Using Multiple Databases for Training in Emotion Recognition: To Unite or to Vote?

Björn Schuller, Zixing Zhang, Felix Weninger, and Gerhard Rigoll

Institute for Human-Machine Communication, Technische Universität München, Germany

{schuller|zixing.zhang|weninger|rigoll}@tum.de

1. Introduction

Cross-corpus emotion recognition—that is, attempting to build classifiers that generalize across application scenarios and acoustic conditions—is highly relevant for engineering of speech emotion recognition systems ‘in the wild’. While there exists an increasing amount of available emotional speech data, optimal ways have still to be found to use it for training of models that generalize very well. The crux, however, is that these data usually come with completely different emotion inventories reaching from Ekman’s ‘big six’ to task specific ones.

Surprisingly, it has been only recently that this task has been adopted, and first results suggest that it is indeed very challenging, not only due to differences on signal level, but particularly also the type of emotion elicitation (e.g., acted emotion vs. spontaneous, non-prototypical emotion) and emotion model used for annotation. While the dimensional space offers us the ability to ‘translate’ these models into, e.g., arousal and valence dimensions, the question arises whether these data are then best ‘put into one bag’ by agglomerating the newly labeled data for training, or, whether it is better to use several classifiers or regressors, one per database, and classify unseen test data based on a majority vote.

First studies exist on enhancing the robustness of cross-corpus emotion recognition by data agglomeration, i.e., combining several emotional speech corpora within the training set [1] and by that reducing the data scarcity problem and extending the variety of acoustic background. In [2], normalization approaches to the speaker, corpus, or both have been exploited to mitigate the divergence between training and testing sets. In this paper, we want to investigate the potential of such data ‘pooling’ [1, 2] as opposed to individually trained machine learners per database and subsequent majority voting on unseen test data instances.

2. Six Emotional Speech Databases

As databases, we chose six among the most frequently used that range from acted over induced to spontaneous affect portrayal. For better comparability of obtained performances among corpora, we additionally map the diverse emotion groups onto the two most popular axes in the dimensional emotion model as in [2, 3]: arousal (i.e., passive (“-“) vs. active (“+”)) and valence (i.e., negative (“-“) vs. positive (“+”)). These mappings are not straight forward—we favor better balance among target classes. We further discretized into the four quadrants (q) 1–4 of the arousal-valence plane for continuous labeled corpora. In the following, each set is shortly introduced including the mapping to binary arousal/valence by “+“ and “-“ per emotion and its number of instances.

The Danish Emotional Speech (DES) database [4] contains professionally acted nine Danish sentences, two words, and chunks that are located between two silent segments of two passages of fluent text. Emotions contain angry (+/-, 85), happy (+/+, 86), neutral (-/-, 85), sadness (-/-, 84), and surprise (+/+). The cINTERFACE (cINTER) [5] corpus consists of recordings of naive subjects from 14 nations speaking pre-defined spoken content in English. The subjects listened to six successive short stories eliciting a particular emotion out of angry (+/-, 215), disgust (-/-, 215), fear (+/-, 215), happy (+/+, 207), sadness (-/-, 210), and surprise (+/+). The Airplane Behaviour Corpus (ABC) [6] is based on induced mood by pre-recorded announcements of a vacation (return) flight, consisting of 13 and 10 scenes. It contains aggressive (+/-, 95), cheerful (+/+, 105), intoxicated (+/-, 33), nervous (+/-, 93), neutral (-/-, 79), and tired (-/-, 25) speech. The Audiovisual Interest Corpus (AVIC) [7] consists of spontaneous speech and natural emotion. In its scenario setup, a product presenter leads subjects through a commercial presentation. AVIC is labelled in “level of interest” (loi) 1–3 having loi1 (-/-, 553), loi2 (+/+, 2279), and loi3 (+/+). The Belfast Sensitive Artificial Listener (SAL) data is part of the final HUMAINE database. We consider the subset used, e.g., in [8] with an average length of 20 minutes per speaker of natural human-SAL conversations. The data has been labeled continu-
We employ acoustic feature vectors of 6,552 dimensions using our open source openEAR toolkit [10] by 39 functionalities of 56 acoustic Low-Level Descriptors (LLDs) including first and second order delta regression coefficients. This feature set corresponds to the “emo-large” configuration delivered with the openEAR toolkit for straightforward reproducibility. Table 3 summarizes the statistical functionals which were applied to the LLDs shown in Table 2 to map a time series of variable length onto a static feature vector.

Further details on the corpora are summarized in Table 1 and found in [3]. Note that in the ongoing, the class distribution of the training partition is balanced by up-sampling (all) instances of the minority class.

### 3. Acoustic Features and Classifiers

We employ acoustic feature vectors of 6,552 dimensions using our open source openEAR toolkit [10] by 39 functionalities of 56 acoustic Low-Level Descriptors (LLDs) including first and second order delta regression coefficients. This feature set corresponds to the “emo-large” configuration delivered with the openEAR toolkit for straightforward reproducibility. Table 3 summarizes the statistical functionals which were applied to the LLDs shown in Table 2 to map a time series of variable length onto a static feature vector.

As to the selection of classifier, we consider Support Vector Machines (SVM) which can provide very good generalization properties and are presently likely the most used classifier in emotion recognition. Further, Random forests (RF), an ensemble learning algorithm combining un-pruned decision tree learners in randomized feature sub-spaces [11] were decided for. Finally, we decided for Naïve Bayes (NB) to complement our chosen classifiers. Thus, for representative results in our experiments, we chose SVM with linear Kernel, complexity 0.05, and pairwise multi-class discrimination based on Sequential Minimal Optimization, RF with 10 trees, and NB as classifiers. Implementations in the Weka toolkit [12] were used for further reproducibility.

### 4. Data Fusion: Voting vs. Pooling

Our goal is to exploit information from different training corpora to enable robust cross-corpus emotion recognition. In particular, for our choice of databases five training databases are available for cross-corpus training and testing in a leave-one-corpus-out manner. We implemented two strategies for data fusion: First, we trained individual classifiers from single databases, and combined their class decisions in late fusion by majority voting. Second, we combined the training material from different databases for training a single classifier; this strategy will be referred to as pooling in the following. Note that since the number of training databases is odd, the majority vote is always well-defined. In addition, we considered (sums of) confidence scores of the RF classifier in the voting process, which arguably leads to more reliable decisions than fusion of binary decisions, as obtained by SVM and NB.
5. Experimental Results

5.1. Data Fusion vs. Single Classifier

We evaluated our experiments in terms of unweighted accuracy (UA) on each test corpus, which is the unweighted average of class-wise recall for each of the ‘positive’ and ‘negative’ classes. In contrast to weighted average recall, UA seems to be better suited to the unbalanced class distribution found in the considered corpora (see Table 1). UA has been the official competition measure for the INTERSPEECH 2009 Emotion Challenge [13]. We compare the results of fusion by pooling and voting against the results obtained by pairwise cross-corpus emotion recognition, as has been investigated, e.g., in [2]: There, classifiers are simply trained on a single database and tested on another. In particular, we also calculated the average UA for each test database obtained in pairwise classification. These evaluation procedures represent fully realistic conditions where the classifier cannot simply adapt to the peculiarities of a single database as in ‘traditional’ intra-corpus evaluation.

Results for each test database, and different classifiers (top: SVM, middle: RF, bottom: NB) are depicted in Figure 1 as one-dimensional scatter plots. The average performance of pairwise cross-corpus recognition is depicted by circles, while triangles indicate the UA obtained by voting and diamonds the one obtained by pooling. Exact values are given in Table 4. On average, it can be seen that both, voting and pooling are superior to the average UA for pairwise classification. Thus, on average one expects a gain by fusing databases instead of selecting the single one that performs best. Comparing the results by data fusion to the individual single-corpus results, data fusion seems most promising for valence recognition, where it often outperforms the best single-corpus classifier. On the other hand, this trend is not as strongly visible in arousal recognition. Still, from the application point of view, this is very interesting: When designing an emotion recognition system, it will be unknown which training database performs best. In that case, using multiple training databases and fusing decisions can dispense with the need that the classifier cannot simply adapt to the peculiarities of a single database as in ‘traditional’ intra-corpus evaluation.

5.2. Pooling vs. Voting

As to comparison of fusion strategies with one another, it seems that pooling (63.4 % UA on average over all test databases) generally outperforms voting (54.4 %) for the SVM classifier, even drastically for the arousal recognition on the DES database (75.7 % vs. 56.2 % UA). However, this can neither be observed for the RF, nor for the NB classifier. A possible explanation might be that the SVM training algorithm automatically weights training instances by selecting them as support vectors; thus, it seems more suited to training on large, heterogeneous datasets. On the other hand, voting with random forests outperforms the voting from other classifiers significantly, both, for valence and arousal; this probably indicates that using confidence scores indeed increases robustness of the voting strategy. Finally though, on average over classifiers, pooling is superior to both, single-database classification of arousal and voting, delivering a relative improvement of UA by 9.0 % over the former. For valence recognition, pooling and voting perform almost equally, and voting is observed slightly better.

5.3. Two-Stage Voting

As the above evaluation revealed very notable differences in the performance of different classifiers, we investigated the performance of a two-stage voting process, where a secondary majority vote among the three classifiers is performed. Again, this majority vote is well-defined in any case. From Table 4, it can be seen that for recognition of arousal, the two-stage vote is on average slightly inferior to the voting by random forests (59.0 % vs 60.0 % UA); for valence, the accuracy of two-stage voting (58.1 %) is even equal to the best possible configuration of classifier and fusion strategy (voting by random forests).

This result, in fact, suggests that when designing an emotion recognizer from multiple databases using fusion by voting – in that case, it is not clear a priori which classifier performs best – a majority vote among classifiers delivers almost equal accuracy to the best classifier. Thus, it will be an interesting issue for future research to evaluate the two-stage scheme for pooled training as well.

6. Conclusions

We proposed and investigated two novel voting strategies to improve cross-corpus acoustic emotion recognition by combination of multiple training databases and classifiers. The results showed that the suggested strategies considerably surpass performance of cross-corpus recognition systems based on single training corpora, which is especially interesting for design of emotion
recognizers ‘in the wild’: It suggests that if it is unknown a priori which kind of training data is best suited to the scenario at hand, fusing a variety of training data is on average better than relying on a single training corpus. Concerning majority voting of individually trained learners as opposed to data agglomeration (pooling) in a single classifier, results largely depend on the classifier architecture. It seems that inclusion of additional selection of suitable training instances such as in [14] will be a powerful tool to boost the performance of data pooling. A very remarkable performance of 63.4% unweighted accuracy in recognition of arousal by SVM is obtained across six databases by data agglomeration.

Summarizing, while we were able to significantly improve the performance of cross-corpus emotion recognition in this study, it remains a challenging field due to the severe diversity not only of acoustic conditions, speakers, content, etc., but foremost also to the diversity of original labeling of data in diverse classes that were mapped to binary arousal and valence classes. Our future efforts will focus on improving confidence measures by not only building on classification scores, but also weighting the vote by finding measures of suitability or similarity of databases to test instances.

7. Acknowledgement

Zixing Zhang’s work is supported by a research grant of the People’s Republic of China. This work has been partly supported by the Federal Republic of Germany through the German Research Foundation under grant no. SCHU 2508/2-1.

8. References


Table 4: Unweighted Accuracy (UA) for cross-corpus binary arousal / valence classification: Average UA of single database classifiers (Avg), and UA of cross-corpus fusion by classifier voting and data pooling, for SVM, Random Forests (RF) and Naive Bayes (NB). Mean UA across classifier type, and UA of two-stage multi-classifier vote (2-Vote).

<table>
<thead>
<tr>
<th>UA [%]</th>
<th>Test on</th>
<th>SVM Vote Pool</th>
<th>Avg</th>
<th>RF Vote Pool</th>
<th>Avg</th>
<th>NB Vote Pool</th>
<th>Avg</th>
<th>Mean Vote Pool</th>
<th>Avg</th>
<th>2-Vote</th>
</tr>
</thead>
<tbody>
<tr>
<td>Arousal</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ABC</td>
<td>50.1</td>
<td>49.2</td>
<td>63.8</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AVIC</td>
<td>50.0</td>
<td>47.4</td>
<td>60.1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DES</td>
<td>54.7</td>
<td>56.2</td>
<td>75.7</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>eNTER</td>
<td>57.4</td>
<td>61.6</td>
<td>62.2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SAL</td>
<td>54.1</td>
<td>59.2</td>
<td>62.4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>VAM</td>
<td>53.2</td>
<td>53.0</td>
<td>55.9</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean</td>
<td>53.3</td>
<td>54.4</td>
<td>63.4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Valence</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ABC</td>
<td>54.0</td>
<td>50.9</td>
<td>60.5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AVIC</td>
<td>51.7</td>
<td>56.6</td>
<td>61.0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DES</td>
<td>50.9</td>
<td>51.7</td>
<td>57.3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>eNTER</td>
<td>52.2</td>
<td>54.8</td>
<td>56.9</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SAL</td>
<td>51.1</td>
<td>52.2</td>
<td>54.8</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>VAM</td>
<td>51.3</td>
<td>49.4</td>
<td>54.9</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean</td>
<td>51.9</td>
<td>52.6</td>
<td>56.7</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

1556