# Selecting Training Data for Cross-Corpus Speech Emotion Recognition: Prototypicality vs. Generalization

*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 investigate strategies for selection of databases and instances for training cross-corpus emotion recognition systems, that is, systems that generalize across different labelling concepts, languages and interaction scenarios. We propose objective measures for prototypicality based on distances in a large space of brute-forced acoustic features and show their relation to the expected performance in cross-corpus testing. We perform extensive evaluation on eight commonly used corpora of emotional speech reaching from acted to fully natural emotion and limited phonetic content to conversational speech. In the result, selecting prototypical training instances by the proposed criterion can deliver a gain of up to 7.5 % unweighted accuracy in cross-corpus arousal recognition, and there is a correlation of .571 between the proposed prototypicality measure of databases and the expected unweighted accuracy in cross-corpus testing by Support Vector Machines.

## 1. Introduction

Cross-corpus emotion recognition from speech requires systems that generalize across different types of labelling, different languages, and different type of data reaching from acted to fully natural emotion and limited phonetic content to conversational speech. While there is a growing amount of emotional speech data available, the question arises how to best exploit them. The dimensional space offers us the ability to map data labelled in emotions as different as joyful and anxious on dimensions as arousal and valence for unified modelling. Still, the challenge remains to find pattern recognition techniques that generalize across interaction and application scenarios.

In this paper, we investigate optimal techniques for selecting training data in cross-corpus emotion recognition. Methods for pruning atypical instances from training have been thoroughly explored in pattern recognition [1] and particularly speech emotion recognition [4]; still, such experiments are usually limited to training and testing on the same data set. On the other hand, first studies on feature selection in cross-corpus emotion recognition suggest that training optimizations do not always generalize across different data sets [5]. In this paper, our goal is to find objective measures for databases and instances that are correlated with the expected accuracy in cross-corpus emotion recognition. Particularly, we deal with the question whether selecting the most 'prototypical' instances and databases for model building enables generalization across corpora.

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 labelling to binary arousal and valence tags is detailed out in Sec. 2. Then, the acoustic feature brute-forcing by our openEAR toolkit and classifier training are briefly presented in Sec. 3. Sec. 4 introduces our strategies for database and instance selection and describes experimental results. Finally, we conclude in Sec. 5.

## 2. Eight Emotional Speech Databases

As databases, we chose eight 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 [12, 14]: arousal (i. e., passive ("-") vs. active ("+")) and valence (i. e., negative ("-") vs. positive ("+")). These mappings are not straightforward—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 [3] 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 *Berlin Emotional Speech Database* (EMOD) [2] features professional actors speaking ten emotionally undefined sentences. 494 phrases are commonly used: angry (+/-, 127), boredom (-/-, 79), disgust (-/-, 38), fear (+/-, 55), happy (+/+, 64), neutral (-/+, 78), and sadness (-/-, 53). The eNTERFACE (eNTER) [10] corpus consists of recordings of naive sub-

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

| Corpus | Lang. | Speech | Emot. | # 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 |
| **EMOD** | German | fixed | acted | 248 | 246 | 352 | 142 | 494 | 0:22 | 5 | 5 | – | studio | 16 |
| **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 |
| **SUSAS** | English | fixed | natural | 701 | 2892 | 1616 | 1977 | 3593 | 1:01 | 4 | 3 | – | noisy | 8 |
| **VAM** | German | free | natural | 501 | 445 | 875 | 71 | 946 | 0:47 | 15 | 32 | 6/17 | noisy | 16 |

jects 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) [13] 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 *Speech Under Simulated and Actual Stress* (SUSAS) database [9] serves as a first reference for spontaneous recordings. Speech is additionally partly masked by field noise in the chosen actual stress speech samples recorded in subject motion fear and stress tasks. SUSAS content is restricted to 35 English air-commands in the speaker states high stress (+/-, 1 202), medium stress (+/-, 1 276), neutral (-/+, 701), and scream (+/-, 414). The *Audiovisual Interest Corpus* (AVIC) [11] 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 contains natural conversations between humans and virtual agents. Per quadrant the samples are: q1 (+/+, 459), q2 (-/+, 320), q3 (-/-, 564), and q4 (+/-, 349). Finally, the *Vera-Am-Mittag* (VAM) corpus [7] consists of recordings taken from a German TV talk show. 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 [5, 12]. Note that in the ongoing, training data is balanced by up-sampling to unit class distribution.

## 3. Acoustic Features and Classifier

We employ acoustic feature vectors of 6 552 dimensions extracted by our open source openEAR toolkit [6], applying 39 functionals to 56 acoustic Low-Level Descriptors (LLDs) including first and second order delta regression coefficients. LLDs comprise spectral, cepstral, voice quality, and pitch features. For straightforward reproducibility, the feature set corresponds to the "emo-large" configuration delivered with the openEAR toolkit[1]. As classifier, we use Support Vector Machines constructed by Sequential Minimal Optimization with a complexity of 0.05. The implementation in the Weka toolkit [8] was used for further reproducibility.

| UA [%] | Test on 7 remaining databases | | | | | |
|---|---|---|---|---|---|---|
| Train on | Arousal | | | Valence | | |
| | min | max | mean | min | max | mean |
| ABC | 52.5 | 73.6 | 59.8 | 47.8 | 58.5 | 53.3 |
| AVIC | 55.0 | 66.6 | 59.5 | 43.7 | 56.6 | 51.7 |
| DES | 58.8 | 80.4 | 66.6 | **49.2** | **64.1** | **54.8** |
| EMOD | 54.9 | 72.9 | 62.5 | 45.6 | 60.5 | 51.3 |
| eNTER | 51.1 | 68.4 | 60.0 | 48.9 | 57.9 | 54.3 |
| SAL | 54.1 | 76.7 | 63.8 | 47.0 | 57.8 | 51.4 |
| SUSAS | 52.2 | 69.5 | 57.1 | 47.1 | 56.3 | 51.7 |
| VAM | **60.6** | **80.6** | **67.7** | 48.8 | 51.3 | 50.2 |

Table 2: Min(imum), max(imum) and mean unweighted accuracy (UA) when training on one of 8 databases and testing on the 7 remaining databases. Binary classification by linear SVM.

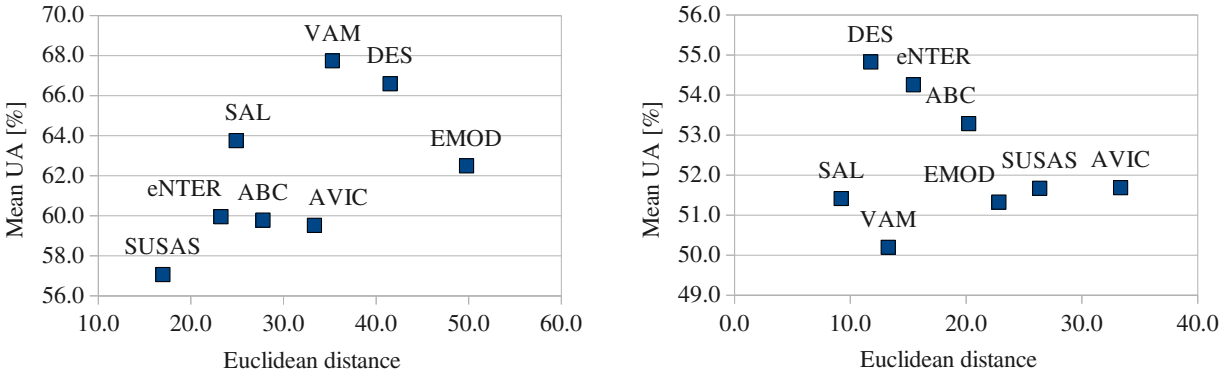
## 4. Database and Instance Selection

We evaluated our experiments in terms of unweighted accuracy (UA), which is the average recall of the '+' and '-' classes, as we are dealing with (sometimes heavily) unbalanced classification problems (cf. Table 1). The feature space was z-normalized to zero mean and unit variance for each corpus.

### 4.1. Database Selection

For each of the eight databases, we evaluated the performance in terms of UA in cross-corpus evaluation on the seven other databases. Results are shown in Table 2. For

[1] http://www.openaudio.eu




(a) arousal +/-

(b) valence +/-

Figure 1: Training database selection: Mean unweighted accuracy (UA) in cross-corpus testing—i. e., training on one database and testing on the remaining seven—and its relation to the Euclidean distance of class centers of '+' and '-' instances (after z-normalization), for arousal (a) and valence (b).

arousal—interestingly—training with the VAM database of spontaneous, natural speech yields highest average UA (67.7 %), minimum UA (60.6 % on SUSAS), and maximum UA (80.6 % on DES). The second best training corpus is the DES database of acted emotions (66.6 % mean UA). In contrast to arousal, recognition of valence seems to be very challenging, resulting in no more than 54.8 % average UA, which is achieved by training with DES. In fact, it is well known that valence recognition from purely acoustic features is challenging, and even more so in cross-corpus testing.

In the following, we investigated the relation between the 'prototypicality' of a database and the expected UA in cross-corpus emotion recognition when using that database for training. As a measure of prototypicality, we calculated the Euclidean distance $d$ of the class center of 'positive' instances, $\overline{\mathbf{x}}_+ = \mathrm{E}\{\mathbf{x}_+\}$, and the one of 'negative' instances, $\overline{\mathbf{x}}_- = \mathrm{E}\{\mathbf{x}_-\}$. Figure 1a shows the results for recognition of positive and negative arousal. Generally, training with databases that exhibit large distances between positive and negative classes delivers higher UA; notably, this prototypicality in the feature space does not exactly correspond to the notion of acted vs. spontaneous emotion: Consider the similar prototypicality measure of the VAM and DES databases. Furthermore, training on the highly prototypical EMOD only delivers mediocre results (62.5 %), seemingly due to insufficient generalization. Overall, the Spearman (rank) correlation between $d(\overline{\mathbf{x}}_+, \overline{\mathbf{x}}_-)$ and the mean UA is $\rho = .571$, which is however not of statistical significance due to the small sample size (8).

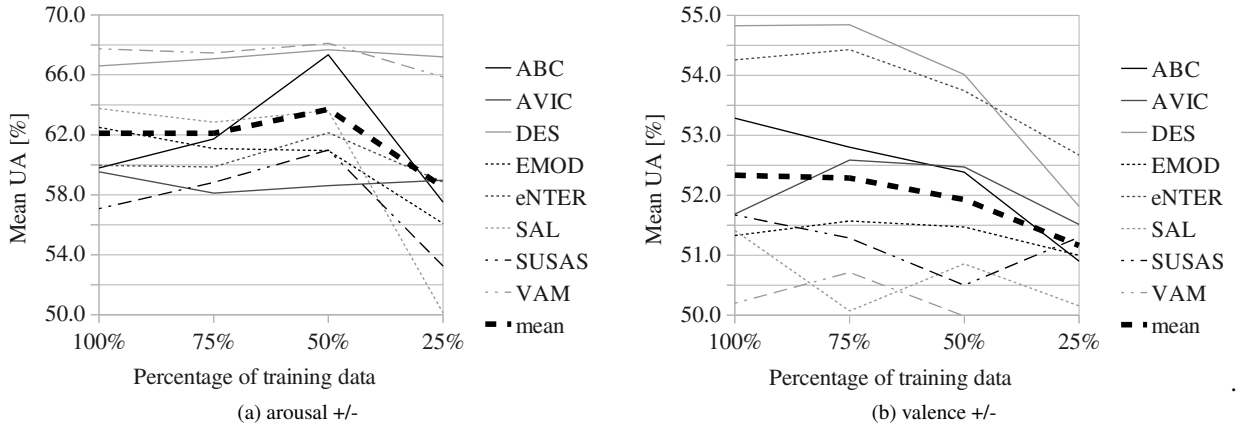
Analogously, results for valence recognition are shown in Figure 1b. As opposed to the arousal case, for valence there is no clear trend as to whether one can expect a gain by using more prototypical databases as training data (Spearman's $\rho = -.060$). Still, the lower recognition rates compared to arousal are reflected in generally smaller distance between the class centers.

### 4.2. Instance Selection

In a second experiment, we evaluated the effect of restricting the training to prototypical instances, for each database. To this end, we computed for each positive instance $\mathbf{x}_+$ the distance to the class center of the negative instances, $d(\mathbf{x}_+, \overline{\mathbf{x}}_-)$. Then, we computed the quartiles of the distribution of $d(\mathbf{x}_+, \overline{\mathbf{x}}_-)$ and selected only the instances corresponding to the fourth quartile (in other words, the 25 % most prototypical positive instances). An analogous procedure was followed for selection of negative instances, which were selected according to the distance $d(\mathbf{x}_-, \overline{\mathbf{x}}_+)$ from the positive class center. For each database, the selected positive and negative instances were joined, and the resulting model was evaluated on the seven other databases. We repeated the experiment using the 50 % (quartiles 3 and 4) and 75 % (quartiles 2–4) most prototypical instances, respectively.

Results are shown in Figure 2. For arousal recognition (Figure 2a), one gains almost 2 % absolute UA on average across the eight databases when using only the 50 % most prototypical instances for training—this result suggests that the 'manual' process of instance selection is complementary with the instance weighting performed in SVM optimization. The improvement is most visible for training with the ABC database, where an absolute gain of 7.5 % UA is achieved. However, generally a drop in performance occurs when further restricting the amount of training data.

Furthermore, cross-corpus *valence* recognition (Figure 2b) cannot generally (i. e., on average) be improved by selecting training instances using the proposed method, despite slight UA gains for the eNTER, EMODB, ABC and VAM databases.




(a) arousal +/-

(b) valence +/-

Figure 2: Training instance selection: Relation between mean UA in cross-corpus testing—i. e., training on one database and testing on the remaining seven—and the amount of training data chosen in order of prototypicality, measured as the Euclidean distance from the class center of the opposite class.

## 5. Conclusions

We have introduced methods for database and instance selection to better exploit the increasing amount of training material available for emotion recognition. Foremost, we could demonstrate that instance selection for binary arousal recognition by our prototypicality measure—based on Euclidean distance from the opposite class center—generalizes across eight commonly used databases of emotional speech. Furthermore, we have shown an effective method to compute prototypicality of databases, which is correlated with average performance in cross-corpus arousal recognition. Yet, these trends are not reflected in cross-corpus *valence* recognition: Neither could instance selection improve performance, nor could a correlation be found between accuracy and database prototypicality.

Future work should focus on unsupervised methods for selection of training data that are suitable for unsupervised or semi-supervised learning from emotional speech.

## 6. Acknowledgments

Zixing Zhang received funding from a research grant by 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.

## 7. References

[1] A. Angelova, Y. Abu-Mostafa, and P. Perona. Pruning training sets for learning of object categories. In *Proc. of CVPR*, pages 494–501, San Diego, CA, USA, 2005.

[2] F. Burkhardt, A. Paeschke, M. Rolfes, W. Sendlmeier, and B. Weiss. A Database of German Emotional Speech. In *Proc. Interspeech*, pages 1517–1520, Lisbon, 2005.

[3] I. S. Engbert and A. V. Hansen. Documentation of the Danish Emotional Speech Database DES. Technical report, Center for PersonKommunikation, Aalborg University, Denmark, 2007.

[4] 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 '10, pages 9–14, New York, NY, USA, 2010. ACM.

[5] F. Eyben, A. Batliner, B. Schuller, D. Seppi, and S. Steidl. Cross-corpus classification of realistic emotions—some pilot experiments. In *Proc. 3rd International Workshop on EMOTION (satellite of LREC): Corpora for Research on Emotion and Affect*, pages 77–82, Valletta, Malta, May 2010. ELRA.

[6] 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)*, pages 576–581, Amsterdam, The Netherlands, 2009. IEEE.

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

[8] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten. The weka data mining software: An update. *SIGKDD Explorations*, 11(1), 2009.

[9] J. Hansen and S. Bou-Ghazale. Getting Started with SUSAS: A Speech Under Simulated and Actual Stress Database. In *Proc. EUROSPEECH-97*, volume 4, pages 1743–1746, Rhodes, Greece, 1997.

[10] 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.

[11] 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*, 27:1760–1774, 2009.

[12] 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*, 1(2):119–131, 2010.

[13] B. Schuller, M. Wimmer, D. Arsic, G. Rigoll, and B. Radig. Audiovisual behavior modeling by combined feature spaces. In *Proc. ICASSP 2007*, volume II, pages 733–736, Honolulu, Hawaii, USA, 2007. IEEE.

[14] A. Stuhlsatz, C. Meyer, F. Eyben, T. Zielke, G. Meier, and B. Schuller. Deep Neural Networks for Acoustic Emotion Recognition: Raising the Benchmarks. In *Proc. of ICASSP*, pages 5688–5691, Prague, Czech Republic, 2011. IEEE.