Gender and Affect Recognition Based on GMM and GMM–UBM modeling with relevance MAP estimation

Rok Gajšek¹, Janez Žibert², Tadej Justin¹, Vitomir Štruc³, Boštjan Vesnicer³, France Mihelič¹

¹Faculty of Electrical Engineering, University of Ljubljana, Slovenia
²Department of Information Sciences and Technology, University of Primorska, Slovenia
³Alpineon Research and Development, Ljubljana, Slovenia

rokgajsek@fe.uni-lj.si, janez.zibert@upr.si

Abstract

The paper presents our efforts in the Gender Sub-Challenge and the Affect Sub-Challenge of the INTERSPEECH 2010 Paralinguistic Challenge. The system for the Gender Sub-Challenge is based on modeling the Mel-Frequency Cepstrum Coefficients using Gaussian mixture models, building a separate model for each of the gender categories. For the Affect Sub-Challenge we propose a modeling schema where a universal background model is first trained an all the training data and then, employing the maximum a posteriori estimation criteria, a new feature vector of means is produced for each particular sample. The feature set used is comprised of low level descriptors from the baseline system, which in our case are split into four sub sets, and modeled by its own model. Predictions from all sub-systems are fused using the sum rule fusion. Aside from the baseline regression procedure, we also evaluated the Support Vector Regression and compared the performance. Both systems achieve higher recognition results on the development set compared to baseline, but in the Affect Sub-Challenge our system’s cross correlation is lower than that of the baseline system, although the mean linear error is slightly superior. In the Gender Sub-Challenge the unweighted average recall on the test set is 82.84%, and for the Affect Sub-Challenge the cross-correlation on the test set is 0.39 with mean linear error of 0.143.

Index Terms: emotion recognition, affect recognition, gender recognition, GMM–UBM, MAP

1. Introduction

The need for incorporating the paralinguistic information in modern human–computer interaction (HCI) systems, on their path toward becoming more user friendly and appearing more human-like, is eminent. Without taking into consideration this implicit (paralinguistic) channel of communication [1], the computer systems will stay perceived by humans as artificial and intrinsically different from how we are used to perceive interaction among ourselves. In human to human interaction we are used to heaving our gender, age group, even our emotional state at the time, recognized and taken into consideration during the exchange of information. Hence, we can expect that these three types of information, which also represent the three sub-challenges of the INTERSPEECH 2010 Paralinguistic Challenge [2], can provide a valuable information which, when incorporated in to the HCI systems, can improve the naturalness of such systems.

In the paper we present our efforts in the Gender and the Affect Sub-Challenges of the INTERSPEECH 2010 Paralinguistic Challenge. Both are based on modeling a frame level based features, in the gender detection sub-challenge using Gaussian Mixture Models (GMMs), and in the affect sub-challenge using GMMs as Universal Background model (UBM) with maximum a posteriori (MAP) estimation criteria. The latter, has been successfully applied to the speaker identification/verification systems [3], and since our past research have lead us to believe that the phenomena of emotions in speech is represented in the acoustical signal in a similar way as are the variations between different speakers [4], [5], our goal was to evaluate the performance of such a system for the affect recognition task. Furthermore, a similar system [6] proved effective in the INTERSPEECH 2009 Emotion Challenge [7]. In the Gender Sub-Challenge the same procedure is evaluated and compared to modeling each classification category with its own GMM.

First, in Sections 2 and 3 each system is described in detail, including the experiments conducted, then in Section 4 results for both, the development set and the “blinded” test set are reported, and finally we give our conclusions.

2. Affect Sub-Challenge

In the Affect sub-challenge our system employs a well known procedure in speaker recognition and verification task [3]. First, a universal background model (UBM) in a form of a Gaussian Mixture Model (GMM) is trained on all the samples from the training set. Then, for each sample, a so-called super vector is extracted by transforming the means of the UBM model based on the maximum a posterior (MAP) criteria. This super vector of transformed means forms the new feature vector for the corresponding data.

2.1. Feature set

For the Affect Sub-Challenge the baseline feature set is comprised of the most prominent and widely used features. Therefore, we decided to use the same low level features as were used in the baseline system, but instead of calculating the functionals and thus, creating one feature vector for each sample, we extract the features on a frame level bases. For the calculation of all the different features we employ the openSMILE component of the openEAR toolkit [8]. The low level features are then split into 4 sub sets:

- I. Prosodic features: F0, F0 Envelope, Voicing Prob., Jitter, Shimmer and PCM loudness.
- II. Mel Frequency Cepstral Coefficients (MFCC) [0-14]
- III. Linear Spectral Frequency pairs (LSP) [0-7]
- IV. Log Mel Frequency Bands (LMFB)[0-7]
For each separate feature the first order deltas are added, yielding a 76 dimension feature vector. Each subset of features is then modeled on its own as described in the following section.

2.2. GMM–UBM modeling

For the UBM modeling, first, a generative model of speech needs to be selected. Two most prominent candidates are GMM and hidden Markov models (HMM). In speaker identification/verification it was shown that HMMs do not offer any advantage over the GMMs, except in the case of strong prior knowledge of the spoken text which can be incorporated in HMM modeling. Since we will not be focusing on recognizing the text we select the GMMs.

The unimodal $D$-dimensional Gaussian density $i$ is given by

$$p_i(x) = \frac{1}{(2\pi)^{D/2}|\Sigma_i|^{1/2}} \exp\left[-\frac{1}{2}(x - \mu_i)^T \Sigma_i^{-1}(x - \mu_i)\right],$$

(1)

where $\mu_i$ is a vector of means and $\Sigma_i$ is a covariance matrix. A mixture Gaussian density is then a weighted linear combination of the $M$ unimodal densities

$$p(x|\lambda) = \sum_{i=1}^{M} w_i p_i(x),$$

(2)

with each Gaussian weighted according to the corresponding $w_i$. The model parameters, means, covariances and weights, are represented for $i$ component of the model as $\lambda = \{\mu_i, \Sigma_i, w_i\}$.

The parameters of the UBM are estimated based on the maximum likelihood (ML) criteria via the expectation–maximization (EM) algorithm [9] on all the training data, in an attempt to cover as much variability as we can. Prior to running the EM algorithm, the initial model of $M$ components is estimated by employing the k-means clustering.

2.3. Relevance MAP

Once the UBM is acquired we use the maximum a posteriori (MAP) estimation criteria (as described in [3]) to adopt the UBM model for each sample. First, the probabilistic alignment of a particular sample $Pr(i|x)$, against all $M$ UBM components is determined according to

$$Pr(i|x) = \frac{w_i p_i(x)}{\sum_{j=1}^{M} w_j p_j(x)}.$$  

(3)

where $p_i(x_i)$ is the density probability function of $x_i$ feature vector for the $i$-th component of the GMM, and $\omega_i$ are the corresponding component’s weights. Next, we calculate the sufficient statistics for updating the means. In general the MAP estimation updates the means, variances and weights of the GMM, but in our system we only focus on updating the mean values of GMMs, hence the statistics required for the Maximization step are

$$n_i = \sum_{t=1}^{T} Pr(i|x_t)$$

(4)

and

$$E_i(x) = \frac{1}{n_i} \sum_{t=1}^{T} Pr(i|x_t)x_t.$$  

(5)

So far the procedure is identical to the Expectation step when using ML criteria in the EM algorithm. The new mean values for the GMM are updated using Eq. (6).

$$\mu_i = \alpha_i^m \mu_i + (1 - \alpha_i^m) \bar{x}_i,$$  

(6)

The adaptation parameter $\alpha_i^m$ which controls the balance between the old values of means and the new estimate is computed as

$$\alpha_i^m = \frac{n_i}{n_i + r},$$  

(7)

where $r$ is the relevance factor, which is the same for all components of the GMM. The value the of relevance factor is chosen experimentally and it usually lays in the interval between 8 and 16. After sufficient iterations of the described procedure are reached, either the means stop changing or the maximum number of iterations is exceeded, we are left with the final vector of the means. The size of this final vector equals the dimension of the original feature vector multiplied with the number of components of the GMM and thus is increasing with the number of GMM components. Also, the amount of available data for training needs to be taken into consideration, as insufficient statistics can lead to singularities in certain components of the GMM when using the maximum likelihood criteria, if the number of components is too big. The effect of different number of components is described in the experiments section 2.5.

2.4. Regression methods

The baseline regression method using Weka’s [10] REP-Trees with Random-Sub-Space learning performs well with our set of MAP estimated features. But since the dimension of our features, when increasing the number of components in the GMM, becomes bigger then the baseline’s feature size, and support vector machines (SVMs) are known to perform well when dealing with large dimension features, we assess the performance of support vector regression (SVR). Using the toolkit LibSVM [11] we evaluated the epsilon-SVR with the kernel in a form of a radial basis function.

The overall results when using SVR did not differ significantly from the baseline regression. Therefore, a combination of both classifiers, depending on the particular feature set, was selected for the final system.

2.5. Experiments

Following the experimental protocol, development and assessment were done using the train and development sets, respectively.

In the Affect Sub-Challenge, for each of the four feature sub sets, described in Section 2.1, a UBM model was trained. For some recordings the pitch detector failed to detect any voiced parts, hence the pitch was zero in all the frames of the corresponding sample. Such cases were discarded when training an UBM on prosodic feature sub-set.

In general, a higher number of GMM components better represent the diversity of the feature space but can also lead to over fitting as well as singularity problems when lacking variability in the training data. Therefore, for each feature subset multiple UBM’s were build with the number of components ranging from 8 to 128. The effect of different types of covariance matrices were also taken into consideration producing two UBM models with diagonal and full covariance matrix for a particular number of components.

After applying the MAP estimation on the means, producing feature vectors for train and development set, both baseline regression and SVR were employed, generating predictions for the development set. Finally, predictions from all four sub-systems were combined using the sum rule fusion. For estimating the fusion parameters a 4-fold cross validation was done on the development set over all possible combinations of different
sub-systems (different number of GMM components, diagonal or full covariance matrix and different regression techniques). For the samples were no voiced frames were detected, three sub-systems (without the “prosodic” sub-system) were fused, again estimating new fusion parameters using a 4-fold cross validation over the development set. A combination of UBM modeling and the use of functionals, as presented in the baseline, for each particular type of feature sub-set was evaluated.

Following the results on the developments set, a final system was trained on the combined train and evaluation data sets, producing predictions for the blinded test set.

3. Gender Sub-Challenge

We performed only gender detection in the Age and Gender Sub-Challenge. Our system was based on Gaussian Mixture Models (GMMs). We trained separately three GMMs, one on each subset of “aGender” corpus [2] corresponding to three provided gender types.

We exploited only acoustic information as was also the case in the baseline system [2]. The acoustic features were based on the mel-frequency cepstral coefficients (MFCCs), whereas in our system 12 MFCCs and short-time energy were extracted from the waveforms, added by corresponding first-order derivatives (ΔMFCCs). The MFCCs were computed by using HTK Toolkit [12]. In addition, cepstral mean- and variance-normalization (CMVN) was performed and silent regions were removed on file-by-file basis. The silent regions were detected by inspecting short-time energy. Those frames that belonged to the lowest 33-percentile of the estimated frame-based short-time energy distribution were discarded from training. The estimation of the GMM parameters was performed separately on three training subsets belonging to each gender. We used 512-mixture, full-covariance GMMs. The GMMs initializations were done by Linde-Buzo-Gray algorithm [12].

Evaluations were performed on the MFCC and the ΔMFCC features, that were obtained from the test waveforms in the same manner as in the training phase. The feature vectors were estimated from each waveform and the log-likelihoods of the data were computed for each gender-based GMM. The data were then classified to the gender, that had the highest log-likelihood score.

Note, that we also built an alternative gender-detection system based on universal background model, MAP adaptations and SVM classification, but the obtained results were not as good as the presented results.

4. Results

4.1. Affect Sub-Challenge

The results for the Affect Sub-Challenge are presented in Table 1. The combination of 4 sub-systems that gave the best cross-correlation coefficient (CC) on the development data is comprised of:

- I. Prosodic: GMM–UBM, 8 components, diagonal covariance, SVR
- II. MFCC: GMM–UBM, 16 components, full covariance, baseline regression
- III. LSP: GMM–UBM, 16 components, diagonal covariance, baseline regression
- IV. LMFB: GMM–UBM, 8 components, full covariance, SVR

It should be noted that although the CC achieved by this system on the development set is superior to the baseline, the mean linear error (MLE) is higher than what is achieved with the baseline. Fusion of some other combinations of sub-systems gave lower MLEs, but since the competition measure is CC, we decided to use for the test set, the system with the superior CC result. Unfortunately, our system’s CC on the test data is 0.390 which is lower than baseline result. Interestingly, the MLE is slightly superior than that of the baseline system. Since the labels for the test set are not known we are unable to analyze where lies the difference in performance between the development and the test results. Obviously, having done a 4-fold cross validation for determining the parameters of fusion on development set could lead to values that are over fitted to the development set. Seeing how the results are noticeably lower for the test set, as compared to the development set, for baseline and our system, we can presume that there is a difference between the test set on one hand and development and train sets on the other. This would imply that our system lacks robustness as oppose to the baseline. Using more training data (incorporating other databases in the train set) for estimating the UBM could help in overcoming the issue of over fitting the general model of feature variability to the original training set. This will be evaluated in the future.

Table 1: Affect Sub-Challenge results

<table>
<thead>
<tr>
<th>Sub-Challenge</th>
<th>CC</th>
<th>MLE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Train vs. Develop</td>
<td>0.604</td>
<td>0.118</td>
</tr>
<tr>
<td>Affect - baseline</td>
<td>0.630</td>
<td>0.123</td>
</tr>
<tr>
<td>Train + Develop vs. Test</td>
<td>0.421</td>
<td>0.146</td>
</tr>
<tr>
<td>Affect - baseline</td>
<td>0.390</td>
<td>0.143</td>
</tr>
</tbody>
</table>

4.2. Gender Sub-Challenge

The evaluation results of the Gender Sub-Challenge are presented in Tables 2 and 3. The results are provided for two tasks, as is suggested in [2]: in the first task GMMs were trained on training data and tested on development data, while in the second task GMMs were built from the training and development data and tested on the test data.

Table 2 shows a confusion matrix of gender recognition on the development set, while the training was performed on the training set of the Gender Sub-Challenge data. The recognition accuracy (recall) of a male speech (m) is above 96%, for a female speech (f) is 88%, and for a children speech (x) is 56%. As expected, the results show a reliable detection of male speech, while there exists a significant confusion of female and children speech.

Table 2: The confusion matrix of a gender detection on a development subset of the Gender Sub-Challenge data. The results are presented in %.

<table>
<thead>
<tr>
<th>Recognized</th>
<th>m</th>
<th>f</th>
<th>x</th>
</tr>
</thead>
<tbody>
<tr>
<td>True</td>
<td>96.18</td>
<td>3.57</td>
<td>0.25</td>
</tr>
<tr>
<td>m</td>
<td>11.85</td>
<td>31.62</td>
<td>56.53</td>
</tr>
<tr>
<td>f</td>
<td>5.57</td>
<td>88.24</td>
<td>6.19</td>
</tr>
</tbody>
</table>
Slightly better results were observed in the second task, where GMMs were built on training and development sets and tested on test data. The achieved recalls in this task were 64%, 91% and 93% for children, female and male speech respectively. The results show quite an improvement of the detection of children and female speech in comparison to the evaluation results on the development data, while the detection of male speech was slightly worse. The improvement of the results was expected due to the larger training set.

Table 3: Gender Sub-Challenge ULJ_FE results by gender-based GMMs trained on MFCC and ΔMFCC features.

<table>
<thead>
<tr>
<th>Sub-Challenge</th>
<th>Task</th>
<th>% UA</th>
<th>% WA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Train vs. Develop</td>
<td>Baseline</td>
<td>77.28</td>
<td>84.60</td>
</tr>
<tr>
<td>Gender</td>
<td>{x,f,m}</td>
<td>80.32</td>
<td>87.83</td>
</tr>
<tr>
<td>Train + Develop vs. Test</td>
<td>Baseline</td>
<td>81.21</td>
<td>84.81</td>
</tr>
<tr>
<td>Gender</td>
<td>{x,f,m}</td>
<td>82.84</td>
<td>87.32</td>
</tr>
</tbody>
</table>

Table 3 shows the overall recognition results of the Gender Sub-Challenge by unweighted and weighted accuracy on average per class. The results are slightly better than the results of the baseline system, presented in [2].

5. Conclusion

In the paper we presented our work for the INTERSPEECH 2010 Paralinguistic Challenge. We participated in 2 of the sub-challenges, the Gender Sub-Challenge and the Affect Sub-Challenge. Our system for gender detection is based on modeling the MFCC features with GMMs, using a separate model for each category. For the affect Sub-Challenge we used a GMM–UBM approach with MAP criteria for final estimation. Results produced by our systems on the development set, are superior to the baseline results for both sub-challenges. In the Gender Sub-Challenge the test results from our system stay superior to the baseline, as for the Affect Sub-Challenge the cross correlation of the baseline system is higher. The results from our GMM based systems do not differ drastically from the baseline systems, we can conclude that modeling frame based features with GMMs gives similar performance as using the functionals of the frame based features.

6. References


