Cross-Corpus Analysis for Acoustic Recognition of Negative Interactions

Iulia Lefter, Harold T. Nefs, Catholijn M. Jonker, and Leon J.M. Rothkrantz
Delft University of Technology, Delft, The Netherlands
Email: I.Lefter@tudelft.nl

***Abstract*—Recent years have witnessed a growing interest in recognizing emotions and events based on speech. One of the applications of such systems is automatically detecting when a situations gets out of hand and human intervention is needed. Most studies have focused on increasing recognition accuracies using parts of the same dataset for training and testing. However, this says little about how such a trained system is expected to perform 'in the wild'. In this paper we present a cross-corpus study using the audio part of three multimodal datasets containing negative human-human interactions. We present intra- and cross-corpus accuracies whilst manipulating the acoustic features, normalization schemes, and oversampling of the least represented class to alleviate the negative effects of data unbalance. We observe a decrease in performance when disjunct corpora are used for training and testing. Merging two datasets for training results in a slightly lower performance than the best one obtained by using only one corpus for training. A hand crafted low dimensional feature set shows competitive behavior when compared to a brute force high dimensional features vector. Corpus normalization and artificially creating samples of the sparsest class have a positive effect.**

***Keywords*—affective computing; cross-corpus evaluation; speech emotion recognition; surveillance**

I. Introduction

Automatic recognition of negative affect based on speech has applications in many contexts: monitoring human-human interactions in public service such as service desks [1], call-centers [2], [3], health care, monitoring conflicts in meetings [4] as well as general public surveillance [5]. Likewise, in the case of virtual agent systems, stress can arise due to task complexity and the inability of the virtual agent to act as expected by the client. Another example is the case of virtual reality therapy systems for anger management of psychiatric patients, or treatment of different phobias, where feedback can be given automatically to the patients based on their behavior.

One of the challenges of deploying trained systems in real-life setting is their limited generalization capabilities [6]. Given the availability of multiple datasets, one option of estimating how well a trained system would perform in practice is to test it on a disjunct dataset. In this paper we present a cross-corpus study using the audio part of three multimodal datasets containing negative interactions. While there have been already a number of studies on cross-corpus performance in speech emotion recognition in general [6], [7], [8], [9], [10], [11], our aim is to have a realistic view on obtainable performances for the special case of monitoring negative interactions, as would be expected in the aforementioned applications. To achieve this, we use three datasets with appropriate content and a degree of realism close to real life situations. We are not discriminating between different emotion categories. Rather, we are interested in evaluating whether a situation is normal, or whether it is escalating towards unwanted behavior. Further, we evaluate a set of design considerations for their generalization performances.

There are various reasons why the generalization of trained systems is challenging. With speech emotion recognition, one of the much debated issues is the use of acted datasets [12]. Some of the early available databases were developed by asking actors to utter predefined texts with different emotions [13] [14]. More realistic approaches make use of reality show recordings [15], Wizard of Oz scenarios of children interacting with a pet robot [16], emotion elicitation by interactions with virtual agents [17], and stress induction by dual-tracking workload computer tasks, or subject motion-fear tasks (subjects in roller-coaster rides) [18].

On the one hand, complaints about using acted data go about the fact that actors tend to exaggerate the emotion portray, and that since the emotion is not real and not spontaneous, the characteristics of the display are different. On the other hand, arguments in favor of using actors given special design considerations acknowledge the fact that real emotions are rare and short lived, and that emotion displays are affected by push (physiologically driven) and pull (social regulation and strategic intention) factors [19]. In the case of recording negative emotions, ethical considerations come into play, which make the recording of real life data challenging.

In this paper we performed a number of cross-corpus experiments using three datasets containing negative human-human interactions. The chosen corpora have a comparable degree of realism: improvisations given a short situation description, that build up spontaneously as the actors interact. The first two selected datasets have been developed for surveillance purposes in the train domain [20], and at service desks [21] respectively. As the third corpus we have chosen a subset of the IEMOCAP dataset [22] which contains improvisations with negative target emotions such as anger and frustration. Recordings of negative interactions are difficult to obtain in real life: negative events are rare, and ethical and privacy

This work was partially funded by EU FP7 project 2013-2016 COMPEIT 'Connected Media and Presence from the European Institute of Technology' and NWO project 432-13-802 'Virtual Reality Aggression Prevention Training for reducing victimization in forensic clinics'.

978-1-4799-9953-8/15/$31.00 ©2015 IEEE

reasons prevent collecting the data. Given these challenges, we believe these datasets are a legitimate choice for our study. Their high degree of realism and similar emotional content increases the suitability for cross-corpus research.

Recording conditions affect the generalization capabilities of emotion recognition systems. Such factors include room acoustics, whether there is a lab or a real-life setup, noise levels, quality and position of microphones used, filters (e.g. telephone). Other factors are the language used and the similarity in emotional content, speaker diversity, cultural differences [23], sparsity of some emotional classes and data sparsity in general, given the common belief in pattern recognition research that there is no data like more data. Even though the number of available datasets has increased, merging them is problematic due to differences in annotation schemes.

Studies of cross-corpus speech emotion recognition consistently indicate the significant inferiority of cross- compared to intra-corpus results [6], [9]. There have been different attempts to increase cross-corpus performances, such as agglomerating multiple datasets into one classifier for training [8], [11], voting based on results of individual classifiers trained on each dataset [24], adding unlabeled data emotional speech and using unsupervised learning techniques [25], using shared-hidden-layer autoencoder for representation learning shared across training and target corpora [26], and employing latent topic models for the extraction of turn level features [10].

Regarding acoustic features, popular approaches exploit either frame level features such as MFCC's and PLP, or focus on the supra-segmental behavior of emotion in speech and apply descriptive statistical functionals on time series such as the fundamental frequency or intensity. Initially, relatively small sized hand crafted feature vectors were used, but the trend is to generate high dimensional features by brute force [27] followed often by feature selection. According to [28], selecting the most promising features is data dependent, so finding an optimal feature set for cross-studies is still unsolved.

The datasets employed in this study differ in envisioned application, recording conditions, and annotation protocol. Furthermore, in all three cases the data is unbalanced, with the most negative cases being least represented. Motivated by the previously mentioned challenges of developing a recognition system that generalizes well, we manipulate of set of factors and evaluate their effect both intra- and cross-corpus. A hand crafted small dimensional feature set is compared with a brute force generated feature set. Two normalization schemes are tested for their ability to alleviate cross-corpus differences. To account for data imbalance we experiment with statistical oversampling of the minority class. We also test the effects of agglomerating two datasets for training. The total of 96 experiments are performed with a Support Vector Machine (SVM) classifier. Given the unbalanced data, the chosen performance measure is unweighted average recall (UA) to which we also refer with the term unweighted accuracy.

The remainder of this paper is organized as follows. In section II we give a detailed overview of the employed datasets in terms of content, recording procedure, and we propose a mapping between their annotations. In section III we introduce the acoustic feature sets, and in section IV we present the experiment setup and the manipulated design alternatives. The obtained intra- and cross-corpus performances given the different setups are presented and discussed in section V, and the paper ends with our conclusion in section VI.

II. Databases

This section describes the content of the three datasets used, as well as their annotations and proposed label mapping.

A. Dataset of Human-Human Service-Desk Interactions (SD)

The audio-visual service desk dataset (SD) proposed in [21] has been specifically designed for surveillance purposes. It contains improvised interactions of actors that only received roles and short scenario descriptions. The dataset contains a high variety of emotional colored manifestations. However, the actors did not receive any indication at all to encourage them to use their voice or body language in any particular way.

The actors had to play the roles of service desk employees and customers, given short role descriptions and short instructions. Four scenarios were played two times, resulting in eight sessions, for which the actors did not see the performance of their colleagues from the other session. Example scenarios are a visitor who is late for a meeting and has to deal with a slow employee, a helpless visitor unable to find a location on a map asking the employee to be escorted but he is being refused, the service desk employee does not want to help because of his lunch break and the employee or a visitor is in a phone conversation and blocking the service desk.

B. Dataset of Train Aggression (TR)

The audio-visual corpus of aggression in the train and train station contexts, together with the annotation scheme has been introduced in [20]. A detailed description of the recording procedure and considerations for scenario development can be found in [29]. To obtain the scenarios used for recording, the authors considered a set of rules defined by the railway company which describe normal behavior, and then generated scenarios which violate those rules. A set of 21 scenarios were generated, consisting of different abnormal behaviors. Examples are harassment, hooligans, theft, begging, noisy football supporters, medical emergency, traveling without ticket, irritation, passing through a crowd of people, rude behavior towards a mother with baby, fight for using the public phone, mocking a disoriented foreign traveler and irritated people waiting at the counter or toilet. As the authors argue [20], the term aggression is used in general for unwanted behavior, and the scenarios described have the role of intrigues. The scenarios were performed by a team of actors.

C. A Subset of the IEMOCAP Dataset (IE)

The IEMOCAP dataset [22] is a large collection of dyadic emotional interactions, containing audio, video as well as motion capture data from the face and body. The IEMOCAP corpus consists of two main categories or recordings: scripted

TABLE I
CHARACTERISTICS OF THE SELECTED CORPORA.

Dataset	Speakers	Language	Recording	Samples	Segmentation	Annotation	Raters	Agreement
SD	8	Dutch, English	noisy	534	utterance	stress	4	0.75
TR	12+	Dutch, English	noisy	1271	utterance	aggression	7	0.64
IE	10	English	lab	2277	2 seconds	negative valence	2-3	0.82

and improvised, which cover a variety of emotions such as happiness, anger, sadness, frustration and neutral state. For our study we have selected the audio of a subset of the IEMOCAP improvisations, which contain negative interaction which target anger and frustrations. These correspond to scenarios 1,4,5 and 8 as presented in [22], Table 1. Content wise, the scenarios contain service desk like interactions (being dissatisfied with the compensations offered by the lost luggage counter, being sent back after waiting a long time in line for not having the right id, being transfered to an operator after 30 minutes of talking with a machine), as well as interactions between friends, out of which one is frustrated due to long term unemployment.

D. Annotations and Label Mappings

All in all, the selection of datasets illustrated a large variety of situations in which negative interactions occur, or in which at least one of the subjects experience negative emotions. Even though the topics covered by the datasets are different, the emotional coloring does overlap. Further, the degree of realism is similar since, even though actors are employed, none of the used material has been clearly predefined or scripted, leaving the actors the freedom to act and respond to their partners.

In Table I we provide an overview of the main characteristics of the selected corpora, in terms of number of speakers, language used, quality of the recordings, segmentation and number of samples, annotated dimension, number of raters and their agreement measured as Krippendorff's alpha for ordinal data for SD and TR, and Cronbach's alpha for IEMOCAP.

One challenge of designing cross-corpus experiments is the different choice of annotation schemes for the existing corpora. In our case, the dataset of aggression in trains (TR) was annotated from a surveillance perspective, on a 3 point scale. Namely, the raters were asked to imagine they are surveillance operators, and rate normal situations with 1, to choose 2 for situations that would draw their attention, and 3 for situations in which they would feel the need to act. The service desk dataset was annotated on stress level on a five point scale, but as the authors mention they regard stress as a general phenomenon in the scene and also have the surveillance domain in mind. The IEMOCAP dataset was rated for emotion categories such as anger or frustration, as well on the emotional dimensions valence, arousal and dominance.

In order to come up with a common annotation mapping for the three datasets we have used the following approach. For the train database we kept exactly the same annotation. For the service desk dataset we mapped labels 2 and 3 to a single class (the new class 2), and labels 4 and 5 to a new class 3, indicating similarly to the train dataset a normal situation, and then two degrees of increased negative situations respectively. For the IEMOCAP subset we have selected the valence annotations, ranging from positive to negative on a 5 point scale. Valence is the most similar to the other two datasets' labels, and also the most appropriate given our interest in detecting when the interaction becomes negative. The selected subset contains the whole range of values for valence. We have used the original annotations by all provided raters and averaged them. In the resulting labels, 1 corresponds to highly positive and 5 to highly negative valence. We have mapped the two degrees of positive (in total 12% of the data) and the neutral valence labels to 1, corresponding to normal situations, and kept the two degrees of negative labels as label 2 and 3 respectively. To ensure that the mapping makes sense not only in theory, the first author has manually checked random samples from each resulting class, in all three datasets. Table II shows the label mapping from the original labels of each dataset to the new 3 point scale ground truth, as well as the obtained class distributions.

It can be observed from the class distributions in II that each of the datasets is unbalanced. In all three cases the least represented class is class 3, the one of highest negative interactions. This situation is similar to what is expected when recording real surveillance data, when most of the time nothing happens and there are only a small proportion of negative incidents. From a pattern recognition perspective, this also adds an extra challenge on achieving high recognition rates for class 3. Given the considered applications, obtaining high recognition rates of this class is extremely important, since negative events should not be missed.

TABLE II
OVERVIEW OF THE CHOSEN LABEL MAPPING AND RESULTING LABEL DISTRIBUTION OF THE SELECTED CORPORA. IN THE RESULTING MAPPING LABEL 1 MEAN A NEUTRAL SITUATION WHILE LABEL 3 MEAN A HIGHLY NEGATIVE / UNWANTED SITUATION.

	Mapping			**Distribution (%)**		
Class / **Dataset**	**1**	**2**	**3**	**1**	**2**	**3**
SD	1	2,3	4,5	38	46	16
TR	1	2	3	54	30	16
IE	1,2,3	4	5	40	47	13

III. ACOUSTIC FEATURES

Traditional approaches to speech emotion recognition can be seen as forming two main categories. In the first approach, frame based features are extracted and then operated at frame

level using among others Gaussian mixture models (GMM) or hidden Markov models (HMM) [30]. The second approach explores the suprasegmental traits of emotion by applying statistical functionals over the frame level features and using classification/regression techniques on the resulting feature sets [19]. The suprasegmental behavior of emotions [31] led to increasing the popularity of supra-segmental approaches.

The suprasegmental approaches started off by generating relatively small feature sets, obtained by applying a set of descriptive statistical functionals such as low order moments or extrema to the frame level features [32]. Recently, the brute force approach feature generation (resulting in 1-50k features) by analytical and evolutionary generation, gained popularity [27]. Frequently feature selection is used to reduce the high dimensionality, but one of the challenges is that the selected features is highly dependent on the chosen corpus [28].

In this paper we chose the supra-segmental approach, and compare a hand crafted small dimensional feature set described in section III-A to the one proposed in the first Interspeech Emotion Challange [33], described in section III-B. We are interested in the behaviors of these sets for both intra- and cross-corpus tasks. Both feature sets cover prosodic, spectral, and voice quality features.

A. Hand Crafted Feature Set (HC)

The hand crafted feature set was inspired from the minimum required feature set for emotion recognition proposed by [34] and the set proposed in [35] . The software tool Praat [36] was used to extract these features.

The hand crafted features set consists of the following 31 features: speech duration (without silences), pitch (mean, standard deviation, max, mean slope with and without octave jumps, and range), intensity (mean, standard deviation, max, slope and range), first four formants (F1-F4) (mean and bandwidth), jitter, shimmer, high frequency energy (HF500) (HF1000), harmonics to noise ration (HNR) (mean and standard deviation), Hammarberg index, spectrum (center of gravity, skewness), long term averaged spectrum (slope).

B. The INTERSPEECH 2009 Emotion Challange Feature Set (IS09)

The feature set proposed for the Interspeech 2009 Emotion Challenge consists of the most promising feature types and functionals covering prosodic, spectral, and voice quality features [33]. The features are extracted using OpenEAR [37].

The 16 low-level descriptors chosen are: zero-crossing-rate (ZCR) from the time signal, root mean square (RMS) frame energy, pitch frequency (normalised to 500 Hz), harmonics-to-noise ratio (HNR) by autocorrelation function, mel-frequency cepstral coefficients (MFCC) 1-12. To each of these, the delta coefficients are additionally computed. Next, 12 functionals (mean, standard deviation, kurtosis, skewness, minimum and maximum value, relative position, and range as well as two linear regression coefficients with their mean square error (MSE)) are applied, resulting in 384 features.

IV. Methodology

A. Classification and Evaluation

For our classification experiments we have chosen a Support Vector Machine (SVM) classifier with a linear kernel and unit c parameter, given its wide use and competitive performance [6]. We use the SVM implementation in the Weka library [38].

For intra-corpus results, the most common approach is cross-validations, with increased popularity on reporting speaker independent results, which leads to methods such as leave-one-speaker/session-out cross-validation. In our case, speaker identities are not available for all three datasets, and therefore we present 10-fold cross validation results, with the note that they might be more optimistic than speaker independent results. However, the speakers within the three corpora are disjunct, therefore, the cross-corpus experiments are speaker independent. In this case we train on one corpus and test on all the other ones, and we agglomerate two corpora for training and test on the remaining one.

In all cases, given the data unbalance, the evaluation measure is the unweighted average accuracy.

B. Normalization

Normalization is known to improve recognition rates in many speech applications. Approaches vary in terms of what data is considered for extracting the normalization parameters, such as speaker-based and corpus-based normalization. They also vary in terms of how the normalization is computed, for example by extracting the means, centering or z-normalizing.

Two normalization approaches are compared in this paper. The first one is a common corpus based z-normalization of all features, which we abbreviate CN. That means that each feature types is normalized by extracting the mean of that feature on the whole corpus, and dividing by the standard deviation.

The second approach is inspired by the work in [39]. Here, the authors propose a normalization scheme by computing the normalization parameters only based on the neutral (non-emotional) labeled samples. We abbreviate this with CNN. Hence, for implementing this normalization scheme we compute the mean and standard deviation of each feature type on only the data with neutral labels, and apply them to the entire corpus similarly to the first normalization scheme. In [39] the normalization scheme was applied per speaker, but since we did not have speaker identity information for all datasets, we did the normalization per corpus.

C. Statistical Oversampling

Data unbalance is a frequent problem that affects classification results. There are different possibilities to mitigate this effect, such as adapting the classifier's cost for the under-sampled class, and resampling the data to achieve more balance. In this paper we experiment with statistical minority oversampling (SMOTE) [40] implemented in Weka [38]. This method generates artificial new samples of the minority class by adding noise to the data. The percentage of new generated data is a parameter that has to be set. We applied feature

generation once for each dataset, with a percentage that would make the number of occurrences of the minority class similar to the one of the second least represented class. Namely, we have oversampled the minority class with 150, 50 and 100 percent of IEMOCAP, train dataset and service desk dataset respectively. For the merged sets of two datasets we have oversampled 100 percent of minority class for each.

V. Results and Discussion

All in all, given the intra- and cross-corpus results, with all combinations of features, normalization schemes, and applying SMOTE or not, we performed 24 intra-corpus experiments and 72 cross-corpus experiments. Our results are summarized in Figure 1, Figure 2 and Figure 3, each of them offering a view that sheds light on one of the manipulated factors. The results are grouped by test set. All figures display intra-corpus results on the left of each test group, shaded with light blue. Cross-corpus results trained on one corpus are not shaded, and cross-corpus results trained on two merged corpora are on the right of each test group, shaded with light yellow.

The expected superiority of intra-corpus compared to cross-corpus results is visible in all three figures, for all three datasets. The highest intra-corpus performance is achieved on the service desk dataset (SD): 76% unweighted average recall. The maximum accuracy on the train dataset (TR) is 70%, and the subset of the IEMOCAP is the most difficult to classify, with a maximum intra-corpus accuracy of 62%. Taking the difference between the maximum intra-corpus result and the maximum (one) cross-corpus result given all tested conditions, we observe absolute decreases of 7%, 13% and 9%, for the Service Desk, Train and IEMOCAP databases respectively. Differencing the best cross-corpus performance with training on one corpus, with the best cross-corpus performance by agglomerating the remaining two corpora, shows a decrease of 2%, 1% and insignificant difference for the service desk, train and IEMOCAP respectively. However, the differences are higher for various setup choices. These results show less dramatic decreases in cross-corpus experiments compared to some cases tested in [6], which might be attributed to the higher similarities in-between the selected corpora.

It is interesting to note that when testing on the IEMOCAP dataset, training on the TR dataset works consistently better than training on the SD dataset under all tested conditions. When testing on the SD data, again training on TR works better than training on IEMOCAP. Finally, testing on the TR is more successful when training is done on SD, rather than IEMOCAP. This leads to the conclusion that data similarities have an impact on how well a system is able to generalize. When no information about the test set is known, it seems reasonable to merge the training datasets, which performs slightly worse than the best case. If more information on the testing is available, it makes sense to use the most similar training set. Another key aspects that affects the cross-corpus results is how the data was labeled, and how appropriate the chosen mapping is.

Figure 1 highlights the differences determined by the choice of acoustic features. When training and testing on the IEMOCAP database, the IS09 features yield consistently better performance than the hand crafted feature set (HC). However the IS09 features generalize worse. For testing on the other two datasets, the HC features show competitive behavior, being with little exception almost always better than IS09. This is an interesting result, showing that a small-dimensional hand crafted feature set is able to perform very well when compared to a high dimensional one.

The two methods of data normalization are compared in Figure 2. As opposed to [39], we did not find that a normalization using only the means and standard deviations of the neutral samples consistently improves the results. A possible cause of this difference can be the fact that we did not perform the normalization per speaker, but per corpus.

As already mentioned, unbalanced data is one of the main problems especially when the application domain is detecting rare events like interactions that get out of hand. To diminish the negative effects of sparsity in one class, we applied statistical oversampling to that class. The results with and without oversampling are presented in Figure 3. With very few exception, oversampling consistently increases the unweighted accuracy. Not shown in the figure are the accuracies for class 3, the most negative cases. In these cases, oversampling generates a dramatic improvement. By averaging over all tested conditions, we observe an average improvement of 17% in the recall of class 3 using oversampling , which given the envisioned applications is extremely important.

VI. Summary and Conclusion

In this paper we performed a cross-corpus study with three datasets containing negative interaction. The purpose of the study was to obtain an indication of expected accuracy for detecting negative interactions in a close to real life setting. Besides, we experimented with two acoustic feature sets, two normalization procedures and with artificially oversampling the minority class. The study is relevant for many real life applications such as surveillance, monitoring in public service and health care applications, and the development of virtual reality therapy systems.

The three datasets used have been collected for different purposes and in different conditions. Nevertheless, all three of them have the content we were interested in, namely negative interactions. The collection procedure for the three corpora was similar: improvisations given short instructions.

As expected, intra-corpus experiment show the best recognition rates for all three corpora. Cross-corpus with training on one database only results in an approximate decrease of 10% in unweighted accuracy, if we consider the best performing condition in each case individually. We observed that training on the dataset with aggression in train (TR) solely in a cross-corpus setup is preferred to the other choice of training sets. However, when testing on the dataset with train aggression in a cross-corpus setup, generates the highest decrease in performance. By merging two datasets for training and testing

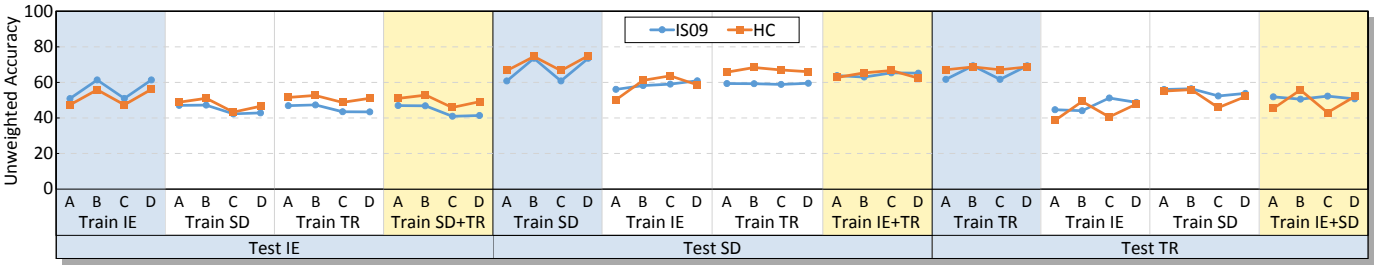

Fig. 1. **Feature effects.** Intra- and cross-corpus unweighted accuracies (UA) given the two selected feature sets. The red plots correspond to results obtained with the hand crafted feature set (HC) described in section III-A and the blue line to results using the IS09 feature set described in section III-B. The four conditions A,B,C and D correspond to different normalization schemes and applications of data oversampling or not (CN = corpus normalization, CNN = corpus normalization with neutral samples, SMOTE0 = no oversampling was applied, and SMOTE1 = oversampling was applied as explained in section IV-C). A = CN and SMOTE0, B = CN and SMOTE1, C = CNN and SMOTE0, D = CNN and SMOTE1.

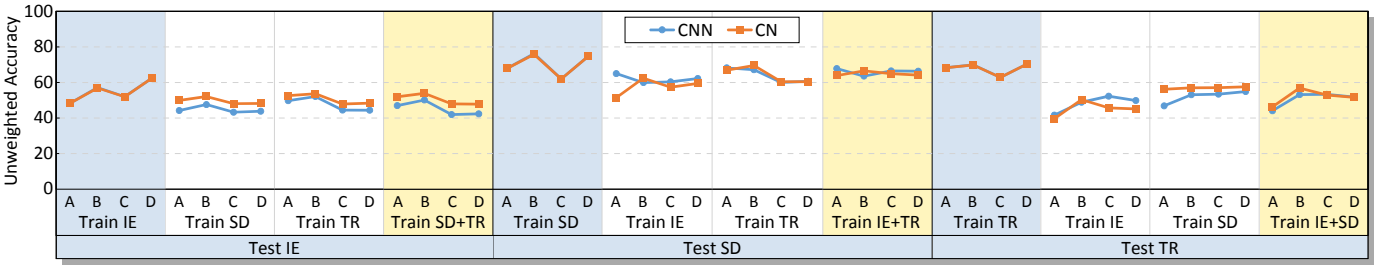

Fig. 2. **Normalization effects.** Intra- and cross-corpus unweighted accuracies (UA) given the two normalization procedures. The red plots correspond to results obtained with corpus normalization (CN), and the blue plots to corpus normalization using only the neutral samples (CNN), as described in section IV-B. The four conditions A,B,C and D correspond to different features and application of data oversampling or not (HC = hand crafted feature set, IS09 = Interspeech challange feature set, SMOTE0 = no oversampling was applied, and SMOTE1 = oversampling was applied as explained in section IV-C). A = HC and SMOTE0, B = HC and SMOTE1, C = ISO09 and SMOTE0, D = IS09 and SMOTE1.

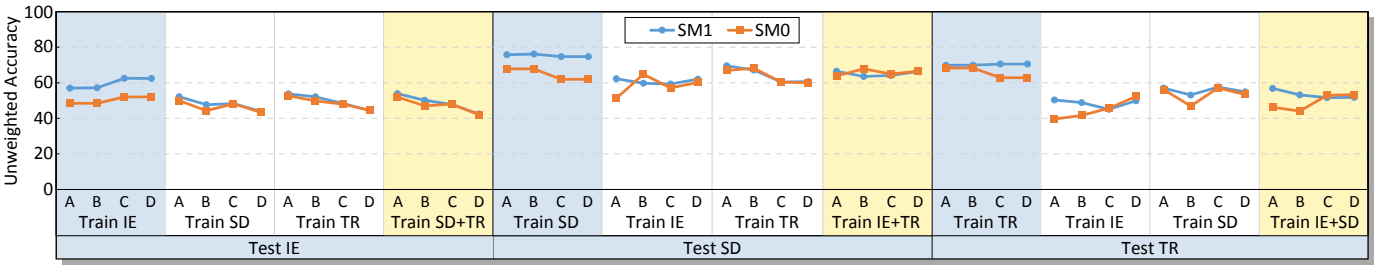

Fig. 3. **Oversampling effects.** Intra- and cross-corpus unweighted accuracies (UA) given applying oversampling (SMOTE1) or not (SMOTE0). The red plots correspond to results obtained with SMOTE0, and the blue plots to SMOTE1, as described in section IV-C. The four conditions A,B,C and D correspond to different features and normalization schemes (HC = hand crafted feature set, IS09 = Interspeech challange feature set, CN = corpus normalization, and CNN = corpus normalization using neutral samples). A = HC and CN, B = HC and CNN, C = ISO09 and CN, D = IS09 and CNN.

on the remaining one, we observe accuracies smaller than the best one obtained with only one dataset. One explanation for this is probably the similarity in behaviors and annotations between datasets. If the application domain is known, it seems best to use the most similar dataset for training. When little is known about what to expect, it makes sense to merge more datasets for training, which is similar to the findings in [24].

We have experimented with two types of acoustic feature sets, a small dimensional hand crafted feature set and a higher dimensional one proposed for the Interspeech 2009 Emotion Challange. Our results show that the small dimensional feature set can achieve comparable and even outperform the higher dimensional feature set in some situations.

Two normalization schemes were tested: corpus normalization (CN), and corpus normalization using neutral samples (CNN), in which the means and standard deviations are computed only based on the neutral (non-emotional) samples of the corpus. Unlike previous research [39] we could not demonstrate the consistent superiority of CNN, which can be caused by normalizing per corpus and not per speaker.

Finally, given that data unbalance is one of the challenges associated with detection of negative events in general, we evaluated the generation of artificial samples from the minority class of the training sets. This method proves to be beneficial and to increase the unweighted average recall in general. Furthermore, it offers a significant increase in recall of the sparsest class - the most negative situations - which given the envisioned applications is of high importance.

978-1-4799-9953-8/15/$31.00 ©2015 IEEE

REFERENCES

[1] I. Lefter, G. Burghouts, and L. Rothkrantz, "Recognizing stress using semantics and modulation of speech and gestures." *Affective Computing, IEEE Transactions on*, 2015, in press.

[2] I. Lefter, L. J. Rothkrantz, D. A. Van Leeuwen, and P. Wiggers, "Automatic stress detection in emergency (telephone) calls," *International Journal of Intelligent Defence Support Systems*, vol. 4, no. 2, pp. 148–168, 2011.

[3] L. Vidrascu and L. Devillers, "Detection of real-life emotions in call centers." in *INTERSPEECH*, vol. 2005, no. 10, 2005, pp. 1841–1844.

[4] S. Kim, F. Valente, and A. Vinciarelli, "Automatic detection of conflicts in spoken conversations: Ratings and analysis of broadcast political debates," in *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*, March 2012, pp. 5089–5092.

[5] P. van Hengel and T. Andringa, "Verbal aggression detection in complex social environments," in *Advanced Video and Signal Based Surveillance, 2007. AVSS 2007. IEEE Conference on*, Sept 2007, pp. 15–20.

[6] B. Schuller, B. Vlasenko, F. Eyben, M. Wollmer, A. Stuhlsatz, A. Wendemuth, and G. Rigoll, "Cross-corpus acoustic emotion recognition: Variances and strategies," *Affective Computing, IEEE Transactions on*, vol. 1, no. 2, pp. 119–131, July 2010.

[7] M. Tahon, A. Delaborde, and L. Devillers, "Real-Life Emotion Detection from Speech in Human-Robot Interaction: Experiments Across Diverse Corpora with Child and Adult Voices." in *INTERSPEECH*, 2011, pp. 3121–3124.

[8] F. Weninger and B. Schuller, "Discrimination of linguistic and non-linguistic vocalizations in spontaneous speech: Intra- and inter-corpus perspectives." in *INTERSPEECH*, 2012.

[9] I. Lefter, L. Rothkrantz, P. Wiggers, and D. Van Leeuwen, "Emotion recognition from speech by combining databases and fusion of classifiers," in *International conference on Text, Speech and Dialogue*, ser. TSD'10. Berlin, Heidelberg: Springer-Verlag, 2010, pp. 353–360.

[10] M. Shah, C. Chakrabarti, and A. Spanias, "Within and cross-corpus speech emotion recognition using latent topic model-based features," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2015, no. 1, 2015. [Online]. Available: http://dx.doi.org/10.1186/s13636-014-0049-y

[11] M. Shami and W. Verhelst, "Automatic classification of expressiveness in speech: a multi-corpus study," in *Speaker classification II*. Springer, 2007, pp. 43–56.

[12] E. Douglas-Cowie, N. Campbell, R. Cowie, and P. Roach, "Emotional speech: Towards a new generation of databases," *Speech communication*, vol. 40, no. 1, pp. 33–60, 2003.

[13] F. Burkhardt, A. Paeschke, M. Rolfes, W. F. Sendlmeier, and B. Weiss, "A database of German emotional speech." in *Interspeech*, vol. 5, 2005, pp. 1517–1520.

[14] I. S. Engberg, A. V. Hansen, O. Andersen, and P. Dalsgaard, "Design, recording and verification of a Danish emotional speech database." in *Eurospeech*, 1997.

[15] M. Grimm, K. Kroschel, and S. Narayanan, "The Vera am Mittag German audio-visual emotional speech database," in *Multimedia and Expo, 2008 IEEE International Conference on*. IEEE, 2008, pp. 865–868.

[16] A. Batliner, S. Steidl, and E. Nöth, "Releasing a thoroughly annotated and processed spontaneous emotional database: the FAU Aibo Emotion Corpus," in *Proc. of a Satellite Workshop of LREC*, 2008, pp. 28–31.

[17] G. McKeown, M. Valstar, R. Cowie, M. Pantic, and M. Schroder, "The SEMAINE database: Annotated multimodal records of emotionally colored conversations between a person and a limited agent," *Affective Computing, IEEE Transactions on*, vol. 3, no. 1, pp. 5–17, 2012.

[18] J. Hansen, S. Bou-Ghazale, R. Sarikaya, and B. Pellom, "Getting started with SUSAS: a speech under simulated and actual stress database." in *EUROSPEECH*, vol. 97, no. 4, 1997, pp. 1743–46.

[19] K. R. Scherer and T. Bänziger, "On the use of actor portrayals in research on emotional expression." in *Blueprint for affective computing: A sourcebook,*, K. R. Scherer, T. Bänziger, and E. B. Roesch, Eds. Oxford, England: Oxford university Press, 2010, pp. 166–176.

[20] I. Lefter, L. Rothkrantz, and G. Burghouts, "A comparative study on automatic audiovisual fusion for aggression detection using meta-information," *Pattern Recognition Letters*, vol. 34, no. 15, pp. 1953 – 1963, 2013.

[21] I. Lefter, G. Burghouts, and L. Rothkrantz, "An audio-visual dataset of human-human interactions in stressful situations," *Journal on Multimodal User Interfaces*, vol. 8, no. 1, pp. 29–41, 2014. [Online]. Available: http://dx.doi.org/10.1007/s12193-014-0150-7

[22] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. Chang, S. Lee, and S. Narayanan, "IEMOCAP: interactive emotional dyadic motion capture database," *Language Resources and Evaluation*, vol. 42, no. 4, pp. 335–359, 2008. [Online]. Available: http://dx.doi.org/10.1007/s10579-008-9076-6

[23] D. Neiberg, P. Laukka, and H. A. Elfenbein, "Intra-, inter-, and cross-cultural classification of vocal affect." in *INTERSPEECH*, 2011, pp. 1581–1584.

[24] 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*, 2011.

[25] Z. Zhang, F. Weninger, M. Wollmer, and B. Schuller, "Unsupervised learning in cross-corpus acoustic emotion recognition," in *Automatic Speech Recognition and Understanding (ASRU), 2011 IEEE Workshop on*, Dec 2011, pp. 523–528.

[26] J. Deng, R. Xia, Z. Zhang, Y. Liu, and B. Schuller, "Introducing shared-hidden-layer autoencoders for transfer learning and their application in acoustic emotion recognition," in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*, May 2014, pp. 4818–4822.

[27] B. Schuller, M. Wimmer, L. Mosenlechner, C. Kern, D. Arsic, and G. Rigoll, "Brute-forcing hierarchical functionals for paralinguistics: A waste of feature space?" in *Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on*. IEEE, 2008, pp. 4501–4504.

[28] T. Vogt and E. Andre, "Comparing feature sets for acted and spontaneous speech in view of automatic emotion recognition," in *Multimedia and Expo, 2005. ICME 2005. IEEE International Conference on*, July 2005, pp. 474–477.

[29] Z. Yang, "Multi-modal aggression detection in trains," Ph.D. dissertation, Delft University of Technology, 2009.

[30] A. Metallinou, A. Katsamanis, and S. Narayanan, "Tracking continuous emotional trends of participants during affective dyadic interactions using body language and speech information," *Image and Vision Computing*, vol. 31, no. 2, pp. 137–152, 2013.

[31] C. M. Lee and S. S. Narayanan, "Toward detecting emotions in spoken dialogs," *Speech and Audio Processing, IEEE Transactions on*, vol. 13, no. 2, pp. 293–303, 2005.

[32] Y. Li and Y. Zhao, "Recognizing emotions in speech using short-term and long-term features." in *ICSLP*, 1998.

[33] B. Schuller, S. Steidl, and A. Batliner, "The INTERSPEECH 2009 emotion challenge." in *INTERSPEECH*, vol. 2009, 2009, pp. 312–315.

[34] P. Juslin and K. Scherer, *In J. Harrigan, R. Rosenthal, and K. Scherer, (Eds.) - The New Handbook of Methods in Nonverbal Behavior Research*. Oxford University Press, 2005, ch. Vocal Expression of Affect, pp. 65–135.

[35] I. Lefter, L. Rothkrantz, and G. Burghouts, "Aggression detection in speech using sensor and semantic information," in *Text, Speech and Dialogue*, ser. Lecture Notes in Computer Science, P. Sojka, A. Horak, I. Kopecek, and K. Pala, Eds. Springer Berlin Heidelberg, 2012, vol. 7499, pp. 665–672.

[36] P. Boersma, "Praat, a system for doing phonetics by computer." *Glot International*, vol. 5, no. 9/10, 2001.

[37] F. Eyben, M. Wollmer, and B. Schuller, "OpenEAR - Introducing the Munich open-source emotion and affect recognition toolkit," in *Affective Computing and Intelligent Interaction and Workshops, 2009. ACII 2009. 3rd International Conference on*, 2009, pp. 1–6.

[38] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The weka data mining software: an update," *ACM SIGKDD explorations newsletter*, vol. 11, no. 1, pp. 10–18, 2009.

[39] C. Busso, A. Metallinou, and S. Narayanan, "Iterative feature normalization for emotional speech detection," in *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*, May 2011, pp. 5692–5695.

[40] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: synthetic minority over-sampling technique," *Journal of artificial intelligence research*, pp. 321–357, 2002.