challenge the transcriptions were not provided, thus we were unable to apply our HMM-based method directly. Nonetheless, since the spoken language of the corpus is the same as in the Intoxication Sub-Challenge, we decided to use the Alcohol Language Corpus (ALC) to train a more complex HMM monophone model, and run the recognition on the Sleepy Language Corpus (SLC) producing transcriptions for all three data sets. This process further evaluates the robustness of our HMM-UBM-MAP approach with regards to the accuracy of the transcriptions.

4.1. Acoustic Features

The same feature sets as described in Sec. 3.1. were used for the Sleepiness Sub-Challenge as well. For the task of phoneme recognition recognition task only, the HTK toolkit’s implementation of MFCCs was used. 0-12 MFCCs were calculated and Cepstral Mean Normalization was applied, after which first and second order delta were added, producing a feature vector of size 39.

4.2. Modeling and Classification

In order to evaluate our HMM-UBM-MAP method we had to obtain transcriptions for the SLC database. Using the HTK toolkit and the ALC dataset we build a 3-state left-to-right monophone HMMs with the combined number of 2994 Gauss-
sians in all the states. Using a bi-gram language model, trained on the ALC dataset, we ran the recognition task on all three SLC datasets. We cannot evaluate the phoneme recognition accuracy on the SLC database, however we can report that the recognition accuracy on the ALC’s dataset, which was used for training, was 51.8%. We can assume that the recognition accuracy was significantly lower for the SLC database due to the mismatch in the acoustic environment between the ALC and SLC corpuses, and the ALC’s reported recognition result being produced on the same data the model was trained on.

Once transcriptions are obtained we can follow a similar procedure as described in the Intoxication Sub-Challenge (Sec. 3.2.). By MAP adaptation of the HMM-UBM model, produced on the train set, a new set of phoneme HMM is estimated per each train and test sample. By combining the means of all states in each sample adapted HMM, we produce an input vector for the SVM classification. As for the Intoxication Sub-Challenge, different number of Gaussian distributions per state in the HMM were evaluated. In contrast to the Intoxication Sub-Challenge, we were able to improve the recognition accuracy on the development set, by reducing the total number of Gaussians in the HMM, as presented in Sec. 5.

Different combinations of feature level fusion between our HMM-UBM-MAP derived features and the baseline subsets were evaluated on the development set. Best individual systems were chosen for evaluation on the test set, as well as fusion of our two best systems and the baseline, by using majority voting.

5. Results

In both Sub-Challenges, different systems and system configurations were evaluated first on the development set, with the train set used for training. After which, the best systems were selected for evaluation on the test set. Here, we combined the train set with the development set for training the system (both for GMMs-UBM estimation, as well as SVM classification) and produce predictions for the test set. All classification tests were done using Weka’s SMO classifier with the same parameter setup as used for the baseline [1].

5.1. Intoxication Sub-Challenge

Results of the initial evaluation on the development set, as well as final results on the test set, for the Intoxication Sub-Challenge are shown in Table 2. First three systems are all based on our HMM-UBM-MAP modeling scheme with only the MFCCs used as acoustic features. The difference lies in the total number of Gaussian distributions used in the states in the HMM. As described in Sec. 3.2 we started with one Gaussian per state (3 states * 47 phonemes = 141), which yields a MAP adapted feature vector of means of size 3666. As it can be seen in Table 2, an increase or decrease in number of Gaussians does not improve the recognition accuracy. We already surpassed the baseline result, but since our method was based only on MFCCs, we were expecting some improvement when adding the F0 based features. First, we added, to our MAP derived feature vector, all the voice related LLD statistical features. This improved the result on the development set to 68.1% unweighted average (UA) recall. However, the best result on development set was achieved by combining our HMM-UBM-MAP derived features with all non-mfcc based IS11 baseline features. The baseline features set comprised 4368 features of which 888 are based on 1-12 MFCCs and their deltas. After excluding those, we took 3480 statistical features from the baseline and added our 3666 features, summing to a final feature vector of size 7146. This combination of features gave the best result of 70.8% UA average recall. It should be noted that since we use SMOTE technique [11] for balancing the minority class in the train data set, the results vary during multiple tests using the same setup.

The system with highest result on the development set was selected for the evaluation on the test set and a 66.6% UA recall was achieved. With this result we surpassed the baseline, but by a smaller amount as one would expect with regards to the results on the development set. During evaluation on the development set we noticed a consistently better recall for the NAL (non-alcoholized) class. In the best system the NAL recall was 80% with AL recall only 61.3%. Therefore, the second system we evaluated on the test set had the minority class (AL) oversampled, by producing 170% synthetic samples (still using SMOTE algorithm) creating a slight bias towards the AL class. As can be seen in the results table, this gave an increase of 0.5% absolute. The final attempt on improving the score was with fusion of our 2 best systems (presented in Table 2) and the baseline system. Predictions were combined using majority voting and another increase of 0.3 % was achieved setting our best UA recall at 67.46%.

<table>
<thead>
<tr>
<th>Sub-Challenge</th>
<th>Intoxication</th>
<th>Features</th>
<th>Train vs. Develop</th>
<th>Train + Develop vs. Test</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>baseline</td>
<td>65.3</td>
<td>69.2</td>
</tr>
<tr>
<td></td>
<td></td>
<td>HMM-UBM-MAP 141</td>
<td>67.6</td>
<td>71.0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>HMM-UBM-MAP 100</td>
<td>67.1</td>
<td>70.6</td>
</tr>
<tr>
<td></td>
<td></td>
<td>HMM-UBM-MAP 200</td>
<td>67.0</td>
<td>70.8</td>
</tr>
<tr>
<td></td>
<td></td>
<td>HMM-UBM-MAP 141</td>
<td>68.1</td>
<td>73.7</td>
</tr>
<tr>
<td></td>
<td></td>
<td>HMM-UBM-MAP 141 + baseline*</td>
<td>70.8</td>
<td>74.4</td>
</tr>
</tbody>
</table>

Table 2: Intoxication Sub-Challenge results for development set and test set.

An HMM-UBM-MAP procedure was evaluated on voice related LLD from the baseline features, calculated on a frame level bases, but results were not satisfactory, thus they were not considered for the final system.

5.2. Sleepiness Sub-Challenge

Following the successful application of our HMM-UBM-MAP procedure in the Intoxication Sub-Challenge similar systems were evaluated in the Sleepiness Sub-Challenge. Since transcriptions were not provided here, and we generated them ourselves by training a simple monophone recognizer on the ALC database, the improvement over the baseline here is lower. Still we are able to show that our type of modeling the MFCCs is superior to the statistical approach, even if the phoneme transcriptions are inaccurate. Initial test with the HMM-UBM-MAP procedure using MFCCs and 141 Gaussian distributions gave worst result than the IS11 feature set baseline. Next, we evaluated changing the number of Gaussians. As with the Intoxication Sub-Challenge, increasing them lead to poorer per-
formance, but interestingly, with the tying of some states, and reducing the number of Gaussians to 100, we noticed a slight increase in performance. Modeling the voice related LLD features with the HMM-UBM-MAP method and combining them with the MFCC based HMM-UBM-MAP features lead to only a slight improvement in precision, but no improvement in UA recall. Finally, the baseline was surpassed on the development set with the combination of the HMM-UBM-MAP derived features based on the frame level MFCCs and the baseline features set, stripped of statistical features based on MFCCs. The same setup gave the best result in the Intoxication Sub-Challenge as well.

Table 3: Sleepiness Sub-Challenge results for development set and test set.

<table>
<thead>
<tr>
<th>Sub-Challenge</th>
<th>Sleepiness</th>
<th>% UA</th>
<th>% WA</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Train vs. Develop</td>
<td></td>
<td></td>
</tr>
<tr>
<td>baseline</td>
<td>67.3</td>
<td>69.1</td>
<td></td>
</tr>
<tr>
<td>HMM-UBM-MAP 141</td>
<td>65.7</td>
<td>63.6</td>
<td></td>
</tr>
<tr>
<td>HMM-UBM-MAP 100</td>
<td>65.9</td>
<td>64.2</td>
<td></td>
</tr>
<tr>
<td>HMM-UBM-MAP 100 (mfcc+f0)</td>
<td>65.9</td>
<td>64.6</td>
<td></td>
</tr>
<tr>
<td>HMM-UBM-MAP 100 + baseline*</td>
<td>67.8</td>
<td>67.8</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Train + Develop vs. Test</td>
<td></td>
<td></td>
</tr>
<tr>
<td>baseline</td>
<td>70.3</td>
<td>73.0</td>
<td></td>
</tr>
<tr>
<td>HMM-UBM-MAP 100 (mfcc+f0)</td>
<td>68.2</td>
<td>71.4</td>
<td></td>
</tr>
<tr>
<td>HMM-UBM-MAP 100 + baseline*</td>
<td>70.7</td>
<td>73.1</td>
<td></td>
</tr>
<tr>
<td>Majority voting fusion</td>
<td>71.3</td>
<td>74.0</td>
<td></td>
</tr>
</tbody>
</table>

*Baseline feature set without the MFCC based functionals.

With the MFCC and the voice related features, modeled by the HMM-UBM-MAP scheme, the performance on the test set was below the baseline with UA recall at 68.22%. But combination of the MFCCs, modeled by the HMM-UBM-MAP scheme, and the non-MFCC based statistical features from the IS11 baseline, gave an improvement of 0.37% absolute. Interestingly, the improvement over the baseline result is consistent with the results on the development set. Similiary to the Intoxication Sub-Challenge, the best result was achieved by majority voting fusion of our two best systems (presented in Table 3) and the baseline.

6. Conclusion

In the paper we presented our efforts in the 2011 Interspeech Speaker State Challenge. Systems for both, the Intoxication and the Sleepiness Sub-Challenge are presented. Both are based on the combination of our HMM-UBM-MAP approach of modeling the MFCCs, and the statistical modeling of various other spectral and voice related features. We consistently outperformed the IS 2011 baseline systems on both the development, as well as the final test set. Especially in the Intoxication Sub-Challenge, where transcriptions were available, we showed that the HMM-UBM-MAP approach significantly outperforms the supra-segmental statistical modeling. In the Sleepiness Sub-Challenge, where phoneme transcriptions were inaccurate, we still managed to get an improvement over the baseline by the HMM-UBM-MAP modeling of the MFCC features.

7. References