Comparative Study on Normalisation in Emotion Recognition from Speech

Ronald Böck(✉), Olga Egorow, Ingo Siegert, and Andreas Wendemuth

Cognitive Systems Group, Otto von Guericke University Magdeburg,
Universitätsplatz 2, 39106 Magdeburg, Germany
ronald.boeck@ovgu.de
http://www.cogsy.de

Abstract. The recognition performance of a classifier is affected by various aspects. A huge influence is given by the input data pre-processing. In the current paper we analysed the relation between different normalisation methods for emotionally coloured speech samples deriving general trends to be considered during data pre-processing. From the best of our knowledge, various normalisation approaches are used in the spoken affect recognition community but so far no multi-corpus comparison was conducted. Therefore, well-known methods from literature were compared in a larger study based on nine benchmark corpora, where within each data set a leave-one-speaker-out validation strategy was applied. As normalisation approaches, we investigated standardisation, range normalisation, and centering. These were tested in two possible options: (1) The normalisation parameters were estimated on the whole data set and (2) we obtained the parameters by using emotionally neutral samples only. For classification Support Vector Machines with linear and polynomial kernels as well as Random Forest were used as representatives of classifiers handling input material in different ways. Besides further recommendations we showed that standardisation leads to a significant improvement of the recognition performance. It is also discussed when and how to apply normalisation methods.

1 Introduction

The detection of affective user states is an emerging topic in the context of human-computer interaction (HCI) (cf. [19,24]), as it is known that besides the pure context additional information on the user's feelings, moods, and intentions is transmitted during communication. For instance [1] discussed that such information should be used in HCI for a more general view on the human interlocutor.

The detection of emotions from speech can be seen as a challenging issue since both, the emotions themselves as well as the way humans utter emotions, introduce variations increasing the difficulty of a distinct assessment (cf. [2,24]). Furthermore, many up-to-date classification methods analyse data based on the distances between the given sample points (cf. [24]). As a consequence of the

P. Horain et al. (Eds.): IHCI 2017, LNCS 10688, pp. 189–201, 2017.
https://doi.org/10.1007/978-3-319-72038-8_15

aforementioned aspects, a data handling which scales the given samples in a comparable way has to be considered, leading to the question of data normalisation before classification. Yet, there are many approaches for data normalisation available (cf. e.g. [26] pp. 45–49) which are used in various studies.

The paper's aim is to investigate and to compare the different normalisation methods and to deduce in which situation they perform best. Since we were mainly interested in the *general trend* of the recognition results we did not argue on pure classification results, but *derived more general statements*. We are aware that a highly optimised classifier outperforms the systems presented in this paper. Nevertheless, in such cases, it is hard to identify general statements we are looking for. Therefore, the presented analyses are based on six normalising methods, dominantly used in the literature, applied to nine benchmark corpora well-known in the community of speech based emotion recognition.

The investigation is guided by the following research questions: **Q1:** Which normalising methods are usually applied in the community? **Q2:** Which normalisation approach provides the best recognition results? **Q3:** At which point can and shall normalisation be applied to the data? **Q4:** Can we derive recommendations stating which method(s) shall be used to achieve a reasonable improvement in the emotion recognition from speech?

Related Work. Normalisation is a pre-processing step which is applied to given material to handle differences caused by various circumstances. According to our knowledge, no comparison study on different normalisation methods based on several benchmark corpora was conducted for emotion recognition from speech. Nevertheless, various approaches are used in the community which are the foundations of this paper. Furthermore, we found that in the literature a heterogeneous terminology is used (cf. e.g. [15,31]). Therefore, we will use in the following a unique naming of normalisation methods.

In general, two papers present an overview on normalisation: in [26] normalisation techniques in the context of speaker verification are presented. For emotion recognition from speech, we found a rather brief overview in [31], highlighting that the same names often refer to different normalisation approaches.

Regarding the different normalisation techniques, the most prominent version is the standardisation (cf. [31]), although it is often just called normalisation. In most cases, papers refer to z-normalisation (cf. [7,9,16,21,22,25]) and further, to mean-variance-normalisation (cf. [29]).

Range normalisation and centering are, to the best of our knowledge, just used in the work of [15,31]. In [31], the authors applied these methods only on six data sets (a subset of corpora presented in Table 1) considering only two affective states and further, they do not vary the classifier.

Another approach highlighted in [15] is the normalisation based on neutral data. This idea is invented in [3], and further elaborated in [4]. In [15], the authors apply this approach on all three presented normalisation methods. As this is a promising approach keeping the differences between various affective states (cf. [3]), we included it in our experiments as well.

Several papers like [11,24,30] do not use any normalisation at all. This practice is related to the statement that "[f]unctionals provide a sort of normalisation over time" [24], assuming that normalisation is implicitely provided by the selected features mainly based on functionals.

In general, the presented works vary in approaches of normalisation, classification techniques, and utilised corpora. Therefore, a direct comparison of results is quite difficult for readers. The closest related papers for comparison are [21,31], as they refer to subsets of the benchmark corpora we analysed. Otherwise, as we were interested in the general characteristics of the normalising methods, we thus did not opt on fully optimised recognition results.

2 Data Sets

This study is focussed on the influence of normalisation approaches on the classification performance. Therefore, we decided to apply the various methods described in the literature to data sets widely used in the community. To cover various characteristics in the experiments, the corpora provide material in various languages, speaker ages and sexes as well as different emotional classes. Further, the material is recorded under different conditions reflecting acted and spontaneous (acoustic) expressions. The individual characteristics of each data set are presented in Table 1 and will be briefly introduced[1] in the following.

Table 1. Overview of the selected emotional speech corpora characteristics including information on number of classes (# C.) and if the corpus provides material for neutral speech (Neu.).

Corpus	Content	Samples	Subjects	Emo.	# C.	Neu.	HH:MM
ABC	German fixed	431	8 (4 f)	acted	6	x	01:15
AVIC	English variable	3 002	21 (10 f)	spont.	3	–	01:47
DES	Danish fixed	419	4 (2 f)	acted	5	x	00:28
emoDB	German fixed	492	10 (5 f)	acted	7	x	00:22
eNTERFACE	English fixed	1 277	42 (8 f)	acted	6	–	01:00
SAL	English variable	1 692	4 (2 f)	spont.	4	–	01:41
SmartKom	German variable	3 823	79 (47 f)	spont.	7	x	07:08
SUSAS	English fixed	3 593	7 (3 f)	spont. + act.	4	x	01:01
VAM	German variable	946	47 (32 f)	spont.	4	–	00:47

The *Airplane Behaviour Corpus (ABC)* (cf. [23]) is developed for applications related to public transport surveillance. Certain moods were induced using a predefined script, guiding subjects through a storyline. Eight speakers – balanced in sex – aged from 25–48 years (mean 32 years) took part in the recording. The 431 clips have an average duration of 8.4 s presenting six emotions.

[1] The explaining text for each corpus is inspired by [27].

The *Audiovisual Interest Corpus* (AVIC) (cf. [20]) contains samples of interest. The scenario setup is as follows: A product presenter leads each of the 21 subjects (ten female) through an English commercial presentation. The level of interest is annotated for every sub-speaker turn.

The *Danish Emotional Speech (DES)* (cf. [8]) data set contains samples of five acted emotions. The data used in the experiments are Danish sentences, words, and chunks expressed by four professional actors (two females) which were judged according to emotion categories afterwards.

The *Berlin Emotional Speech Database (emoDB)* (cf. [2]) is a studio recorded corpus. Ten (five female) professional actors utter ten German sentences with emotionally neutral content. The resulting 492 phrases were selected using a perception test and contain in seven predefined categories of acted emotional expressions (cf. [2]).

The *eNTERFACE* (cf. [18]) corpus comprises recordings from 42 subjects (eight female) from 14 nations. It consists of office environment recordings of pre-defined spoken content in English. Overall, the data set consists of 1277 emotional instances in six induced emotions. The quality of emotional content spans a much broader variety than in emoDB.

The *Belfast Sensitive Artificial Listener (SAL)* (cf. [6]) corpus contains 25 audio-visual recordings from four speakers (two female). The depicted HCI-system were recorded using an interface designed to let users work through a continuous space of emotional states. In our experiments we used a clustering provided by [21] mapping the original arousal-valence space into 4 quadrants.

The *SmartKom* (cf. [28]) multi-modal corpus provides spontaneous speech including seven natural emotions in German and English given a Wizard-of-Oz setting. For our experiments, we used only the German part.

The *Speech Under Simulated and Actual Stress* (SUSAS) (cf. [14]) dataset contains spontaneous and acted emotional samples, partly masked by field noise. We chose a corpus' subset providing 3593 actual stress speech segments recorded in speaker motion fear and stress tasks. Seven subjects (three female) in roller coaster and free fall stress situations utter emotionally coloured speech in four categories.

The *Vera-Am-Mittag* (VAM) corpus consists of audio-visual recordings taken from a unscripted German TV talk show (cf. [12]). The employed subset includes 946 spontaneous and emotionally utterances from 47 participants. We transformed the continuous emotion labels into four quadrants according to [21].

3 Normalising Methods

We reviewed the literature according to normalisation methods utilised in speech based emotion recognition and found four main approaches, but no direct comparison amongst them. Furthermore, it can be seen that the utilised methods are named differently by various authors although employing the same approaches. Therefore, we structured the methods and harmonised the naming.

Generally, we defined x as the input value representing, for instance, a speech feature, μ as the corresponding mean value, and σ as the corresponding variance.

Standardisation is an approach to transform the input material to obtain standard normally distributed data ($\mu = 0$ and $\sigma = 1$). The method is computed as given in Eq. 1.

$$x_s = \frac{x - \mu}{\sigma} \tag{1}$$

Range Normalisation is also called normalisation and is thus often confused with common standardisation. Therefore, we chose the term *range normalisation* that implies the possibility to vary the transformation interval. In Eq. 2 the interval is specified by $[a, b]$ and further $x_{\min}$ and $x_{\max}$ are the minimal and maximal values per feature. In contrast to standardisation (cf. Eq. 1) the mean and variance are not used by the approach.

$$x_n = a + \frac{(x - x_{\min})(b - a)}{x_{\max} - x_{\min}} \tag{2}$$

In our experiments we chose the interval $[-1, 1]$ for range normalisation.

The *Centering* approach frees the given input data from the corresponding mean (cf. Eq. 3). Therefore, the transformation results in a shift of input data.

$$x_c = x - \mu \tag{3}$$

Neutral Normalisation is an approach where normalisation parameters are computed based on neutral data, only. It is described in [4], and a logical extension of the idea to use neutral speech models for emotion classification (cf. [3]). Neutral normalisation is used for normalisation purpose in [15]. The methods works as follows: The parameters μ and σ or $x_{\min}$ and $x_{\max}$, respectively, for each feature are obtained based on the samples annotated as neutral and are further applied on samples with other emotional impressions. In our experiments this was done separately for each aforementioned normalisation method, namely standardisation, range normalisation, and centering.

Application of normalisation methods is as follows: The described normalising methods were applied to the training material as well as to the testing samples. For the test set two practices are possible and both were examined in our experiments. The first option assumed that both sets are known. Therefore, each set can be normalised separately, where accordingly optimal parameters (i.e. μ and σ, for instance) were used. In the second option, the necessary parameters were extracted only on the training set and applied to the testing set. In this case, it is assumed that the test samples are unknown, and thus no parameter estimation can be previously operated.

4 Experimental Setup

To evaluate the influence of normalisation, we conducted a series of classification experiments. Since one of our objectives was to obtain *reproducible* results

comparable to other studies, we decided to employ established feature sets and classifiers.

The *emobase* feature set is well-known in the community of emotion recognition from speech. This set comprises 988 functionals (e.g. mean, minimum, maximum, etc.) based on acoustic low-level descriptors (e.g. pitch, mel-frequency cepstral coefficients, line spectral pairs, fundamental frequency, etc.) [10]. The features are extracted on utterance level, resulting in one vector per utterance.

We decided to employ two different kinds of *classifiers*: the distance-based Support Vector Machine (SVM) and the non-distance-based Random Forest (RF). We expected that normalisation would provide significant improvement if using SVM, and no or only little improvement if using RF. For SVM, we used the LibSVM implementation developed by [5] implemented in WEKA [13]. For RF, we also rely on WEKA.

Since the data sets used in the experiments are very diverse, it would be difficult to impossible to fine-tune the classifiers to fit all the data. Therefore, we decided to use standard parameters for both, SVM and RF, without further fine-tuning. In the case of SVM, we chose a linear kernel (referred to as lin-SVM) and a polynomial kernel with a degree of 3 (referred to as pol-SVM), both with cost parameter $C = 1.0$. In the case of RF, we used 32 features per node, as the square root of the number of input features (in our case 988) is often used as default value in different RF implementations, and 1000 trees.

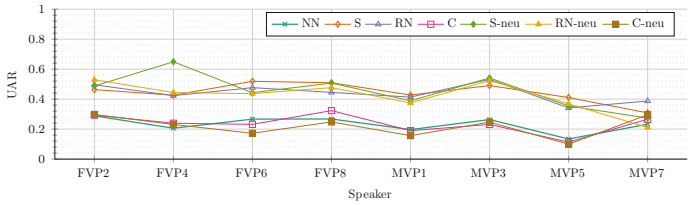
We evaluated the classifiers in a Leave-One-Speaker-Out (LOSO) manner, using the Unweighted Average Recall (UAR) of all emotions per speaker as evaluation metric.

5 Results

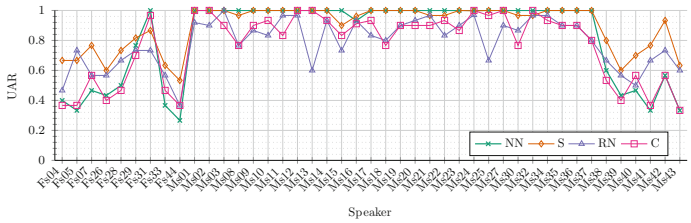
Figure 1 shows the results at a glance for lin-SVM on two of the nine investigated corpora (ABC and eNTERFACE). For the ABC corpus, we could see that some normalising methods such as standardisation performed better than others for nearly all speakers. For the eNTERFACE corpus, we see that the performance of the same normalising method varies remarkably depending on the speaker.

Table 2. Classification results (UAR, averaged over all nine corpora, in %) for all normalising methods (NN - non-normalised, S(-neu) - standardisation (with neutral), RN(-neu) - range normalisation (with neutral), C(-neu) - centering (with neutral)). The best classification result is highlighted for each classifier.

	NN	S	RN	C	S-neu	RN-neu	C-neu	Mean (w/o NN)
UAR for lin-SVM	39.1	49.6	45.9	38.7	47.3	45.1	32.6	43.2 ± 6.4
UAR for pol-SVM	37.4	40.1	22.9	33.5	42.9	27.4	30.3	32.9 ± 7.6
UAR for RF	44.9	47.5	43.2	46.1	45.5	43.2	45.2	45.1 ± 1.7



(a) ABC



(b) eNTERFACE

Fig. 1. UAR per speaker in (a) ABC and (b) eNTERFACE for lin-SVM.

In Table 2, the results are shown in a more detailed way, comparing the mean UAR, averaged over all nine corpora for all normalising methods and classifiers. For two of the three classifiers, standardisation outperformed other methods – and in the case of lin-SVM, neutral standardisation worked even better. Also, we see that standardisation and neutral standardisation were the only two normalising methods that always led to an improvement of the classification results.

An interesting point could be found by looking at the mean and standard deviation of all normalising methods presented in Table 2: For both SVM classifiers, normalising data in any kind changed the results (on average, +4.1% for lin-SVM and −4.5% for pol-SVM, absolute) more than in the case of RF (only 0.2%). There were also noticeable differences between the normalising methods, resulting in a higher standard deviation for both SVM classifiers compared to RF. Both observations support our hypothesis that in the case of SVM, changing the distance between data points by applying any normalising method would influence the classification results, whereas in the case of RF, normalisation would not change the classification results significantly.

There is another interesting point concerning the results using pol-SVM: Applying range normalisation significantly impairs the classification, leading to an UAR drop of 14.5% absolute. Our hypothesis concerning this phenomenon was that there is a non-linear effect induced by the combination of the polynomial kernel and high-dimensional data. To investigate this phenomenon, we conducted a series of additional experiments using polynomial kernels of increasing degrees. The results are shown in Table 3. We could see that the increasing degree of the kernel led to a drop in performance – for higher degrees the performance

Table 3. Mean UAR (in %) with variance on emoDB and SAL for SVMs with polynomial kernel (pol-SVM) presenting the anomaly between usage of range normalisation (RN) and higher polynomial degrees (d1 ... d6). For reference the results on non-normalised material using degrees 1 and 6 are shown.

	NN-d1	...	NN-d3	...	NN-d6
UAR (emoDB)	47.7 ± 7.4	...	45.3± 8.3	...	37.5 ± 6.1
UAR (SAL)	27.2 ± 1.6	...	24.9 ± 2.7	...	25.5 ± 2.3
	RN-d1	RN-d2	RN-d3	...	RN-d6
UAR (emoDB)	55.0 ± 7.5	20.0 ± 2.1	14.3 ± 0.0	...	14.3 ± 0.0
UAR (SAL)	30.5 ± 6.2	25.6 ± 0.6	25.0 ± 0.1	...	25.0 ± 0.0

decreases to chance level. This effect does not occur on non-normalised data, so we could conclude that it is related to or caused by range normalisation.

For a closer look on multi-corpus evaluation, the classification results in terms of UAR, obtained employing lin-SVM, are presented in Table 4. Since the data was not normally distributed, we executed the Mann-Whitney-U-Test (cf. [17]) to calculate significance for all classification outcomes. For five of the nine corpora, the improvements of normalised over non-normalised data were statistically significant ($p < 0.1$). But even for the cases where the improvements were not significant, normalising data led to at least some improvements: For all corpora except SAL, standardisation or standardisation on neutral data achieves the best results (cf. Table 4). In the case of SAL, range normalisation achieved the best results – but is only 0.2% better than standardisation. Otherwise, using inappropriate normalising methods could also impair the results. For example, in the case of AVIC, eNTERFACE, and SUSAS, all normalising methods except for standardisation led to minor decreases, although not statistically significant.

Table 4. Results achieved (UAR in %) using lin-SVM on normalised data and non-normalised baseline. Best results are highlighted gray, results below the baseline are given in *italic*. Significance levels: $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$

UAR	NN	S	RN	C	S-neu	RN-neu	C-neu
ABC	23.2	44.4***	43.9***	23.6	45.6***	42.1**	*21.8*
AVIC	46.6	47.5	*44.2*	*45.5*			
DES	30.3	50.5*	41.0	30.3	47.5*	44.2*	*27.4*
emoDB	47.4	77.2***	75.6***	51.4	72.4***	70.6***	48.9
eNTERFACE	81.9	89.3	*78.3*	*76.5*			
SAL	23.8	31.2	31.4	25.6			
SmartKom	16.5	19.0*	16.8	16.6	19.1*	17.4	*16.2*
SUSAS	53.4	54.4	52.0	50.4	52.0	*51.3*	*48.7*
VAM	28.6	32.5*	30.4	*28.1*			

Concerning normalising training and test set either using independently calculated parameters or using parameters calculated on both data sets, we could state that there is no significant difference in terms of UAR. There were some fluctuations in the results depending on the considered corpus, but the differences occurred in both directions and did not show a trend towards one option, and they were within the standard deviation. For example, in the case of AVIC, the maximum difference in the UAR achieved using independent versus combined parameters is 1.5% in favour of the former – with a standard deviation of 6.6% and 8.3% for independently and non-independently calculated normalisation parameters, respectively.

6 Discussion

In the current section the experimental results (cf. Sect. 5) are reflected considering the questions Q1 to Q4.

For question Q1, we analyse various works reflecting the state-of-the-art in the community (cf. Sect. 1). From these, we find that mainly two different approaches are used, namely standardisation and (range) normalisation. Less frequently centering is applied to data sets for normalisation purposes. Further, as presented in [3], the normalisation parameters can also be estimated based on emotionally neutral samples. This is tested in our experiments as well. We also find a slight trend towards standardisation in the literature.

Given this overview, we select the three most prominent methods for the experiments, namely standardisation, range normalisation, and centering (cf. Sect. 3). Further, they are also applied in the context of neutral normalisation if possible. Based on our results, the aforementioned trend towards standardisation is valid, since for eight benchmark corpora (cf. Table 1) standardisation produces an improvement in the recognition performance. The same statement holds for neutral normalisation, where standardisation shows the best performance as well (cf. question Q2). In our experiments we apply the LOSO validation strategy. Therefore, we have the opportunity to analyse the recognition performance in a speaker-independent way. As shown in Fig. 1 for ABC and eNTERFACE, the recognition results depend on the speaker to be tested. Of course, this effect is seen on the other corpora as well. Nevertheless, we find a relation between normalisation methods and the performance. For corpora containing mainly acted speech samples, a clustering of particular normalisation methods can be seen (cf. the gap between lines in Fig. 1(a)). In contrast for data sets providing more spontaneous emotions such clustering is not feasible. Further, the different methods are closer to each other in absolute numbers (cf. Fig. 1(b)). From our point of view, this is related to the lower expressivity of emotions uttered in spontaneous conversations, and hence, no particular normalisation approach is able to improve the recognition performance. As presented in Table 4, we can conclude that standardisation provides the best results across the nine benchmark corpora. In the case of SAL, range normalisation outperforms standardisation by 0.2%, absolute, only. Based on the Mann-Whitney-U-Test, we show that the

improvement of recognition performance is significant for five corpora (at least $p < 0.1$). For this, we test the significance against the non-normalised classification as well as against the second best results if the difference is low (cf. e.g. SmartKom in Table 4). This statistical significance emphasises the importance of suitable normalisation during the classification process.

Regarding the question how the normalisation shall be applied (cf. Q3), we tested two possible options: For the first one, the test set is normalised independently from the training set, for the second one, we normalise the test set using parameters obtained on the training set. The final results show that the differences in the recognition results are marginal with no statistical significance for either method. Therefore, both options are useful for testing purposes, and thus there is no need to refrain from using separately normalised test samples.

From our experiments, we can derive some recommendations for the application of normalisation approaches (cf. question Q4). First, in a multi-corpus evaluation based on a LOSO strategy standardisation is reasonable since in most cases (six of nine) this leads to a (significant) improvement of classification performances. This is also an indicator that normalisation improves even classification results based on feature sets mainly consisting of functionals (cf. *emobase* in Sect. 4). From our perspective this levels the statement of [24] that functionals already provide a kind of normalisation. Secondly, there is no need to favour either handling approach for test sets as no statistical significance in the differences in performance can be seen. Finally, the classifier influences the effect obtained by normalisation as well. From Tables 2 and 3 we can see that lin-SVM achieved better results than the other two classifiers across corpora. For RF, it was expected that normalisation has almost no influence since the classification is not distance based, resulting in lower standard deviations across corpora (cf. Table 2). In contrast, pol-SVM collapses with higher degrees (cf. Table 3), especially in the case of using range normalisation. We assume that this is related to a non-linear effect between the polynomial degree and the normalisation method. This will be further elaborated in future research.

7 Conclusion

In this paper, we have shown that normalising data in emotion recognition from speech tasks can lead to significant improvements. The extent of these improvements depends on three factors – these are the *general trends* we already discussed in Sect. 1. First of all, we have shown that standardisation works best in almost all cases: Applying it improved the recognition results for all nine corpora, for six corpora it proved to be the best normalising method. Secondly, the results depend on the used classifier: We have shown that, using lin-SVM, significant improvements are possible when applying standardisation as well as range normalisation. But for pol-SVM, range normalisation does not work well. The final factor is the data itself: For some corpora such as emoDB, improvements of up to 30% absolute are possible, for other corpora like SmartKom, only slight improvements of less than 3% absolute are achieved. From these findings we

can conclude that standardisation in most cases leads to substantially improved classification results.

Acknowledgments. We acknowledge continued support by the Transregional Collaborative Research Centre SFB/TRR 62 "Companion-Technology for Cognitive Technical Systems" (www.sfb-trr-62.de) funded by the German Research Foundation (DFG). Further, we thank the project "Mod3D" (grant number: 03ZZ0414) funded by 3Dsensation (www.3d-sensation.de) within the Zwanzig20 funding program by the German Federal Ministry of Education and Research (BMBF).

References

1. Biundo, S., Wendemuth, A.: Companion-technology for cognitive technical systems. KI - Künstliche Intelligenz **30**(1), 71–75 (2016)
2. Burkhardt, F., Paeschke, A., Rolfes, M., Sendlmeier, W., Weiss, B.: A database of German emotional speech. In: INTERSPEECH-2005, pp. 1517–1520, Lisbon, Portugal (2005)
3. Busso, C., Lee, S., Narayanan, S.S.: Using neutral speech models for emotional speech analysis. In: INTERSPEECH-2007, pp. 2225–2228. ISCA, Antwerp, Belgium (2007)
4. Busso, C., Metallinou, A., Narayanan, S.S.: Iterative feature normalization for emotional speech detection. In: Proceedings of the ICASSP 2011, pp. 5692–5695. IEEE, Prague, Czech Republic (2011)
5. Chang, C.C., Lin, C.J.: LIBSVM: A library for support vector machines. ACM Trans. Intell. Syst. Technol. (TIST) **2**, 1–27 (2011)
6. Douglas-Cowie, E., Cowie, R., Cox, C., Amier, N., Heylen, D.: The sensitive artificial listner: an induction technique for generating emotionally coloured conversation. In: LREC Workshop on Corpora for Research on Emotion and Affect, pp. 1–4. ELRA, Paris, France (2008)
7. El Ayadi, M., Kamel, M.S., Karray, F.: Survey on speech emotion recognition: Features, classification schemes, and databases. Pattern Recogn. **44**(3), 572–587 (2011)
8. Engbert, I.S., Hansen, A.V.: Documentation of the Danish emotional speech database DES. Technical report Center for PersonKommunikation, Aalborg University, Denmark (2007)
9. Eyben, F., Scherer, K., Schuller, B., Sundberg, J., Andre, E., Busso, C., Devillers, L., Epps, J., Laukka, P., Narayanan, S., Truong, K.: The geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing. IEEE Trans. Affect. Comput. **7**(2), 190–202 (2016)
10. Eyben, F., Wöllmer, M., Schuller, B.: Opensmile - the munich versatile and fast open-source audio feature extractor. In: Proceedings of the MM-2010, pp. 1459–1462. ACM, Firenze, Italy (2010)
11. Eyben, F., Batliner, A., Schuller, B., Seppi, D., Steidl, S.: Cross-corpus classification of realistic emotions - some pilot experiments. In: LREC Workshop on Emotion: Corpora for Research on Emotion and Affect, pp. 77–82. ELRA, Valetta, Malta (2010)
12. Grimm, M., Kroschel, K., Narayanan, S.: The vera am mittag German audio-visual emotional speech database. In: Proceedings of ICME 2008, pp. 865–868. IEEE, Hannover, Germany (2008)

13. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.H.: The WEKA data mining software: an update. ACM SIGKDD Explor. Newsl. **11**(1), 10–18 (2009)
14. Hansen, J., Bou-Ghazale, S.: Getting started with SUSAS: A speech under simulated and actual stress database. In: Proceedings of EUROSPEECH-1997, vol. 4, pp. 1743–1746. ISCA, Rhodes, Greece (1997)
15. Lefter, I., Nefs, H.T., Jonker, C.M., Rothkrantz, L.J.M.: Cross-corpus analysis for acoustic recognition of negative interactions. In: Proceedings of the ACII 2015, pp. 132–138. IEEE, Xi'an, China (2015)
16. Lefter, I., Rothkrantz, L.J.M., Wiggers, P., van Leeuwen, D.A.: Emotion recognition from speech by combining databases and fusion of classifiers. In: Sojka, P., Horák, A., Kopeček, I., Pala, K. (eds.) TSD 2010. LNCS (LNAI), vol. 6231, pp. 353–360. Springer, Heidelberg (2010). https://doi.org/10.1007/978-3-642-15760-8_45
17. Mann, H.B., Whitney, D.R.: On a test of whether one of two random variables is stochastically larger than the other. Ann. Math. Stat. **18**(1), 50–60 (1947)
18. Martin, O., Kotsia, I., Macq, B., Pitas, I.: The eNTERFACE'05 audio-visual emotion database. In: Proceedings of the Workshop on Multimedia Database Management. IEEE, Atlanta, USA (2006)
19. Picard, R.: Affective Computing. MIT Press, Cambridge (2000)
20. Schuller, B., Müller, R., Hörnler, B., Höthker, A., Konosu, H., Rigoll, G.: Audiovisual recognition of spontaneous interest within conversations. In: Proceedings of the 9th ICMI, pp. 30–37. ACM, Nagoya, Japan (2007)
21. Schuller, B., Vlasenko, B., Eyben, F., Rigoll, G., Wendemuth, A.: Acoustic emotion recognition: A benchmark comparison of performances. In: Proceedings of the ASRU 2009, pp. 552–557. IEEE, Merano, Italy (2009)
22. Schuller, B., Vlasenko, B., Eyben, F., Wöllmer, M., Stuhlsatz, A., Wendemuth, A., Rigoll, G.: Cross-corpus acoustic emotion recognition: Variances and strategies. IEEE Trans. Affect. Comput. **1**(2), 119–131 (2010)
23. Schuller, B., Arsic, D., Rigoll, G., Wimmer, M., Radig, B.: Audiovisual behavior modeling by combined feature spaces. In: Proceedings of the ICASSP-2007, pp. 733–736. IEEE, Honolulu, USA (2007)
24. Schuller, B., Batliner, A., Steidl, S., Seppi, D.: Recognising realistic emotions and affect in speech: State of the art and lessons learnt from the first challenge. Speech Commun. **53**(9–10), 1062–1087 (2011)
25. Schuller, B., Zhang, Z., Weninger, F., Rigoll, G.: Selecting training data for cross-corpus speech emotion recognition: Prototypicality vs. generalization. In: Proceeeings of the Afeka-AVIOS Speech Processing Conference, Tel Aviv, Israel (2011)
26. Schwartz, R., Kubala, F.: Hidden markov models and speaker adaptation. In: Laface, P., De Mori, R. (eds.) Speech Recognition and Understanding: Recent Advances, Trends and Applications, pp. 31–57. Springer, Heidelberg (1992)
27. Siegert, I., Böck, R., Vlasenko, B., Wendemuth, A.: Exploring dataset similarities using PCA-based feature selection. In: Proceedings of the ACII 2015, pp. 387–393. IEEE, Xi'an, China (2015)
28. Steininger, S., Schiel, F., Dioubina, O., Raubold, S.: Development of user-state conventions for the multimodal corpus in smartkom. In: Proceedings of the Workshop on Multimodal Resources and Multimodal Systems Evaluation, pp. 33–37. ELRA, Las Palmas, Spain (2002)
29. Tahon, M., Devillers, L.: Towards a small set of robust acoustic features for emotion recognition: challenges. IEEE/ACM Trans. Audio Speech Lang. Process. **24**(1), 16–28 (2016)

30. Tahon, M., Devillers, L.: Acoustic measures characterizing anger across corpora collected in artificial or natural context. In: Proceedings of the 5th International Conference on Speech Prosody. ISCA, Chicago, USA (2010)
31. Zhang, Z., Weninger, F., Wöllmer, M., Schuller, B.W.: Unsupervised learning in cross-corpus acoustic emotion recognition. In: Nahamoo, D., Picheny, M. (eds.) Proceedings of the ASRU 2011, pp. 523–528. IEEE, Waikoloa, HI, USA (2011)