Exploring Dataset Similarities using PCA-based Feature Selection

Ingo Siegert*, Ronald Böck*, Bogdan Vlasenko† and Andreas Wendemuth*
*Cognitive Systems Group, Otto von Guericke University Magdeburg, Germany
Email: {firstname.lastname}@ovgu.de
†Idiap Research Institute Martigny, Switzerland
Email: bvlasenko@idiap.ch

***Abstract*—In emotion recognition from speech, several well-established corpora are used to date for the development of classification engines. The data is annotated differently, and the community in the field uses a variety of feature extraction schemes. The aim of this paper is to investigate promising features for individual corpora and then compare the results for proposing optimal features across data sets, introducing a new ranking method. Further, this enables us to present a method for automatic identification of groups of corpora with similar characteristics. This answers an urgent question in classifier development, namely whether data from different corpora is similar enough to jointly be used as training material, overcoming shortage of material in matching domains. We compare the results of this method with manual groupings of corpora. We consider the established emotional speech corpora AVIC, ABC, DES, EMO-DB, ENTERFACE, SAL, SMARTKOM, SUSAS and VAM, however our approach is general.**

I. Introduction

In automatic speech recognition as well as in emotion recognition from speech, data material is an important resource. Todays emotional speech recognition is mostly data-driven: by using labelled speech samples a classifier is trained that can afterwards be used to classify unknown or unseen data [1]. To successfully train a classifier by this method, usually a high amount of data is needed. But, most datasets does not comprise such an amount. Thus, a combination of data from different corpora is needed. Though, two problems have to be addressed: 1) the combination of different emotional classes and 2) the selection of acoustically similar datasets. The first issue is already addressed in [2] by using an emotional clustering into broader classes of high and low arousal or positive and negative valence. But the second issue has so far not detailed investigated for emotional speech corpora.

Although various emotional speech datasets are available, they differ in various properties, e.g. recording quality, acoustic setting, and type of emotions [3]. Thus, for the combination of different datasets, one has to choose similar corpora in advance. The question arises: which properties should be taken into account as basis for similarity? This is especially important as emotions are phenomena that are decoded within the acoustic signal, together with spoken content, nonverbal utterances and background noise. Also the expressiveness of emotions can be different which also makes it hard to combine different corpora with reliable recognition results [4]. Thus, for training the classifier as similar as possible data in terms of quality and character should be used.

Only a few publications investigate similarities across emotional corpora. In [5] acoustic features were examined for a few corpora to distinguish anger. The authors could show that their investigated loudness and spectral features tend to be higher for anger across all corpora. But this investigation needed substantial manual effort, as it was conducted on voiced parts only and a manual pre-selection of the anger parts had to be performed. Another publication [6] measures the similarity between different corpora by implementing a sequential forward floating search (SFFS) algorithm. They also incorporate a Principal Component Analysis (PCA), but only to visualise the distribution of the corpora along the first two components. To measure the similarity the authors used the SFFS algorithm together with a cross-correlation to indicate the similarity. Unfortunately the SFSS has high computation costs, so an exhaustive analysis across several datasets is quite expensive. The authors of [7] pursued a feature selection from two corpora, showing that acted and spontaneous corpora share specific features. But as this investigation was limited to two datasets, a general statement could not be made.

In the following we present a method that determines the similarity across several corpora by analysing the features' variances in the data samples using PCA. This allow us to resign extensive feature selection algorithms, as the PCA directly computes the main influencing feature variances within all corpora [8]. In contrast to previous studies, we investigate nine well known emotional corpora to put our analyses on a broader basis. Using rank analyses, we are able to identify groups of similar corpora. At the same time this study allows us to make a statement about which features should be used as a minimum set for certain corpora groups.

The remainder of the paper is structured as follows: Section II shortly describes the utilized datasets and emphasized specific differences. In Section IV, we shortly describe the PCA and our feature ranking as well as our similarity measure. Section V presents the grouping of similar corpora. Afterwards, in Section VI their minimal feature set is discussed. Finally, Section VII gives an outlook for further research.

II. Corpora

For our experiments we chose nine among the most popular emotional speech corpora. They cover a broad variety in types

TABLE I
OVERVIEW OF THE SELECTED EMOTION CORPORA.

Corpus	Content	All	Subjects	Emotion	Quality	Audio channel	Length (HH:MM)
ABC	German fixed	431	8 (4 female)	acted	studio	16 kHz 16 bit	01:15
AVIC	English variable	3002	21 (10 female)	spontaneous	normal	44.1 kHz 16 bit	01:47
DES	Danish fixed	419	4 (2 female)	acted	normal	20 kHz 16 bit	00:28
emoDB	German fixed	494	10 (5 female)	acted	studio	16 kHz 16 bit	00:22
eNTERFACE	English fixed	1277	42 (8 female)	acted	noisy	16 kHz 16 bit	01:00
SAL	English variable	1692	4 (2 female)	spontaneous	normal	16 kHz 16 bit	01:41
SmartKom	German variable	3823	79 (47 female)	spontaneous	noisy	16 kHz 16 bit	07:08
SUSAS	English fixed	3593	7 (3 female)	mixed	noisy	8 kHz 16 bit	01:01
VAM	German variable	946	47 (32 female)	spontaneous	normal	16 kHz 16 bit	00:47

of emotional content, reaching from acted speech, covering story guided speech and spontaneous emotions with fixed spoken content, and finally corpora with further variety with respect to the number of subjects involved, spontaneity, and free language. The total amount of all presented emotional speech corpora is approx. 14h. An overview on different properties of the chosen databases is given in Table I.

A. Acted emotions

The *Airplane Behaviour Corpus (ABC)* [9] is developed 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. 8 speakers in gender balance from 25-48 years (mean 32 years) took part in the recording. The 431 clips have an average duration of 8.4s.

The *Danish Emotional Speech (DES)* [10] database is a representative of acted emotions. The data used in the experiments are Danish sentences, words and chunks that are expressed by four professional actors, two males and two females.

The *Berlin Emotional Speech Database (emoDB)* [11] is a further well known studio recorded corpus. Ten (five female) professional actors speak ten German emotionally neutral sentences in terms of content. It contain 494 phrases [12], where both naturalistic and pre-identified emotions are present.

The *eNTERFACE* [13] 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 database consists of 1277 emotional instances. This database has been chosen for this investigation, as it represents a simulated database with laymen. Thus, although the affective material is acted, the quality of emotional content spans a much broader variability, than in emoDB.

B. Spontaneous emotions

To add spontaneous emotion samples of non-restricted spoken content, we further decided for the *Audiovisual Interest Corpus* (AVIC) [14]. In its scenario setup, a product presenter leads one of 21 subjects (10 female) through an English commercial presentation.

The *Belfast Sensitive Artificial Listener (SAL)*, used e.g. in [15], contains 25 audio-visual recordings in total from 4 speakers (2 female) with an average length of 20 minutes per subject. The depicted natural human-computer conversations were recorded using an interface designed to let users work through a range of emotional states.

The *SmartKom* [16] multi-modal corpus recorded spontaneous speech and natural emotions for German and English via a Wizard-Of-Oz setting. For our evaluations we use the German dialogues, only. The dataset is structured into sessions of approximately 4.5 min length per subject.

The *Vera-Am-Mittag* (VAM) database [17] consists of audio-visual recordings taken from a popular authentic and unscripted German TV talk show. The set used includes 946 spontaneous and emotionally coloured utterances from 47 talkshow guests. For annotation of the speech material, the audio recordings were manually segmented to the utterance level, whereas each utterance contained at least one phrase.

C. Mixed emotions

The *Speech Under Simulated and Actual Stress* (SUSAS) database [18] contains both spontaneous and acted emotional instances. As additional challenge, the speech signal is partly masked by field noise. We decided for the 3593 actual stress speech segments recorded in speaker motion fear and stress tasks. Seven subjects, three of them female, in roller coaster and free fall actual stress situations are included in this dataset.

III. METHODS

A. Feature Extraction

Since we compare nine different corpora (cf. Section II) in terms of relevant features reflecting the corpora's characteristics, we had to decide which candidate features should be observed at all. In particular, the experiments need to be general and reproduceable in a broader sense. For this, a basic feature set, reflecting a variety of characteristics, was selected. A possible feature set for all nine corpora was already proposed by Schuller et al. including 6552 Low-Level-Descriptors and dedicated functionals [2]. In the corresponding paper, neither feature selection nor feature dimension reduction was conducted. From [7] a feature selection is known, considering 1280 features and observing two corpora of the presented nine only, namely emoDB and SmartKom. Hence, we can conclude that already a reduced set of relevant features (roughly 90 to 160) was identified taking into account the characteristics of the particular two corpora in [7]. As we can see, there is a huge variety in the number of selected features but also in

the type of features [19]. Further, from the literature review in [20], we can state that for several of the observed corpora reasonable results were achieved by various research groups using considerably less than 6552 features. According to [7], we also decided to start with a larger feature collection. Since we intended reproducibility, we selected the feature set proposed by Eyben et al. in the context of the openEAR project [21]. The feature set is called "emobase" and contains overall 952 characteristics extracted by the OpenSMILE toolkit [21], used as feature set for various recognition experiments [22], [23]. Besides time and class information, the features are based on Low-Level-Descriptor as Cepstrum, MFCCs, pitch, LPC, LSP, etc. and corresponding first order functionals like extrema, moments, and percentiles. Neglecting the non-feature information, as the class identifier in the extracted characteristics, we end up with a set of 949 candidate features for all corpora. To perserve to individual corpus characteristics, no feature normalisation is applied.

B. Shared Features

Looking at the extracted features which are shared amongst the nine corpora (cf. Section II), we investigated the meaning of the features. In particular, we are interested whether these features are already well-known from literature or novel combinations yet not considered. Moreover, this indicates whether high-order features might be more meaningful. The shared features are presented in Table II. Since we extracted these with OpenSMILE [21], Table II briefly describes the naming. We refer to the OpenSMILE Book [24] for further details.

TABLE II
BRIEF DESCRIPTION ACCORDING TO THE OPENSMILE BOOK [24] OF MEANINGFUL FEATURES SHARED AMONGST THE NINE CORPORA.

Name	Meaning
mfcc	Mel-Frequency Cepstral Coefficient
sma	smoothing with moving average filter; window length of 3
de	delta coefficient
[i]	ith coefficient
range	max - min value of the contour
iqr1-3	inter-quartile range of quartile 1 and 3
stddev	standard deviation of contours' values
linregerrA	difference of linear approximation and current contour

C. Principal Component Analysis

The PCA is an approach which can be used to identify reasonable dimensions in a feature space given a set of samples. In general, PCA computes the principal components u_k, as a linear combination of original features x_j, while examining the variance in the data. The first principal component is oriented in the direction of the largest variance. The following components are furthermore oriented in the direction of the decreasingly ordered further variances. The most relevant combination is the projection of the features to the eigenvector $\vec{w}^k$ associated with the highest eigenvalue λ_k of the feature correlation matrix $\mathbf{C}$ and thus, provides the best information. This results in a weighting coefficient w_j^k for each component of the linear combination. Finally, the linear combination can be expressed for the kth dimension of the projected feature vector $\vec{u}$ as follows:

$$u_k = \sum_j w_j^k x_j \tag{1}$$

The weighting reflects the contribution of the original features to the linear combination, and thus is related to the original variance of the data samples. Based on these coefficients, a feature reduction can be established.

For conducting the PCA we used the WEKA toolkit [25] which provides a framework allowing for the necessary analyses. As we intended to establish a relation between the number of relevant principal components covering all nine corpora or subsets thereof, we computed all components and applied a feature ranking afterwards (cf. Section III-D and [8]). This means, a reduction to the most reasonable features was not done by WEKA but by ourself. For this, we achieved more flexibility and could influence the feature reduction directly.

D. Feature Ranking

To perform the feature ranking, we rely on a method presented in [8]. Only the first dimension of the projected feature vector, $\vec{u}_1$, associated with the first principal component, having the largest λ_1, is used. This is valid under the assumption (which has to be secured) that the following λ_i $(i = 2, 3, \ldots)$ are much smaller, since in this case, the data distribution can be thought of as being *principally extended* in the direction of $\vec{w}^1$. Additionally, the features x_j of the selected component are ranked by their absolute weight $|w_j^1|$. By this method, a large $|w_j^1|$ generates a large contribution of the feature x_j to the projection u_1. The ranking can be further adjusted by eliminating all $|w_j^1|$ below a predefined threshold ϑ. We did not apply such a threshold value, as we want to analyse all features and the development of joined features over the whole number of 949 features. Thus, we rearranged all projections of the first principal component for each utilised dataset, resulting in a ranked list as given in Table III.

TABLE III
EXCERPT OF THE RANKED FEATURE LIST OF PCA-BASED FEATURE SELECTION FOR ONE CORPUS (ABC)

$\lambda_i = 297.6$	w_j^1	x_j
1. feature	0.055	mfcc_sma_de[2]_linregerrA
2. feature	0.055	mfcc_sma_de[2]_stddev
3. feature	0.054	mfcc_sma_de[7]_linregerrA
4. feature	0.054	mfcc_sma_de[6]_linregerrA
...	...	...
949. feature	0	mfcc_sma[10]_linregc1

For ABC used in Table III, we can indeed verify that $\lambda_1 = 297.6$, followed with considerable distance by $\lambda_2 = 64.1$.

To extract the features that are used over all corpora, we go step-wise from the first to the last feature through the ranked lists of all corpora and store the features with their occurrences. At that moment, where a feature has been seen in all corpora, this (joined) feature is stored together with the rank of its last occurrence. As we go stepwise through the list

of ranked features, the weight $|w_j^1|$ of each feature has already been considered and can be neglected. If we use more than one principal component from a corpus, i.e. the second eigenvalue is not remarkably smaller then the first one, we expand the ranked list to have as many columns as components are used. If in this expanded list a feature x_j is occurring repeatedly, only the instance with the best weight, i.e. $\max_k |w_j^k|$, is stored, the remaining occurrences are omitted.

IV. Defining a Measure for Similarity

A. The Influence of the Number of Eigenvalues

At first, we have a look on the eigenvalues of all components of the utilised corpora, to make an assumption which components have to be considered for our feature selection and similarity measure. The ten highest eigenvalues for all 949 components of all nine corpora are depicted in Fig. 1.

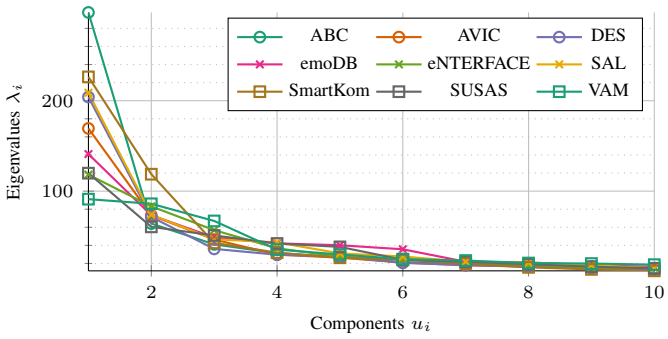

Fig. 1. The values of the first then eigenvalues of the PCA components for all nine investigated corpora

From Fig. 1 it can be seen that for all corpora except VAM the eigenvalue of the first component is remarkably larger than the next eigenvalue. But on all corpora the first three eigenvalues comprise one-third of the variance of the feature distribution. As we are interested in a measure for the similarity of the different datasets, we concentrate on the first eigenvalue, as the first component covers most of the feature variances and thus is responsible for the characterisation of the data samples within each corpora [6]. At this point, we can raise the hypothesis that VAM is remarkably different from all other corpora. Next we have to find a simple value that is able to express this difference. From the definition of the PCA we know that the original feature dimensions are recombined along their contribution to the overall variance of that corpus. If two datasets have similar characteristics, that is expressed by the extracted features, than the PCA should rearrange the features into a similar component. Thus, the rank of the first joined features will be quite low. To approve this assumption, we analyse the development of joined features over all corpora.

B. Analysis of Joined Features over all Corpora

At first, we analysed how the number of joined features evolve when we go through our ranked feature list over all corpora. The evolvement when combining up to the third PCA component of each corpus is given in Fig. 2. As we have stated in Fig. 1, the first three eigenvalues of VAM are very close. Therefore, we also included the evolvement when using the first component of each corpus together with the second and third component from VAM only, denoted as $\lambda_1 + VAM_3$.

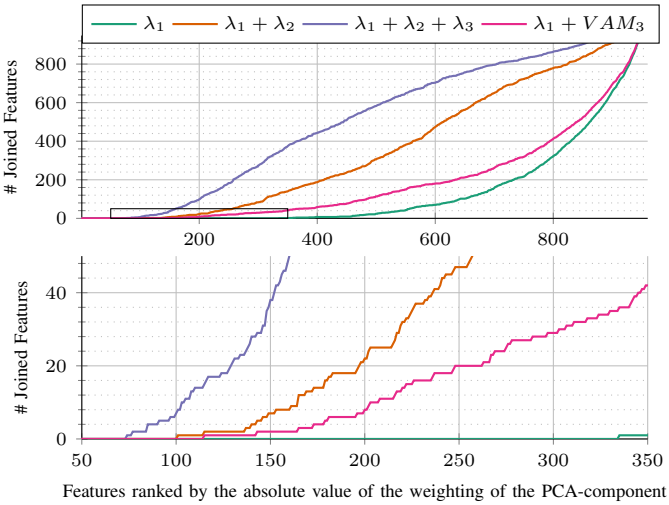

Fig. 2. Development of joined features over all corpora dependent of the ranked feature list for different numbers of considered PCA-components

From Fig. 2 it can be seen that by including more components for each corpus, the number of the rank where the same feature has occurred in all corpora is decreasing as expected. In the case of using only the component with the highest eigenvalue (λ_1), at rank 335 all corpora share one feature (mfcc_sma_de[12]_range), whereas for the best two eigenvalues, this happens at rank 101 (mfcc_sma_de[2]_iqr1-3). For the best three eigenvalues we can observe a shared feature at rank 74 (mfcc_sma_de[1]_stddev). The last case ($\lambda_1 + VAM_3$), is quite similar to $\lambda_1 + \lambda_2$. In this case all corpora share one feature at rank 115 (mfcc_sma_de[2]_linregerrA).

C. Similarity Measure: The Rank of the first Joined Feature

The specialty of the VAM corpus can also be seen, when we combine eight corpora and highlight the rank when the first feature is joined over all eight corpora, Table IV. From this table it can be seen if VAM is omitted the first feature that is joined over all remaining corpora is remarkable lower (115) than for all other corpus combinations where VAM is included (mean of 327). The same effect can be observed when using combinations of corpora sub-groups for 3 to 7 corpora. All combinations, where VAM is contained, have a significant higher rank than all combinations without VAM (non-parametric ANOVA [26], $p < 0.001$). Thus, we can state that VAM seems to be a corpus that has no similarity to all other investigated corpora. The similarity could be explored by just analysing the rank of the first joined feature exposed by a PCA based feature selection. Thus, this rank seems to be a very good indicator for the similarity of different datasets and will be used in the following.

TABLE IV

NUMBER OF THE RANK WHEN THE FIRST JOINED FEATURE IS PRESENT IN EIGHT CORPORA USING ONLY THE FIRST EIGENVALUE. FOR COMPARISON THE RANK OF THE FIRST JOINED FEATURE OVER ALL CORPORA IS GIVEN

Corpora **not** used	Rank	Corpora **not** used	Rank
emoDB	335	SAL	335
eNTERFACE	335	DES	334
VAM	115	AVIC	335
SUSAS	277	ABS	335
SmartKom	323		
all		335	

Next, we investigated whether we also find a corpus that has a quite low rank in combination with one other corpus. Therefore, we build sub-groups of only two corpora where additionally, we keep hold of one corpus (base-corpus) and extract the rank of the first joined feature for the combinations with all other corpora. We depict the results in Fig. 3. We excluded VAM, as we have already shown that this corpus is quite different from all other investigated corpora. Thus, we get 7 values for each corpus but the distributions between different corpora are not independent. In Fig. 3 it is shown that SmartKom has the lowest median rank of 3. Thus, it can be assumed that this corpus is quite similar to all other corpora as it has a quite low rank in many groupings.

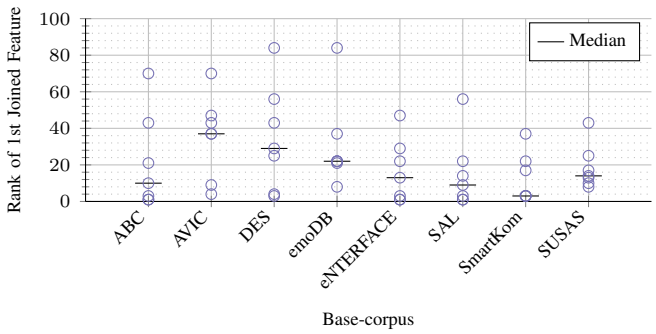

Fig. 3. Rank of joined feature for groups of two corpora, where the base-corpus is fix. VAM is excluded from this comparison

V. RESULTS FOR CORPUS-SIMILARITY

A. Results by Manual Selection using Corpus Description

In the following, we investigate the rank, when combining different corpora by using various properties. The properties are depicted in Table I. We manually distinguish by recording characteristics, type of emotions, and language. The resulting rank of the first joined feature for all manual groupings is depicted in Table V. As most of the corpus properties are shared among three to four datasets, we also investigate whether the similarity rank changes if we use just a selection of three corpora out of the groups with four corpora. In Table V, the groupings are then ordered by first displaying the group with four datasets and then subgroups thereof.

When using the corpus properties and our feature-based corpus similarity measure we can find six groups (M1-M6)

TABLE V

NUMBER OF THE RANK AS SIMILARITY FEATURE FOR DIFFERENT MANUALLY SELECTED GROUPS OF CORPORA. THE LOWEST RANK FOR EACH PROPERTY IS DENOTED WITH MX.

	Type of Emotion		
M1	acted	ABC DES emoDB eNTERFACE	96
	acted	ABC DES emoDB	96
	acted	DES emoDB eNTERFACE	96
	acted	ABC DES eNTERFACE	59
	acted	ABC emoDB eNTERFACE	22
M2	spont	AVIC SAL SmartKom SUSAS	84
	spont	AVIC SAL SmartKom	37
	spont	AVIC SAL SUSAS	47
	spont	AVIC SmartKom SUSAS	55
	spont	SAL SmartKom SUSAS	17
	Quality		
M3	studio	emoDB ABC	21
	normal	AVIC DES SAL	65
M4	noisy	eNTERFACE SmartKom SUSAS	21
	Language		
M5	German	ABC emoDB SmartKom	22
	English	AVIC eNTERFACE SAL SUSAS	79
	English	AVIC eNTERFACE SAL	47
	English	AVIC eNTERFACE SUSAS	55
M6	English	eNTERFACE SAL SUSAS	14

of three corpora that are very similar. These groups are basically formed by two clusters of corpora: ABC, emoDB, and eNTERFACE with acted emotions; SAL, SmartKom and SUSAS having spontaneous emotions. In this case the rank is in the same regions as for groups of two corpora (cf. Fig. 3).

B. Results by Automatic Grouping over all Datasets

Additionally, we also conducted experiments, where we select a sub-set of three, four and five corpora out of all eight corpora and analysed the rank of the first joined feature. Again, we omitted VAM. The groupings with the lowest rank are presented in Table VI. Those groupings with ranks up to the rank of the best subset of six corpora are indicated by Ax.

By automatic grouping we could reach similar ranks than for the manual grouping based on specific corpus properties. For the three and four corpora sets, we could even expose better ranks. Thus, these corpora share a specific characteristic that is not covered directly by the corpus description.

VI. DISCUSSION

Regarding the VAM we can state this corpus as a special case. The explanation for this observation can be found in the corpus description itself. VAM is a dataset generated from a TV show excerpt. In this case the sound recording took place in a big studio rather than in a lab and several directional microphones are mixed together. Thus the acoustics is quite different from all other datasets. The dataset contains large amount of cross-talk segments.

When comparing Fig. 3 with Table V and Table VI, we see that ABC, SAL, and SmartKom are mostly parts of the groupings with the smallest rank. These corpora also had the smallest rank compared to all other corpora on a one-to-one basis. Thus, we can draw the conclusion, that these corpora share characteristics with all other datasets.

TABLE VI
NUMBER OF THE RANK AS SIMILARITY FEATURE FOR DIFFERENT AUTOMATICALLY SELECTED GROUPS OF CORPORA. FOR COMPARISON THE MEAN AND STANDARD DERIVATION IS GIVEN. THE LOWEST RANK FOR EACH SUB-GROUP IS DENOTED WITH AX.

	Sub-set of Three Corpora	
A1	ABC eNTERFACE SAL	1
A2	ABC SAL SmartKom	3
A3	ABC eNTERFACE SmartKom	3
A4	eNTERFACE SAL SmartKom	3
	ABC SAL SUSAS	14
	ABC eNTERFACE SUSAS	14
	ABC SmartKom SUSAS	21
	ABC emoDB SUSAS	21
	Mean (Std)	48 (30)
	Sub-set of Four Corpora	
A5	ABC eNTERFACE SAL SmartKom	3
A6	ABC eNTERFACE SAL SUSAS	14
	ABC SAL SmartKom SUSAS	21
	ABC eNTERFACE SmartKom SUSAS	21
	eNTERFACE SAL SmartKom SUSAS	21
	Mean (Std)	73 (34)
	Sub-set of Five Corpora	
A7	ABC eNTERFACE SAL SmartKom SUSAS	21
A8	ABC emoDB eNTERFACE SAL SmartKom	22
	Mean (Std)	92 (28)
	Sub-set of Six Corpora	
A9	ABC emoDB eNTERFACE SAL SmartKom SUSAS	25
	Mean (Std)	105 (18)
	Sub-set of Seven Corpora	
	ABC AVIC emoDB eNTERFACE SAL SMARTKOM SUSAS	101
	Mean (Std)	112 (7)

Using our method, we see that the automatic grouping could find a group of six corpora with similar datasets, denoted A9. This group has a low rank in comparison to all other groupings. The other groupings, denoted as M1-6 in Table V and A1-8 in Table V are just sub-sets of this set. As we did not find the A9 grouping by selecting similar corpus properties, we can state that these corpora have a quality property that is not directly given in the corpus description. The first joined feature of A9 is mfcc_sma_de[2]_linregerrA. The evolvement of joined features over the rank of selected features for A9 is depicted in Fig. 4. By combining those corpora a total amount of 12:27 hours of speech can be used for training.

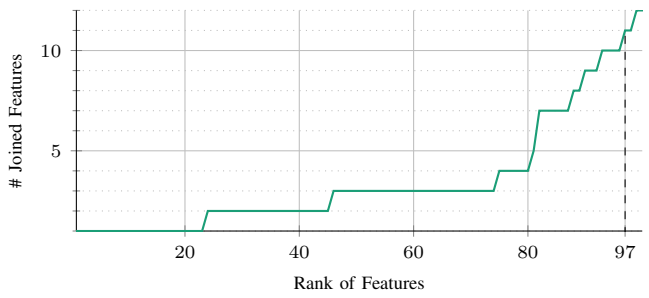

Fig. 4. Development of joined features over the six most similar datasets (emoDB, eNTERFACE, SAL, SmartKom, SUSAS) up to rank 100

If we select the first 98 features from the corpora included in A9, we have 10 features in common. These features are given in Table VII. They can serve as a minimal set of features for a joined corpus emotion recognition.

TABLE VII
THE FIRST TEN JOINED FEATURES FOR THE SIX MOST SIMILAR DATASETS.

Rank	Feature
25	mfcc_sma_de[5]_iqr1-3
47	mfcc_sma_de[9]_stddev
76	mfcc_sma_de[9]_linregerrA
82	mfcc_sma_de[7]_linregerrA
83	mfcc_sma_de[12]_linregerrA
83	mfcc_sma_de[7]_stddev
89	mfcc_sma_de[2]_linregerrA
91	mfcc_sma_de[8]_linregerrA
94	mfcc_sma_de[3]_linregerrA
98	mfcc_sma_de[2]_iqr1-3

Interestingly, the ten most common features throughout the six most similar datasets (A9) are MFCCs. These features have been proven to show good recognition results on various emotional speech corpora of different type, language and content [27], [28], [3]. Thus, our investigation reveals the importance of MFCCs also for speech based emotion recognition. MFCCs are the standard acoustic features used for dynamic frame-level emotion recognition.

VII. OUTLOOK

In this paper we presented a method for the calculation of the similarities between several corpora using a PCA-based feature selection method. This approach is quite general and thus can be applied to any set of corpora.

Until now no classification experiments using the selected corpora are pursued, in particular with the automatically selected groupings and feature sets. Cross-dataset experiments will be done in a forthcoming investigation to verify the performance of the assumed corpus similarity. In the future we would like to provide intra– and inter-corpora evaluation for different groups of similar datasets.

Thereby, we will also test whether the proposed features of the similar sets are generally suitable to form a minimalistic set for speech-based emotion recognition. Additionally, a principled reason why these features are selected among these corpora has to be investigated. Therefore, a comparison with a PCA-based feature selection on each dataset alone. Also using the weights learned by a linear Support Vector Machine could help to distinguish important features from others

As a further aspect also the relation between features and class-labels can be investigated. For this the same experiments using an Linear-Discriminant-Analysis will be perfomed. In this case the features are analysed according to the highest class-seperating variance.

ACKNOWLEDGMENT

The work presented in this paper was done within 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).

REFERENCES

[1] C. M. Bishop, *Pattern Recognition and Machine Learning*, 2nd ed. Berlin, Heidelberg, Germany: Springer, 2011.

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

[3] D. Ververidis and C. Kotropoulos, "Emotional speech recognition: Resources, features, and methods," *Speech Commun*, vol. 48, pp. 1162–1181, 2006.

[4] Z. Zhang, F. Weninger, M. Wöllmer, and B. Schuller, "Unsupervised learning in cross-corpus acoustic emotion recognition," in *Proc. of the IEEE ASRU-2011*, Waikoloa, USA, 2011, pp. 523–528.

[5] M. Tahon and L. Devillers, "Acoustic measures characterizing anger across corpora collected in artificial or natural context," in *Int. Conf. on Speech Prosody (SP 2010)*, Chicago, USA, Mai 2010.

[6] M. Brendel, R. Zaccarelli, B. Schuller, and L. Devillers, "Towards measuring similarity between emotional corpora," in *Proc. of the 7th LREC*, Valletta, Malta, 2010, pp. 58–64.

[7] T. Vogt and E. André, "Comparing feature sets for acted and spontaneous speech in view of automatic emotion recognition," in *Proc. of the 2005 IEEE ICME*, Amsterdam, The Netherlands, 2005, pp. 474–477.

[8] J.-L. Xu, B.-W. Xu, W.-F. Zhang, and Z.-F. Cui, "Principal component analysis based feature selection for clustering," in *Int. IEEE Conf. on Machine Learning and Cybernetics*, vol. 1, Kunming, China, July 2008, pp. 460–465.

[9] B. Schuller, D. Arsic, G. Rigoll, M. Wimmer, and B. Radig, "Audiovisual behavior modeling by combined feature spaces," in *Proc. of the IEEE ICASSP-2007*, Honolulu, USA, 2007, pp. 733–736.

[10] I. S. Engbert and A. V. Hansen, "Documentation of the danish emotional speech database des," Center for PersonKommunikation, Aalborg University, Denmark, Tech. Rep., 2007.

[11] F. Burkhardt, A. Paeschke, M. Rolfes, W. Sendlmeier, and B. Weiss, "A database of german emotional speech," in *Proc. of the INTERSPEECH-2005*, Lisbon, Portugal, 2005, pp. 1517–1520.

[12] B. Vlasenko, B. Schuller, K. Tadesse Mengistu, G. Rigoll, and A. Wendemuth, "Balancing Spoken Content Adaptation and Unit Length in the Recognition of Emotion and Interest," in *Proc. of the INTERSPEECH-2008*, Brisbane, Australia, 2008, pp. 805–808.

[13] O. Martin, I. Kotsia, B. Macq, and I. Pitas, "The enterface'05 audio-visual emotion database," in *Proc. of the IEEE Workshop on Multimedia Database Management*, Atlanta, USA, 2006.

[14] B. Schuller, R. Müller, B. Hörnler, A. Höthker, H. Konosu, and G. Rigoll, "Audiovisual recognition of spontaneous interest within conversations," in *Proc. of the 9th ACM ICMI*, 2007, pp. 30–37.

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

[16] S. Steininger, F. Schiel, O. Dioubina, and S. Raubold, "Development of user-state conventions for the multimodal corpus in smartkom," in *Proc. of the Workshop on Multimodal Resources and Multimodal Systems Evaluation*, Las Palmas, 2002, pp. 33–37.

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

[18] J. Hansen and S. Bou-Ghazale, "Getting started with susas: A speech under simulated and actual stress database," in *Proc. of EUROSPEECH-1997*, vol. 4, Rhodes, Greece, 1997, pp. 1743–1746.

[19] 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 Commun*, vol. 53, pp. 1062–1087, 11 2011.

[20] I. Siegert, D. Philippou-Hübner, K. Hartmann, R. Böck, and A. Wendemuth, "Investigation of speaker group-dependent modelling for recognition of affective states from speech," *Cognitive Computation*, vol. 6, no. 4, pp. 892–913, 2014.

[21] F. Eyben, M. Wöllmer, and B. Schuller, "Openear - introducing the munich open-source emotion and affect recognition toolkit," in *Proc. of the 3rd IEEE ACII*, Amsterdam, The Netherlands, 2009, pp. 576–581.

[22] A. Tickle, S. Raghu, and M. Elshaw, "Emotional recognition from the speech signal for a virtual education agent," *J. Phys.: Conf. Ser.*, vol. 450, p. 012053, 2013.

[23] T. Pfister and P. Robinson, "Speech emotion classification and public speaking skill assessment," in *Human Behavior Understanding*, ser. LNCS, A. Salah, T. Gevers, N. Sebe, and A. Vinciarelli, Eds. Springer Berlin Heidelberg, 2010, vol. 6219, pp. 151–162.

[24] F. Eyben, F. Weninger, M. Wöllmer, and B. Schuller, *openSMILE – the Munich open Speech and Music Interpretation by Large Space Extraction toolkit*. Technical University Munich, 2013, no. 2.

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

[26] W. Kruskal and W. A. Wallis, "Use of ranks in one-criterion variance analysis," *J Am Stat Assoc*, vol. 47, pp. 583–621, 1952.

[27] A. Cullen and N. Harte, "Feature sets for automatic classification of dimensional affect," in *Proc. of the 23nd IET Irish Signals and Systems Conf.*, Maynooth, Ireland, 2012, pp. 1–6.

[28] P. Dumouchel, N. Dehak, Y. Attabi, R. Dehak, and N. Boufaden, "Cepstral and long-term features for emotion recognition," in *Proc. of the INTERSPEECH-2009*, Brighton, UK, 2009, pp. 344–347.