Using multiple databases for training in emotion recognition: to unite or to vote?

Zixing Zhang, Felix Weninger, Björn Schuller, Gerhard Rigoll

Angaben zur Veröffentlichung / Publication details:

Zhang, Zixing, Felix Weninger, Björn Schuller, and Gerhard Rigoll. 2011. "Using multiple databases for training in emotion recognition: to unite or to vote?" In *INTERSPEECH 2011 - 12th Annual Conference of the International Speech Communication Association, Florence, Italy, August 27-31. 2011*, edited by Piero Cosi, Renato De Mori, Giuseppe Di Fabbrizio, and Roberto Pieraccini, 1553–56. ISCA Archive.
https://doi.org/10.21437/Interspeech.2011-468.

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

Abstract

We present an extensive study on the performance of data agglomeration and decision-level fusion for robust cross-corpus emotion recognition. We compare joint training with multiple databases and late fusion of classifiers trained on single databases, employing six frequently used corpora of natural or elicited emotion, namely ABC, AVIC, DES, eNTERFACE, SAL, VAM, and three classifiers i. e. SVM, Random Forests, Naïve Bayes to best cover for singular effects. On average over classifier and database, data agglomeration and majority voting deliver relative improvements of unweighted accuracy by 9.0 % and 4.8 %, respectively, over single-database cross-corpus classification of arousal, while majority voting performs best for valence recognition.

Index Terms: Emotion Recognition, Data Agglomeration, Cross-Corpus

10.21437/Interspeech.2011-468

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 all '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. We perform our evaluation on six selected databases that are among the most frequently used in the field.

We structured the remainder of this contribution as follows: The data sets for experimentation are described and the mapping of their original and diverse emotion labeling to binary arousal and valence tags is detailed out in Sec. 2. Then, the acoustic feature brute-forcing by our openEAR toolkit and classifier selection will be presented in Sec. 3. Sec. 4 introduces our strategies for optimal exploitation of multiple datasets for optimal classification results as implemented in the experiments which are described in Sec. 5. Finally, we conclude in Sec. 6.

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 (+/+, 79). The eNTERFACE (eNTER) [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 (+/+, 215). 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 (+/+, 170). 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-

Table 1: Overview of the selected emotion corpora (Lab: labelers, Rec: recording environment, f/m: (fe-)male subjects).

Corpus	Language	Speech	Emotion	# Arousal		# Valence		# All	h:mm	# m	# f	# Lab	Rec	kHz
				-	+	-	+							
ABC	German	fixed	acted	104	326	213	217	430	1:15	4	4	3	studio	16
AVIC	English	free	natural	553	2449	553	2449	3002	1:47	11	10	4	studio	44
DES	Danish	fixed	acted	169	250	169	250	419	0:28	2	2	–	studio	20
eNTER	English	fixed	induced	425	852	855	422	1277	1:00	34	8	2	studio	16
SAL	English	free	natural	884	808	917	779	1692	1:41	2	2	4	studio	16
VAM	German	free	natural	501	445	875	71	946	0:47	15	32	6/17	noisy	16

Table 2: 33 Low-Level Descriptors (LLD) used.

Feature Group	Features in Group
Raw Signal	Zero-crossing-rate
Signal energy	Logarithmic
Pitch	Fundamental frequency F_0 in Hz via Cepstrum and Autocorrelation (ACF). Exponentially smoothed F_0 envelope.
Voice Quality	Probability of voicing ($\frac{ACF(T_0)}{ACF(0)}$)
Spectral	Energy in bands 0–250 Hz, 0–650 Hz, 250–650 Hz, 1–4 kHz 25 %, 50 %, 75 %, 90 % roll-off point, centroid, flux, and rel. pos. max. / min.
Mel-spectrum	Band 1–26
Cepstral	MFCC 0–12

Table 3: 39 functionals applied to LLD contours.

Functionals	#
Respective rel. position of max./min. value	2
Range (max.-min.)	1
Max. and min. value - arithmetic mean	2
Arithmetic mean, Quadratic mean, Centroid	3
Number of non-zero values	1
Geometric, and quadratic mean of non-zero values	2
Mean of absolute values, Mean of non-zero abs. values	2
Quartiles and inter-quartile ranges	6
95 % and 98 % percentile	2
Std. deviation, variance, kurtosis, skewness	4
Zero-crossing rate	1
# of peaks, mean dist. btwn. peaks, arth. mean of peaks, arth. mean of peaks - overall arth. mean	4
Linear regression coefficients and error	4
Quadratic regression coefficients and error	5

ously in real time with respect to valence and activation using a system based on FEELtrace. The annotations were normalized to zero mean globally and scaled so that 98 % of all values are in the range from -1 to +1. The 25 recordings have been split into turns using an energy based Voice Activity Detection. Labels for each obtained turn are computed by averaging over the complete turn. Per quadrant the samples are: q1 (+/+, 459), q2 (-/+, 320), q3 (-/-, 564), and q4 (+/-, 349). Finally, the *Vera-Am-Mittag* (VAM) corpus [9] consists of recordings taken from a German TV talk show. The audio recordings were manually segmented to the utterance level, whereas each utterance contained at least one phrase. The labeling bases on a discrete five point scale for valence, activation, and dominance. Samples among quadrants are q1 (+/+, 21), q2 (-/+, 50), q3 (-/-, 451), and q4 (+/-, 424).

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

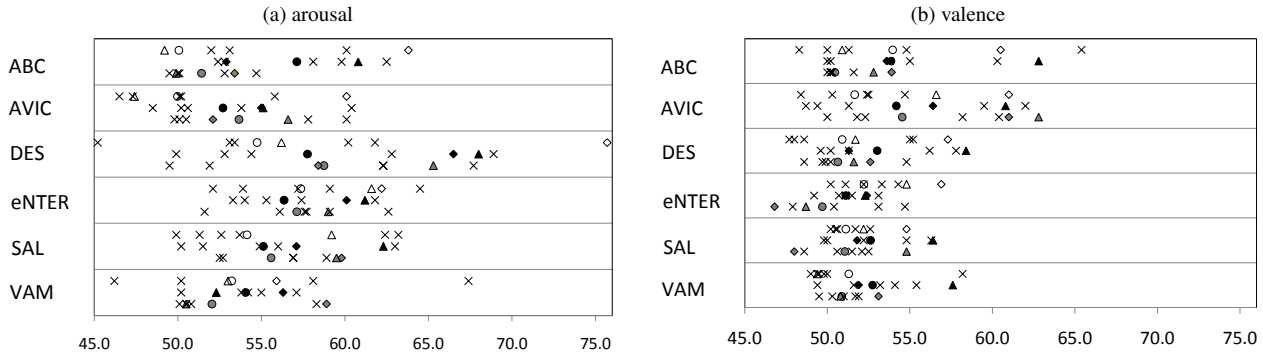
We employ acoustic feature vectors of 6 552 dimensions using our open source openEAR toolkit [10] by 39 functionals of 56 acoustic Low-Level Descriptors (LLDs) including first and second order delta regression coefficients. This feature set corresponds to the "emo-large" configuration delivered with 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.

Figure 1: *Distributions of unweighted accuracies for cross-corpus binary arousal / valence classification of six test databases: Single-database classifiers (crosses), average of single-database classifiers (circles), and classifier fusion by voting (triangles) and pooling (diamonds). The top row per test database depicts results obtained by SVM, the middle one by RF, and the bottom one by 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-corpora 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 dispose of the need for extensive validation experiments with different training sets.

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

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).*

UA [%]	**SVM**			**RF**			**NB**			**Mean**			**2-Vote**
Test on	**Avg**	**Vote**	**Pool**	**Avg**	**Vote**	**Pool**	**Avg**	**Vote**	**Pool**	**Avg**	**Vote**	**Pool**	
Arousal													
ABC	50.1	49.2	63.8	57.1	60.8	52.9	51.4	49.9	53.4	52.9	53.3	56.7	55.5
AVIC	50.0	47.4	60.1	52.7	55.1	55.0	53.7	56.6	52.1	52.1	53.0	55.7	53.9
DES	54.7	56.2	75.7	57.8	68.0	66.5	58.7	65.3	58.4	57.1	63.2	66.9	68.7
eNTER	57.4	61.6	62.2	56.4	61.2	60.1	57.1	59.0	59.1	56.9	60.6	60.5	61.3
SAL	54.1	59.2	62.4	55.1	56.3	57.1	55.6	59.5	59.8	55.0	60.3	59.8	63.3
VAM	53.2	53.0	55.9	54.1	52.3	56.3	52.0	50.5	58.9	53.1	51.9	57.0	51.0
Mean	53.3	54.4	**63.4**	55.5	**60.0**	58.0	54.8	56.8	**57.0**	54.5	57.1	**59.4**	59.0
Valence													
ABC	54.0	50.9	60.5	53.9	62.8	53.6	50.5	52.8	53.9	52.8	55.5	56.0	61.0
AVIC	51.7	56.6	61.0	54.2	60.8	56.4	54.5	62.8	61.0	53.5	60.1	59.5	65.7
DES	50.9	51.7	57.3	53.0	58.4	51.3	50.6	51.6	52.6	51.5	53.9	53.7	58.2
eNTER	52.2	54.8	56.9	51.1	52.3	52.4	49.7	48.7	46.8	51.0	51.9	52.0	52.2
SAL	51.1	52.2	54.8	52.6	56.4	51.8	51.1	54.8	48.0	51.6	54.5	51.5	56.7
VAM	51.3	49.4	49.4	52.7	57.6	51.9	50.9	50.8	53.1	51.6	52.6	51.5	54.8
Mean	51.9	52.6	**56.7**	52.9	**58.1**	52.9	51.2	**53.6**	52.6	52.0	**54.7**	54.0	58.1

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

[1] I. Lefter, L. J. M. Rothkrantz, P. Wiggers, and D. A. van Leeuwen, "Emotion recognition from speech by combining databases and fusion of classifiers," in *Proc. of Text and Speech and Dialogue*, Berlin, German, 2010.

[2] B. Schuller, B. Vlasenko, F. Eyben, M. Wöllmer, A. Stuhlsatz, A. Wendemuth, and G. Rigoll, "Cross-Corpus Acoustic Emotion Recognition: Variances and Strategies," *IEEE Transactions on Affective Computing*, vol. 1, pp. 119–131, 2010.

[3] B. Schuller, B. Vlasenko, F. Eyben, G. Rigoll, and A. Wendemuth, "Acoustic Emotion Recognition: A Benchmark Comparison of Performances," in *Proc. IEEE ASRU*, Merano, Italy, 2009, pp. 552–557.

[4] I. S. Engbert and A. V. Hansen, "Documentation of the Danish Emotional Speech Database DES," Center for PersonKommunikation, Aalborg University, Denmark, Tech. Rep., 2007.

[5] O. Martin, I. Kotsia, B. Macq, and I. Pitas, "The eNTERFACE'05 Audio-Visual Emotion Database," in *Proc. IEEE Workshop on Multimedia Database Management*, Atlanta, 2006.

[6] B. Schuller, M. Wimmer, D. Arsic, G. Rigoll, and B. Radig, "Audiovisual behavior modeling by combined feature spaces," in *Proc. ICASSP 2007*, vol. II. Honolulu, Hawaii, USA: IEEE, 2007, pp. 733–736.

[7] B. Schuller, R. Müller, F. Eyben, J. Gast, B. Hörnler, M. Wöllmer, G. Rigoll, A. Höthker, and H. Konosu, "Being Bored? Recognising Natural Interest by Extensive Audiovisual Integration for Real-Life Application," *Image and Vision Computing Journal*, vol. 27, pp. 1760–1774, 2009.

[8] M. Wöllmer, F. Eyben, S. Reiter, B. Schuller, C. Cox, E. Douglas-Cowie, and R. Cowie, "Abandoning emotion classes - towards continuous emotion recognition with modelling of long-range dependencies," in *Proc. 9th Interspeech 2008*. Brisbane, Australia: ISCA, 2008, pp. 597–600.

[9] M. Grimm, K. Kroschel, and S. Narayanan, "The Vera am Mittag German Audio-Visual Emotional Speech Database," in *Proc. IEEE ICME*, Hannover, Germany, 2008, pp. 865–868.

[10] F. Eyben, M. Wöllmer, and B. Schuller, "openEAR – Introducing the Munich Open-Source Emotion and Affect Recognition Toolkit," in *Proc. Affective Computing and Intelligent Interaction (ACII)*. Amsterdam, The Netherlands: IEEE, 2009.

[11] T. K. Ho, "The Random Subspace Method for Constructing Decision Forests," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, pp. 832–844, 1998.

[12] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The WEKA data mining software: An update," *SIGKDD Explorations*, vol. 11, no. 1, 2009.

[13] B. Schuller, S. Steidl, and A. Batliner, "The INTERSPEECH 2009 Emotion Challenge," in *Proc. Interspeech*. Brighton, UK: ISCA, 2009, pp. 312–315.

[14] C. E. Erdem, E. Bozkurt, E. Erzin, and A. T. Erdem, "RANSAC-based training data selection for emotion recognition from spontaneous speech," in *Proc. of the 3rd international workshop on Affective Interaction in Natural Environments (AFFINE)*, Firenze, Italy, 2010, pp. 9–14.