Unsupervised Learning in Cross-Corpus Acoustic Emotion Recognition

Zixing Zhang, Felix Weninger, Martin Wöllmer, Björn Schuller

Institute for Human-Machine Communication, Technische Universität München
D-80333 München, Germany
{zixing.zhang|weninger|woellmer|schuller}@tum.de

***Abstract*—One of the ever-present bottlenecks in Automatic Emotion Recognition is data sparseness. We therefore investigate the suitability of unsupervised learning in cross-corpus acoustic emotion recognition through a large-scale study with six commonly used databases, including acted and natural emotion speech, and covering a variety of application scenarios and acoustic conditions. We show that adding unlabeled emotional speech to agglomerated multi-corpus training sets can enhance recognition performance even in a challenging cross-corpus setting; furthermore, we show that the expected gain by adding unlabeled data on average is approximately half the one achieved by additional manually labeled data in leave-one-corpus-out validation.**

***Index Terms*—speech emotion recognition, unsupervised learning**

I. Introduction

Obtaining large amounts of realistic data for speech emotion recognition is currently considered one of the most important issues in the field [1]. In fact, it is an ever-lasting belief in pattern recognition that 'there is no data like more data'. Yet, compared to automatic speech recognition where many corpora comprise hundreds of hours of transcribed speech, databases annotated in emotional categories are still sparse—in particular publicly available ones. Semi-supervised and unsupervised learning can be a promising approach to remedy the issue of data sparsity: Assuming sufficiently robust automatic emotion recognition engines, unlabeled data can be classified and integrated into an iterative re-training process. Such unlabeled data is practicably available in 'infinite' amount: One could not only profit from many existing conversational speech corpora that contain emotional colouring, but add data from the media. Notably, studies dealing with unsupervised adaptation of acoustic and language models in automatic speech recognition [2], [3] suggest that addition of unlabeled data in training is competitive with labeled data, even more so if one considers the enormous efforts usually required for manual annotation of speech data. Further, in speech recognition, recent real-life studies as the Google Voice Search show that unsupervised learning has in fact already turned into common practice. As a rule of thumb, roughly ten times the amount of unlabeled data is needed there in comparison to labeled data in order obtain the same gain as with labeled data.

So far, first studies in unsupervised learning for the related field emotion recognition in speech show promising results [4], [5], [6]; yet, these studies are usually limited to single target domains respectively application scenarios. In contrast, in this contribution we aim at a large-scale cross-corpus study such as recently carried out in [7], [8]. Evaluation in cross-corpus experiments, 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'. Peculiar challenges of cross-corpus emotion recognition are that emotional corpora usually come with completely different emotion inventories reaching from Ekman's 'big six' to task specific ones, and that they differ not only on the acoustic level, but particularly also the type of emotion elicitation (e. g., acted emotion vs. spontaneous, non-prototypical emotion).

The crucial question in our study is not whether adding training data by unsupervised learning yields a performance improvement—this has been theoretically proven [9] and repeatedly confirmed in practice [4], [2], [3]. Instead, we investigate how agglomeration of unlabeled data compares with agglomeration of labeled data in cross-corpus emotion recognition, as data agglomeration in general has been proven useful for this task [8], [10].

In the remainder of this paper, we describe the six emotional speech databases that we used for evaluation (Section II). Particularly, we describe how we use a dimensional representation to 'translate' (i. e., map) various emotion models into binary arousal and valence dimensions in order to allow for data agglomeration. We further briefly describe our brute-force large-space extraction of acoustic features and the chosen classifier set-up in Section III. Then, we investigate optimal intra- and inter-corpus normalization for our cross-corpus experiments before evaluating agglomeration of unlabeled and labeled data in Section IV before concluding this paper in Section V.

II. Six Emotional Speech Databases

From the most frequently used publicly available databases, we choose six well known corpora, i. e., ABC, AVIC, DES, eNTERFACE, SAL, and VAM. These corpora cover a broad variety of data from acted speech (DES) over simulated emotions (ABC, eNTERFACE) to spontaneous emotions (AVIC, VAM), and from strictly limited textual context (DES) over more variation (eNTERFACE) to full variance (AVC, AVIC, SAL, VAM). Three languages (English, German, and Danish) belonging to the same family of Germanic languages are contained. Nevertheless, the speaker characteristics, the recording conditions, as well as annotators vary greatly among these databases. An overview over these six corpora is shown in table I. In the following, we briefly introduce the six databases.

A. ABC

The Airplane Behaviour Corpus (ABC) [11] is an audiovisual emotion database. It is crafted for the special target application of public transport surveillance. In order to induce a certain mood, a script was used, which lead the subjects through a guided storyline: Prerecorded announcements by five different speakers were automatically played back controlled by a hidden test-conductor. As a general framework, a vacation flight was chosen, consisting of several scenes such as start, serving of wrong food, turbulences, falling asleep, conversation with a neighbor, or touch-down. The general setup consisted of an airplane seat for the subject, positioned in front of a blue screen. Eight subjects in gender balance from 25 to 48 years (mean 32 years) took part in the recording. The language throughout recording is German. A total of 11.5 h video was recorded and annotated independently after pre-segmentation by three experienced male labelers within a closed set. The average length of the 431 clips is 8.4 s.

B. AVIC

The Audiovisual Interest Corpus (AVIC) [12] is another audiovisual emotion corpus. In its scenario setup, a product presenter leads one of 21 subjects (10 female) through an English commercial presentation. The level of interest is annotated for every turn reaching from boredom (subject is bored with listening and talking about the topic, very passive, does not follow the discourse; this state is also referred to as level of interest (loi) 1, i. e., loi1), over neutral (subject follows and participates in the discourse, it cannot be recognized, if she/he is interested or indifferent in the topic; loi2) to joyful interaction (strong wish of the subject to talk and learn more about the topic; loi3). Additionally, the spoken content and non-linguistic vocalizations are labeled in the AVIC set. For our evaluation we use all 3 002 phrases, in contrast to only 996 phrases with high inter-labeler agreement as e. g. employed in [12].

C. DES

The Danish Emotional Speech (DES) [13] database has been chosen as one of the ‘traditional representatives’ for our study, because it is easily accessible and well annotated. The data used in the experiments are nine Danish sentences, two words and chunks that are located between two silent segments of two passages of fluent text, for example: “Nej” (No), “Ja” (Yes), “Hvor skal du hen?” (Where are you going?). The set used contains 419 speech utterances (i. e., speech segments between two silence pauses) which are expressed by four professional actors, two males and two females. Speech is expressed in five emotional states: anger, happiness, neutral, sadness, and surprise. Twenty judges (native speakers from 18 to 58 years old) verified the emotions with a score rate of 67 %.

D. eNTERFACE

The eNTERFACE [14] corpus is a further public audiovisual emotion database. It consists of induced anger, disgust, fear, joy, sadness, and surprise speaker emotions. 42 subjects (eight female) from 14 nations are included. It consists of office environment recordings of pre-defined spoken content in English. Each subject was instructed to listen to six successive short stories, each of them eliciting a particular emotion. They then had to react to each of the situations by uttering previously read phrases that fit the short story. Five phrases are available per emotion as “I have nothing to give you! Please dont hurt me!” in the case of fear. Two experts judged whether the reaction expressed the emotion in an unambiguous way. Only if this was the case, the sample was added to database. Overall, the database consists of 1 277 samples.

E. SAL

The Belfast Sensitive Artificial Listener (SAL) data is part of the final HUMAINE database [15]. This subset contains 25 recordings in total from 4 speakers (2 male, 2 female) with an average length of 20 minutes per speaker. The data contains audiovisual recordings from natural human-computer conversations that were recorded through a SAL interface designed to let users work through a range of emotional states. The data has been labeled continuously in real time by four annotators with respect to valence and activation using the feel-trace system: The annotators used a sliding controller to annotate both emotional dimensions separately whereas the adjusted values for valence and activation were sampled every 10 ms to obtain a temporal quasi-continuum. To compensate linear offsets that are present among the annotators, the annotations were normalized to zero mean globally. Further, to ensure common scaling among all annotators, each annotator’s labels were 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. A total of 1 692 turns is accordingly contained in the database. Labels for each turn are computed by averaging the frame level valence and activation labels over the complete turn. Apart from the necessity to deal with continuous values for time and emotion, the great challenge of the SAL database is the fact that one must deal with all data.

F. VAM

The Vera-Am-Mittag (VAM) corpus [16] consists of audiovisual recordings taken from a German TV talk show. The set contains 946 spontaneous and emotionally colored utterances from 47 guests of the talk show which were recorded from unscripted, authentic discussions. The topics were mainly personal issues such as friendship crises, fatherhood questions, or romantic affairs. To obtain non-acted data, a talk show in which the guests were not being paid to perform as actors was chosen. The speech extracted from the dialogues contains a

TABLE I
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	2 449	553	2 449	3 002	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	1 277	1:00	34	8	2	studio	16
SAL	English	free	natural	884	808	917	779	1 692	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 II
MAPPING THE CLASSES OF VARIOUS DATABASES TO A BINARY AROUSAL (POSITIVE OR NEGATIVE).

Corpus	Positive	Negative
ABC	aggressive, cheerful, intoxicated, nervous	neutral, tired
AVIC	loi2, loi3	loi1
DES	angry, happy, surprise	neutral, sad
eNTERFACE	anger, fear, happiness, surprise	disgust, sadness
SAL	q1, q4	q2, q3
VAM	q1, q4	q2, q3

TABLE III
MAPPING THE CLASSES OF VARIOUS DATABASES TO A BINARY VALENCE (POSITIVE OR NEGATIVE).

Corpus	Positive	Negative
ABC	cheerful, neutral, intoxicated	aggressive, nervous, tired
AVIC	loi2, loi3	loi1
DES	happy, surprise neutral	angry, sad
eNTERFACE	happiness, surprise	anger, fear, disgust, sadness
SAL	q1, q2	q3, q4
VAM	q1, q2	q3, q4

large amount of colloquial expressions as well as nonlinguistic vocalizations and partly covers different German dialects. For annotation of the speech data, the audio recordings were manually segmented to the utterance level, whereas each utterance contained at least one phrase. A large number of human labelers was used for annotation (17 labelers for one half of the data, six for the other). The labeling bases on a discrete five point scale for three dimensions mapped onto the interval of [-1,1]: The average results for the standard deviation are 0.29, 0.34, and 0.31 for valence, activation, and dominance. The averages for the correlation between the evaluators are 0.49, 0.72, and 0.61, respectively. The correlation coefficients for activation and dominance show suitable values, whereas the moderate value for valence indicates that this emotion primitive was more difficult to evaluate, but may partly also be a result of the smaller variance of valence.

G. Mapping and Clustering

Since four of the six databases are annotated in terms of emotion categories, a mapping was defined to generate labels for binary arousal/valence from the emotion categories in order to generate a unified set of labels that can be used for cross-corpus experiments. This mapping is given in tables II and III.

III. ACOUSTIC FEATURES AND CLASSIFICATION

We employ acoustic feature vectors of 6 552 dimensions using our open source openEAR toolkit [17]. In total, we use 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 V summarizes the statistical functionals which were applied to the LLDs shown in Table IV to map a time series of variable length onto a static feature vector.

TABLE IV
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

As classifier, we consider Support Vector Machines (SVM) which can provide very good generalization properties and are presently one of the most used classifier in emotion recognition. 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. Implementations in the Weka toolkit [18] were used for further reproducibility.

IV. EXPERIMENTS

Our evaluation measure is unweighted accuracy (UA), i. e., the unweighted average of the recalls of the ‘positive’ and ‘negative’ classes, which has been the official competition measure of the first of its kind INTERSPEECH 2009 Emotion Challenge [1]. We evaluate on the six emotional databases following a cross-corpus leave-one-corpus out (LOCO) strategy, i. e., one corpus is used as test set while the remaining five are

TABLE V
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

TABLE VI
NORMALIZATION IN LEAVE-ONE-CORPUS-OUT CROSS-CORPUS BINARY AROUSAL / VALENCE CLASSIFICATION: TEST ON 6 DATABASES AND TRAINING ON 5 REMAINING DATABASES. UNWEIGHTED ACCURACY (UA) FOR CENTERING (C), MIN-MAX NORMALIZATION (M) AND Z-NORMALIZATION (Z) ON CORPUS BEFORE AND AFTER DATA AGGLOMERATION (AGG.), AND BOTH. W-MEAN: MEAN WEIGHTED BY NUMBER OF INSTANCES AS OPPOSED TO MEAN: MEAN OVER THE RESULTS OF THE CORPORA WITHOUT WEIGHTING BY THE NUMBER OF INSTANCES WITHIN THE CORPORA.

UA [%]	**Before agg.**			**After agg.**			**Both**	
Test on	**C**	**M**	**Z**	**C**	**M**	**Z**	**C**	**Z**
				Arousal				
ABC	63.1	64.5	66.6	64.3	60.2	61.0	63.4	65.5
AVIC	55.9	55.1	62.0	55.9	59.0	62.7	55.8	61.4
DES	76.1	79.1	78.3	74.9	66.3	74.4	77.9	80.1
eNTER.	62.7	60.0	61.6	61.7	57.8	61.6	63.3	60.8
SAL	60.0	55.4	61.6	64.4	51.2	64.7	62.9	63.3
VAM	64.6	58.2	69.2	65.8	58.3	67.4	67.4	69.7
W-Mean	60.5	58.2	63.9	62.9	57.5	64.1	61.6	64.0
Mean	63.7	62.1	66.6	65.2	58.8	65.3	65.1	66.8
				Valence				
ABC	63.6	62.2	62.3	63.3	58.0	59.7	63.6	62.3
AVIC	61.8	51.7	57.8	61.8	50.1	60.0	61.8	57.9
DES	57.0	56.3	59.7	57.0	61.1	57.9	56.8	59.7
eNTER.	57.4	56.0	58.2	56.5	55.2	57.4	57.4	58.2
SAL	54.3	50.0	53.4	54.3	51.1	55.5	54.3	53.4
VAM	54.4	51.5	52.0	56.4	53.0	54.0	54.5	52.0
W-Mean	58.4	52.8	56.6	58.5	52.5	57.7	58.4	56.6
Mean	58.1	54.6	57.2	58.2	54.8	57.4	58.1	57.3

used for (supervised or unsupervised) training. Training data is always agglomerated on the instance level by simply joining databases for training ('pooling') since in [10], this strategy has shown superior classification performance in comparison with late decision fusion for cross-corpus LOCO evaluation with SVM.

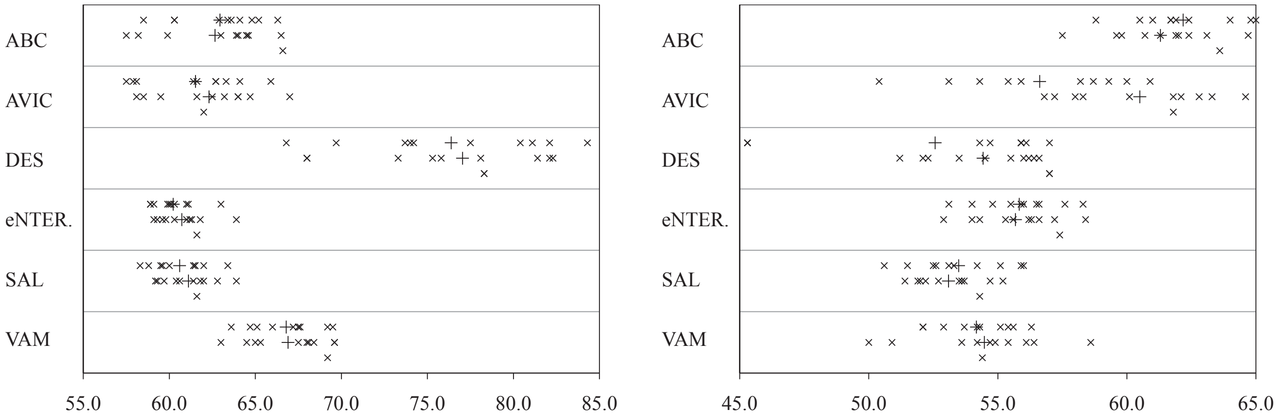
A. Normalization

In automatic speech and speaker recognition, methods such as cepstral mean subtraction or joint factor analysis are widely used to mitigate the diversities among speakers and acoustic environments. In our case of cross-corpus emotion recognition, differences do not only exist within corpora among different speakers (intra-corpus) but also in between corpora (inter-corpus) due to various recording settings and languages (cf. Table I); consequently, the impact of normalization techniques on cross-corpus recognition rates has been demonstrated [7]. In this paper, we investigate three kinds of normalization methods: centering, normalization and standardization. Centering is equal to simple subtraction of the feature-wise mean. Min-max (range) normalization forces the range of each feature to the interval [-1, 1] by linear scaling while z-normalization (sometimes referred to as standardization) refers to linear scaling to zero mean and unit variance. Thus, z-normalization is more robust to outliers than min-max normalization. These three methods can be applied to each corpus separately (i. e., before data agglomeration) or after building a joint training set from multiple databases. In Table VI we compare the mean unweighted accuracy (UA) across databases in LOCO evaluation with the three above named normalization methods. Since the databases vary greatly in size (cf. Table I), we present both, the unweighted mean and the mean UA weighted by number of instances in the database.

When applying normalization per corpus before data agglomeration, it can be seen that z-normalization delivers a vast improvement for arousal recognition both over min-max normalization and centering: The (unweighted) mean UA is 66.6 % for z-normalization compared with 62.1 % (min-max normalization) and 63.7 % (centering). For valence, interestingly, simple centering delivers best results (58.1 %), and min-max normalization severely deteriorates the results (54.6 %). The largest improvement in accuracy of arousal classification by standardization instead of centering or min-max normalization is found for the AVIC and VAM databases of spontaneous speech. These results are mirrored to a great extent in the results for normalization *after* data agglomeration, and in general the per-corpus normalization cannot be outperformed. We also investigated a combination of normalization both before and after agglomeration; the mean UA in arousal recognition in that case is 65.1 % for centering and 66.8 % for z-normalization. For valence recognition, this 'double' normalization yielded almost identical results as normalization before agglomeration. In all further experiments, we used z-normalization on the corpus level for arousal and centering for valence recognition. Note that the per corpus normalization strategy is very convenient in practice as it does not require re-training when adding further databases to the training set.

B. Unsupervised vs. Supervised Learning

To determine the potential of unsupervised learning for emotion recognition, we considered three different experimental settings: First, we agglomerated ('pooled') together three corpora for training and tested on one database (corresponding to 'Pool 3' in table VII). This results in ten possible training set permutations for each of the six test sets. Second, we agglomerated three corpora for training and two corpora for unsupervised adaptation, and tested on the remaining corpus (i. e., we used three corpora to build models that are used to generate predictions for the two further corpora which



(a) Arousal

(b) Valence

Fig. 1. Distributions of unweighted accuracies for cross-corpus binary arousal/valence classification of six test databases. Crosses refer to the individual training set combinations, while the plus sign refers to the average performance. The top row per test database depicts results obtained by pooling three training corpora, the row in the middle refers to pooling three corpora and fusing two corpora for unsupervised adaptation, and the bottom row represents pooling five training corpora. Note that in the last case no permutations are possible, as six corpora are used and five are pooled for training, while in the other cases several permutations exist.

in turn are used for unsupervised learning). This series of experiments is denoted by 'Pool 3 + 2' in table VII. Note that due to the varying size of the corpora, this covers both settings where little labeled data is available as a 'seed', and an 'unsupervised adaptation' scenario where the amount of unlabeled data is rather small compared with the available labeled data. Finally, as a reference for supervised learning, we considered agglomerating together five databases for training, again testing on the remaining corpus. Table VII shows the unweighted accuracies (UA) obtained for the two-class arousal and valence classification task when evaluating the three training scenarios. Using a set of three databases for training leads to an average UA of 64.7 % and 55.8 % for arousal and valence, respectively. Unsupervised adaptation with two additional corpora increases average recognition performance to 65.1 % and 56.6 %, respectively. The most impressive gain is seen for the AVIC database of spontaneous speech: Here, unsupervised training even slightly outperforms supervised training for arousal recognition, and gives a boost in accuracy of almost 4 % absolute for valence (compared with 5 % for supervised training). Still, as expected, the best average result is obtained when using the labels of all five corpora for training (UA of 66.6 % and 58.1 %, respectively). Figure 1 depicts the distributions of UA for the six test databases. The plus sign indicates the UA averaged over all test sets. The top row per test database depicts results obtained by agglomerating three training corpora, the middle row refers to agglomerating 3 corpora and fusing two further corpora for unsupervised learning, and the bottom row shows the results for agglomerating all five training corpora with known ground truth. From Figure 1, it can be seen that unsupervised learning outperforms the baseline setting (i. e., using only three corpora without further data agglomeration) in 5 of 6 cases for arousal but only 3 of 6 cases for valence, which can probably be attributed to generally insufficient robustness of cross-corpus valence recognition from acoustic features. Overall, in terms of (weighted) mean UA in arousal and valence recognition, addition of unlabeled training data delivers roughly half of the gain that can be expected from adding labeled training data, as in previous studies in speech recognition [3].

TABLE VII

MEAN AND THE MAXIMUM UNWEIGHTED ACCURACY (UA) OF SUPERVISED AND UNSUPERVISED TRAINING FOR CROSS-CORPUS BINARY AROUSAL / VALENCE CLASSIFICATION. POOL 3: AGGLOMERATION OF THREE CORPORA; POOL 5: AGGLOMERATION OF FIVE CORPORA; POOL 3 + 2: AGGLOMERATION OF THREE LABELED CORPORA AND TWO UNLABELED CORPORA FOR UNSUPERVISED LEARNING; W-MEAN: MEAN WEIGHTED BY NUMBER OF INSTANCES.

UA [%]	Pool 3		Pool 3 + 2		Pool 5
Test on	Mean	Max.	Mean	Max.	Value
			Arousal		
ABC	62.9	66.3	62.7	66.5	66.6
AVIC	61.5	65.9	62.3	67.0	62.0
DES	76.4	84.3	77.0	86.1	78.3
eNTER.	60.2	63.0	60.7	63.9	61.6
SAL	60.6	63.4	61.1	63.9	61.6
VAM	66.8	69.5	66.9	69.6	69.2
W-Mean	62.6	66.3	63.2	67.1	63.9
Mean	64.7	68.7	65.1	69.5	66.6
			Valence		
ABC	62.2	65.0	62.3	64.7	63.6
AVIC	56.6	60.9	60.5	64.6	61.8
DES	52.6	57.0	54.4	56.6	57.0
eNTER.	55.8	58.3	55.7	58.4	57.4
SAL	53.5	56.0	53.1	55.2	54.3
VAM	54.2	56.3	54.5	58.6	54.4
W-Mean	55.6	58.9	57.1	60.4	58.4
Mean	55.8	58.9	56.6	59.7	58.1

V. CONCLUSION

In this paper, we evaluated the suitability of unsupervised learning in a large-scale study on cross-corpus acoustic emotion recognition, investigating six different emotional databases as test sets and ten permutations of labeled and unlabeled training databases per test set. The results show that adding unlabeled data to agglomerated multi-corpus training sets can enhance recognition performance across emotion models, emotion elicitation methods and acoustic conditions. While the results are still clearly below the gain that can be expected when adding labeled data, the fact that manual labeling of emotional speech data is highly costly while large amounts of emotional speech data per se are publicly available (e. g., TV talk shows) makes consideration of unsupervised learning a promising approach for the future, even for cross-language and cross-application scenarios.

What seems interesting is that unlabeled data addition could surpass addition of the same data when labeled by humans in 'lucky' combinations. In the overall two times (arousal / valence) six testing cases, this was ten times the case. On average, the maximum obtained in unsupervised manner (column 'Pool 3 + 2', 'Max.' in Table VII) significantly exceed the use of the same data in human labeled version (column 'Pool 5') at the 10^{-3} (arousal) and 10^{-2} level (valence) in a one-sided z-test based on the weighted mean over all corpora (for comparison: adding unlabaled data (column 'Pool 3 + 2', 'Mean') significantly exceeds not adding (column 'Pool 3', 'Mean') at the common 0.05 level on average for valence). Our explanation for this is that some databases are better suited for training in the first place and thus lead to more reliable and consistent labels for the added speech without human label. It thus should be the aim to find good initial training data in order to profit from this effect. A starting point for the selection of 'good' training data is introduced in [19].

In addition, as mentioned in the introduction, currently running large-scale voice services such as the Google Voice Search show that roughly ten times the amount of data is needed to reach equivalent gains as by using labeled data. Here, we limited ourselves to comparing the same amount of added unlabeled data to adding the exact same data and by that same amount of labeled data. The obvious next step is thus to add considerably more unlabeled data not coming from emotional speech databases per se, but stemming from the richly available further resources.

Further promising directions can be found in considering classifier agreement in unsupervised learning—this somewhat corresponds to instance selection by agreement of human annotators as often done in emotion recognition [12] and is often employed in unsupervised machine learning techniques such as co-training [9]. Next, we focused on acoustic emotion recognition. Obviously, experience on combining acoustic and linguistic analysis in unsupervised manner will be of interest. Also, in speech recognition unsupervised *testing* has recently become a new practice in order to ensure consistency of unsupervised trained models and to overcome 'noise' of varying human transliteration styles. This seems particularly promising in our field of emotion recognition, where the ground truth annotation differs severely already among human raters and additionally needed to be mapped to a unifying scheme by employing arousal and valence dimensions.

Acknowledgment

Zixing Zhang received funding from a research grant by the People's Republic of China.

References

[1] B. Schuller, A. Batliner, S. Steidl, and D. Seppi, "Recognising realistic emotions and affect in speech: State of the art and lessons learnt from the first challenge," *Speech Communication*, vol. Special Issue on Sensing Emotion and Affect - Facing Realism in Speech Processing, 2011.

[2] D. Hakkani-Tur, G. Tur, M. Rahim, and G. Riccardi, "Unsupervised and active learning in automatic speech recognition for call classification," in *Proc. of ICASSP*, 2004, pp. 429–432.

[3] G. Tur and A. Stolcke, "Unsupervised Language Model Adaptation for Meeting Recognition," in *Proc. of ICASSP*, 2007, pp. 173–176.

[4] A. Mahdhaoui and M. Chetouani, "A new approach for motherese detection using a semi-supervised algorithm," in *Proc. of IEEE International Workshop on Machine Learning for Signal Processing*, Paris, France, 2009.

[5] J. Liu, C. Chen, J. Bu, M. You, and J. Tao, "Speech emotion recognition using an enhanced co-training algorithm," in *Proc. of IEEE International Conference on Multimedia and Expo (ICME)*, Hangzhou, China, 2007, pp. 999–1002.

[6] B. Maeireizo, D. Litman, , and R. Hwa, "Co-training for predicting emotions with spoken dialogue data," in *Proc. of Annual Meeting of the Association for Computational Linguistics*, 2004, pp. 202–205.

[7] 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, no. 2, pp. 119–131, 2010.

[8] 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, Germany, 2010.

[9] A. Blum and T. Mitchell, "Combining labeled and unlabeled data with co-training," in *Proc. of the Workshop on Computational Learning Theory (COLT)*. Morgan Kaufmann, 1998, pp. 92–100.

[10] B. Schuller, Z. Zhang, F. Weninger, and G. Rigoll, "Using Multiple Databases for Training in Emotion Recognition: To Unite or to Vote?" in *Proc. INTERSPEECH*, Florence, Italy, 2011, to appear.

[11] B. Schuller, M. Wimmer, D. Arsic, G. Rigoll, and B. Radig, "Audiovisual behaviour modeling by combined feature spaces," in *Proc. ICASSP*, 2007, pp. 733–736.

[12] 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, Special Issue on Visual and Multimodal Analysis of Human Spontaneous Behavior*, vol. 27, pp. 1760–1774, 2009.

[13] I. S. Engberg, A. V. Hansen, O. Andersen, and P. Dalsgaard, "Design, recording and verification of a Danish emotional speech database," in *Proc. Eurospeech*, Rhodes, 1997, pp. 1695–1698.

[14] O. Martin, I. Kotsia, B. Macq, and I. Pitas, "The enterface'05 audio-visual emotion database," *IEEE Workshop on Multimedia Database Management*, 2006.

[15] E. Douglas-Cowie, R. Cowie, I. Sneddon, C. Cox, O. Lowry, M. McRorie, J.-C. Martin, L. Devillers, S. Abrilan, A. Batliner, N. Amir, and K. Karpousis, "The HUMAINE Database: Addressing the Collection and Annotation of Naturalistic and Induced Emotional Data," in *Affective Computing and Intelligent Interaction*. Springer, 2007, pp. 488–500.

[16] M. Grimm, K. Kroschel, and S. Narayanan, "The Vera am Mittag German Audio-Visual Emotional Speech Database," in *Proc. IEEE International Conference on Multimedia and Expo (ICME)*, Hannover, Germany, 2008, pp. 865–868.

[17] F. Eyben, M. Wöllmer, and B. Schuller, "openEAR - Introducing the Munich Open-Source Emotion and Affect Recognition Toolkit," in *Proc. ACII*, Amsterdam, 2009, pp. 576–581.

[18] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. Witten, "The WEKA Data Mining Software: An Update," *SIGKDD Explorations*, vol. 11, 2009.

[19] B. Schuller, Z. Zhang, F. Weninger, and G. Rigoll, "Combining labeled and unlabeled data with co-training," in *Proc. 2011 Afeka-AVIOS Speech Processing Conference*, Tel Aviv, Israel, 2011.