Comparing Features for Acoustic Anger Classification in German and English IVR Portals

Tim Polzehl[1], Alexander Schmitt[2], and Florian Metze[3]

[1] Quality and Usability Lab der Technischen Universität Berlin / Deutsche Telekom Laboratories Ernst-Reuter-Platz 7, D-10587 Berlin, Germany
tim.polzehl@telekom.de
[2] Dialogue Systems Group / Institute of Information Technology, University of Ulm Albert-Einstein-Allee 43, D-89081 Ulm
alexander.schmitt@uni-ulm.de
[3] Language Technologies Institute, Carnegie Mellon University, 5000 Forbes Avenue, Pittsburgh, U.S.A.
fmetze@cs.cmu.edu

Abstract. Acoustic anger detection in voice portals can help to enhance human computer interaction. In this paper we report about the performance of selected acoustic features for anger classification. We evaluate the performance of the features on both a German and an American English dialogue voice portal database which contain "real" speech, i.e. non-acted, continuous speech of narrow-band quality. Deploying a large-scale feature extraction we determine the optimal set of features for each language. To obtain the ranking we use an Information-Gain Ratio filter. Analyzing the most promising features we notice a predominance of MFCC and loudness features. However, for the English database also pitch features proved importance. We further calculate classification scores for our setups using discriminative training and Support-Vector Machine classification. The developed systems show that Emotion Recognition in both English and German language can be processed very similarily.

1 Introduction

Detecting emotions in Human Computer Interactive communication is gaining more and more attention in the speech research community. Moreover, classifying human emotions by means of automated speech understanding analysis is gaining performance figures to a level that makes it applicable not only for basic research but also opens up opportunities in deployment systems. Emotion detection in Interactive Voice Response (IVR) Dialogue systems can be used to monitor quality of service or to adapt emphatic dialogue strategies [19, 17]. Especially anger detection can deliver useful information to both the customer and the carrier of IVR platforms. It indicates potentially problematic turns or slots to the carrier so he can monitor and refine the system. It can further serve

as trigger to switch between tailored dialogue strategies for emotional conditions to better react to the user's behavior [11, 3]. Some carriers have also been experimenting with re-routing the customers to the assistance of a human operator when problems occur. Problems and uncertainties arise from the imbalance in complexity between human computer interaction and models trained for these interactions. The difficulty is to capture the various and divers human expression patterns that convery emotional information by automated measurements.

This paper analyzes the importance of different acoustic and prosodic measurements, i.e. we examine expressive patterns that are based on vocal intonation. Applying our system from [12] we capture these expressions extracting low-level audio descriptors, e.g. pitch, loudness, MFCC, spectrals, formants and intensity. Current state of the art acoustic recognition systems operate with feature vectors of static length, i.e. statistics are applied to the descriptors that are calculated from the whole turn length. These statistics mostly encompass moments, extrema, linear regression coefficients and ranges of the respective acoustic contours. Other systems also model the course of acoustic contours by HMMs or other dynamic methods [16], but are mostly outperformed by static approaches.

We gain insight into the importance of our features by ranking them due to their Information-Gain Ratio. Looking at high-ranked features we report on their distribution and numbers in total as well as in relation to each other. We compare our features on two different corpora, i.e. an English and a German corpus both holding telephony conversations with IVR systems.

2 Related Work

Offering as much as 97% accuracy for recognition of angry utterances in a 7 class recognition test performed by humans the TU Berlin EMO-DB [5] bases on speech produced by German speaking professional actors. Here it is important to mention that the database contains 10 pre-selected sentences all of which are conditioned to be interpretable in 6 different emotions and neutral speech. All recordings have wideband quality. When classifying for all emotions and neutral speech automatically Schuller [15] resulted in 92% accuracy. For this experiment he chose only a subset of the EMO-DB speech data that, judged by humans, exceeded a recognition rate of 80% and a naturalness evaluation value of 60%. Eventually, 12% of all utterances selected contained angry speech. He implemented a high number of acoustic audio descriptors such as intensity, pitch, formants, Mel-frequency Cepstral Coefficients (MFCCs), harmonics to noise ratio (HNR), and further information on duration and spectral slope. He compared different classification algorithms and obtained best scores with Support Vector Machines (SVM).

A further anger detection experiment was carried out on the DES database [8] which contains mostly read Dutch speech and also includes free text passages. All recordings are of wideband quality as well. The main difference to the EMO-DB is that the linguistic content had not been controlled entirely during recordings.

The people chose their words according to individual topics. The accuracy for human anger detection for this corpus resulted in 75%. This accuracy bases on a five class recognition test. Schuller results in 81% accuracy when classifying for all emotions. Voting for maximum prior probability class would reach an accuracy of 31% only.

Note that also these results base on acted speech data, containing consciously produced emotions, performed by professional speakers. Human recognition rates were obtained by comparing impressions of the labelers during the perception test with the intended emotions of actors' performances. In cases where there is no professional performance, i.e. when looking at natural speech utterances, we need to rely on the labels of the testers only. To obtain a measurement for consistency of such corpora the inter labeler agreement measurement can be applied. It is the ratio of the chance level corrected proportion of times that the labelers agree to the maximum proportion of times that the labelers could agree. The inter labeler agreement of two labelers is given by Cohen's Kappa. We apply Davies extension of Cohen's Kappa [6] for multiple labelers to give a value of coherence among the labelers.

Lee and Narayanan [10] as well as Batliner [1] used realistic IVR speech data. These experiments use call center data, which is of narrow-band quality. Also the classification tasks were facilitated. Both applied binary classification, i.e. Batliner discriminates angry from neutral speech, Lee and Narayanan classify for negative versus non-negative utterances. Given a two class task it is even more important to know the prior probability of class distribution. Batliner reaches an overall accuracy of 69% using Linear Discriminative Classification (LDC). Unfortunately no class distribution or inter labeler agreement for his corpus is given. Lee and Narayanan reached a gender dependent accuracy of 82% for female and 88% for male speakers. He measured inter labeler agreement with 0.45 for male and 0.47 for female speakers, which can be interpreted as moderate agreement. For both gender classes, constant voting for non-negative class would mean to achieve roughly 75% accuracy already. Note that an accuracy measurement allows for false bias since it follows the majority class to a greater extent than it follows other classes. If the acoustic models fit the majority class to a greater extent this would lead to overestimated accuracy figures. We therefore emphasize the use of balanced performance measurements, such as the f1 measure, which will be discussed in Section 6.

3 Corpora

Nearly all studies on anger detection on narrow-band speech are based on a singular corpus making a generalization of the results difficult. Our aim in this study is to compare the performance of different features when trained and tested on different languages. When comparing existing works on anger detection one has to be aware of essential conditions in underlying database design. The most restricted settings would certainly have prearranged sentences performed by professional speakers (one at a time) recorded in audio studios tol-

erating almost no background noise and performing close capturing of speech signals. Real life speech does not have this setting. The databases we used do have background noise, people do cross- and off-talk, they are free in choice of words and would never pronounce themselves as clearly as trained speakers do. The German database roughly captures 21 hours recordings from a German Interactive Voice Response (IVR) portal. The data can be subdivided into 4683 dialogs, averaging 5.8 turns per dialog. For each turn, 3 labelers assigned one of the following labels: *not angry*, *not sure*, *slightly angry*, *clear anger*, *clear rage* or marked the turns as *non applicable* when encountering garbage. The labels were mapped onto two cover classes by clustering according to a threshold over the average of all voters' labels as described in [4]. Following Davies extension of Cohen's Kappa [6] for multiple labelers we obtain a value of $\kappa = 0.52$ which corresponds to moderate inter labeler agreement. Finally, our training setup contained 1761 angry turns and 2502 non-angry turns. The test setup included 190 angry turns and 302 non-angry turns which roughly corresponds to a 40/60 split of anger/non-anger distribution in the sets. The average turn length after cleaning out initial and final pauses is 1.8 seconds. The English database originats from a US-American IVR portal capable of fixing Internet-related problems jointly with the caller. Three labelers divided the corpus into *angry*, *annoyed* and *non-angry* utterances. The final label was defined based on majority voting resulting in 90.2% neutral, 5.1% garbage, 3.4% annoyed and 0.7% angry utterances. 0.6% of the samples in the corpus were sorted out since all three raters had different opinions. While the number of angry and annoyed utterances seems very low, 429 calls (i.e. 22.4% of all dialogues) contained annoyed or angry utterances. In order to be able to compare results of both corpora we matched the conditions to the conditions of the German database, i.e. we collapsed *annoyed* and *angry* to *angry* and created a test and training set according to the 40/60 split. The resulting training set consists of 1396 non-angry and 931 angry turns while the final test set comprises 164 non-angry utterances and 81 utterances of the anger class. The inter labeller agreement in the final set resulted $\kappa = 0.63$, which also resembles moderate agreement. The average turn length after cleaning out initial and final pauses is 0.8 seconds. Details of both corpora are listed in Table 3.

4 Prosodic and Acoustic Modeling

Our prosodic and/or acoustic feature definition provides a broad variety of information about vocal expression patterns that can be useful when classifying speech metadata. Our approach is structured into an audio descriptor extraction unit followed by a unit that calculates various statistics on both the descriptors and certain subsegments of them.

4.1 Audio Descriptor Extraction

The audio descriptors can be sub-divided into 7 groups: *pitch, loudness, MFCC, spectrals, formants, intensity* and *other* features. All descriptors are extracted using 10ms frame shift.

	German	English
Domain	Mobile	Internet Support
Number of Dialogs in Total	4682	1911
Duration in Total	21h	10h
Average Number of Turns per Dialog	5.7	11.88
Number of Raters	3	3
Speech Quality	Narrow-band	Narrow-band
Deployed Subsets for Anger Recognition		
Number of Anger Turns in Trainset	1761	931
Number of Non-Anger Turns in Trainset	2502	1396
Number of Anger Turns in Testset	190	81
Number of Non-Anger Turns in Testset	302	164
Average Utterance Length w/o Initial or Final Turn Pauses in Seconds	1.80	0.84
Average Duration Anger in Seconds	3.27 $\pm$2.27	1.87 $\pm$0.61
Average Duration Non-Anger in Seconds	2.91 $\pm$2.16	1.57 $\pm$0.66
Cohen's Extended Kappa	0.52	0.63

Table 1. Database comparison of both corpora.

Regarding the group of perceptually motivated acoustic measurements we extract *pitch* by autocorrelation as described in [2]. To avoid octave jumps in pitch estimation we post-process a range of possible pitch values using relative thresholds between voiced and unvoiced candidates. Remaining octave confusions between sub-segments of a turn are further processed by a rule-based path finding algorithm. After converting pitch into the semitone domain we apply piecewise cubic interpolation and smoothing by local regression using weighted linear least squares.

Another perceptively motivated measurement is the *loudness* as defined by [9]. This measurement operates on a Bark filtered version of the spectrum and finally integrates the filter coefficients to a single loudness value in sone units per frame. We further filter for the Mel Domain. After filtering a discrete cosine transformation (DCT) gives the values of the Mel frequency cepstral coefficients (*MFCC*). We extract a number of 16 coefficients and keep the zero coefficient. Although MFCCs are most commonly used in speech recognition tasks they often give excellent performance in emotion detection tasks as well.

Further spectral features are the center of spectral mass gravity (Centroid), the 95% roll-off point of spectral energy and the the magnitude of spectral change over time, also known as spectral flux. These features will be referred to as *spectrals* in the following experiments.

Due to telephony speech quality we extract 5 formant center frequencies estimate the formants' bandwidths. Taken directly from the speech signal we extract the contour of *intensity*.

Referred to as *other* features we calculate the harmonics-to-Noise Ratio (HNR) and the Zero-Crossing-Rate and the average amplitude of the time signal. We

add a single coefficient for the correlation between pitch and intensity as an individual feature. Finally, taken from the relation of pitched and non-pitched speech segments we calculate durational features such as pause lengths and the average expansion of voiced segments.

4.2 Statistic Feature Definition

The statistic unit derives means, moments of first to fourth order, extrema and ranges from the respective contours in the first place. Special statistics are then applied to certain descriptors. Pitch, loudness and intensity are further processed by a DCT. Applying DCT to these contours directly we model the behavior over time. There exist different norms of DCT calculation. We refer to a DCT type III which is defined as:

$$X_k = \frac{1}{2}x_0 + \sum_{n=1}^{N-1} x_n \cos[\frac{\pi}{N}n(k+\frac{1}{2})] \qquad k = 0, \ldots, N-1 \tag{1}$$

A high correlation of a contour with the lower coefficients indicates a rather slowly moving time behavior while mid-range coefficients would rater correlate with fast moving audio descriptors. Higher order coefficients would correlate with micro-prosodic movements of the respective curves, which corresponds to a kind of shimmer in the power magnitude or jitter in pitch movement.

However, a crucial task is the time normalization. Dealing with IVR speech we usually deal with very short utterances that often have command-like style. We suppose, every turn is a short utterance of one prosodic entity. Consequently we calculate our statistics to account for whole utterances, i.e. we apply static feature length modeling. Although this seems suboptimal for longer utterances that might hold more than one emotion in a single turn we keep this approach for the current experiments due to the short average turn length.

In order to exploit the temporal behavior at a certain point in time we appended first (D) and second order (DD) derivatives to the contours and calculated statistics on them alike.

As some features tend to give meaningful values only when applied to specific segments, such as voiced or unvoiced segments, we developed an extended version of the speech-silence detection proposed by [13]. After having found certain voiced points we move to the very first and the very last point now looking for adjacent areas of relatively high zero-crossing rates. Also any non-voiced segment in between the outer borders is classified into high and low zero-crossing regions corresponding to unvoiced or silent speech segments. Eventually, we calculate features on basis of voiced and/or unvoiced sounds both separately and jointly. In order to capture magnitudes of voiced to unvoiced relations we also compute this quotient as a ratio measurement. We apply it to audio descriptors such as intensity and loudness to obtain:

- Ratio of mean of unvoiced to mean of voiced points
- Ratio of median of unvoiced to median of voiced points

– Ratio of maximum of unvoiced to maximum of voiced points

In some utterances we notice an absence of unvoiced sounds. In fact the English database includes less unvoiced sounds than the German does. This can be due many reasons. First, standard English language usually entails a lower level of pressure when producing unvoiced sounds, e.g. fricatives and especially the glottal "h" sound. Also the phonological strong aspiration is normally expected to occur with less pressure in English. Thus in English language these sounds may be harder to be found. Moreover they may be harder to detect from ZCR and our detection algorithm may fail. Secondly, this can also refer to a difference in speaking style. The average utterance length of English samples shows nearly half the length of German utterances. This could indicate a more command-like speaking style, e.g. omitting words that are not necessary, consequently being less outspoken. After all, 16% of all utterances in the German train set and 22% of all utterances in the German test set were of no unvoiced sound share. For the English database these figures raised to 27% and 33% respectively.

All in all, we obtained some 1450 features. Table 4.2 shows the different audio descriptors and the number of features calculated from them. Table 4.2 also shows figures of f1 performance, which will be discussed in the Section Ranking and Section Classification. Note that the different number of features can take bias on the performance comparison. Further experiments will report on individual feature performance by applying feature ranking.

Feature Group	**Number of Features**	**f1 Performance on German DB**	**f1 Performance on English DB**
pitch	240	67.7	72.9
loudness	171	68.3	71.2
MFCC	612	68.6	68.4
spectrals	75	68.4	69.1
formants	180	68.4	67.8
intensity	171	68.5	73.5
other	10	56.2	67.2

Table 2. Feature Groups and Performance on German and English Databases.

5 Feature Ranking

In order to gain insight about which of our features can be useful for the given classification task we applied a filter-based ranking scheme, i.e. Information-Gain-Ratio (IGR) [7]. This measure evaluates the gain in information that a single feature contributes in adding up to an average amount of information needed to classify for all classes. It is based on the Shannon Entropy H [18] for a

class distribution $P(p_1, \ldots, p_K)$ of P samples which is measured in bit unit and defined as

$$H = -\sum_{i=1}^{K} p_i \cdot log_2(p_i) \tag{2}$$

Now let Ψ be the totality of our samples and $\Psi_i \in \Psi$ the subset of elements that belongs to class index i. The average information needed in order to classify a sample out of Ψ into a classes $i_1 \ldots i_K$ is given by

$$H(\Psi) = -\sum_{i=1}^{K} p_i \cdot log(p_i) \qquad with \qquad p_i = \frac{|\Psi_i|}{|\Psi|} \tag{3}$$

To estimate the contribution of a single feature every unique value is taken as partition point. For non-discrete features discretization has to be executed. Let $\Psi_{x,j}$ with $j = 1 \ldots J$ bins be the partion blocks of Ψ_x, holding values of a single feature x, the amount of information contributed by this feature is given by

$$H(\Psi|x) = \sum_{j=1}^{J} \frac{|\Psi_{x,j}|}{|\Psi|} \cdot H(\Psi_{x,j}) \tag{4}$$

The Information Gain (IG) of a feature is then given as its contribution to reach the average needed information for classification.

$$IG(\Psi, x) = H(\Psi) - H(\Psi|x) \tag{5}$$

The Information Gain Ratio accounts for the fact that IG is biased towards features with high number of individual values in their span. IGR normalizes IG by the amount of total information that can be drawn out of J splits.

Table 3 presents the 15 top-ranked features for the English and the German corpus according to IGR. To obtain a more general and independent ranking we performed 10-fold cross validation. The ranking presented accounts for the average ranking throughout the folds. For the English database almost all features are of loudness descriptor origin predominantly capturing the moments of the contour or its maximum and range applied to the original contour, not its derivatives. The picture is much more diverse when we look at the German ranks. Although the loudness features that are present are of the same kind as those on the English set we note also formant, MFCC and intensity descriptors.

Figure 1 shows the relative distributions of the feature sets grouped to their audio descriptor's origin when expanding the feature space from 50 top-ranked features to 500 top-ranked features. Comparing ranks we notice that the top 50 ranks of the English database are occupied by intensity, spectrals and predominatly loudness features only. Pitch, formants and MFCC descriptors did not generated top rank features within the top 100 ranks. However, beyond this

German Database	English Database
intensity DCT coeff2	loudness max
loudness std	loudness std of voiced points
loudness max	loudness std
5th formant bandwidth std	loudness mean
5th formant std	loudness inter-quartile range
intensity err. of lin.reg over voiced points on DD	loudness mean voiced points
loudness std of voiced points	intensity skewness of voiced points
loudness DCT coeff1 on DD	loudness inter-quartile range of voiced points on D
intensity err. lin.reg over voiced points of D	loudness median
loudness inter-quartile range	loudness median over voiced points
loudness DCT coeff2 on D	loudness DCT coeff16
MFCC coeff15 std over whole utterance	loudness std voiced points on D
loudness mean voiced points	loudness DCT coeff26
pitch lin.reg over D of whole utterance	loudness DCT coeff12
MFCC coeff1 min on voiced segments	loudness max unvoiced points

Table 3. Top 15 ranked features for German and English databases.

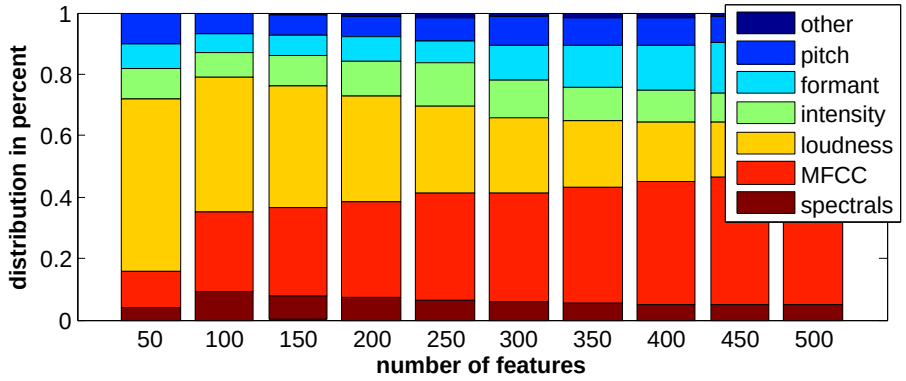
point pitch features become much more important for the English database than for the German.

Table 2 already suggests that the audio descriptor groups are of more counterballanced importance for the German database than they are for the English one. Also the feature distribution in the top-ranks suggest a more heterogeneous distribution in the German set. In General it seems as for the German set loudness and MFCCs are building the most important descriptors. The more features the more important becomes the MFCC contours. Note that also here the absolute number of MFCCs features affects the distribution more and more when the feature space expands.

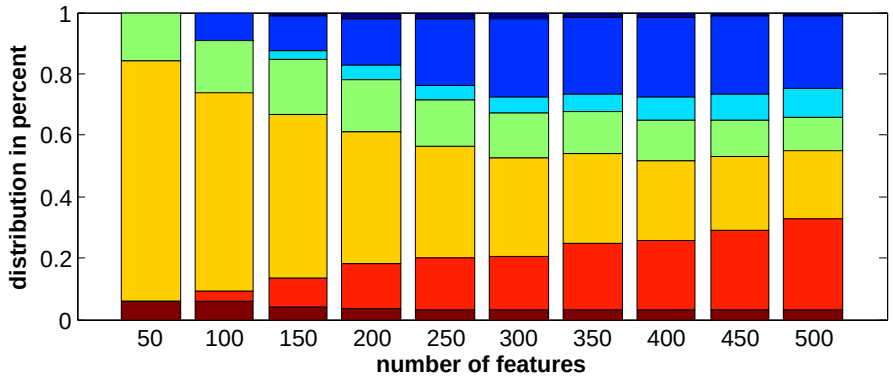
For the English database it seems as the more features are included, the more the picture resembles a three-fold situation, i.e. loudness, MFCc and pitch are of most importance. Also cross-comparing the languages it seems that loudness is of higher impact for the English language as there are consistently more loudness features among all sizes of feature spaces for the English language. On the opposite, MFCC descriptors are more important to the German language. Note that these charts do not tell about how good the classification would be. This issue is discussed in the following Section.

6 Classification

In order to compare results from different feature sets we calculate classification success using the f1 measurement. The f1 measurement is defined as the arithmetic mean of F-measures from all classes. The F-measure accounts for the



(a) German corpus



(b) English corpus

Fig. 1. Percentage of features within each feature group when considering the 50 to 500 top ranked features. Number of features in total = 1450.

harmonic mean of both precision and recall of a given class. Note that an accuracy measurement would allow for false bias since it follows the majority class to a greater extent than it follows other classes. Since our class distribution is unbalanced and our models tend to fit the majority class to a greater extent this would lead to overestimated accuracy figures. To obtain classification results we apply 10-fold speaker independent cross validation on the training set. We also keep an holdout set (test set) for evaluation.

For classification we use a Support Vector Machines (SVM). SVMs view data as two sets of vectors in a multi-dimensional space. Building a miximal margin in between the classes the algorithm constructs a separating hyperplane in that space. We used an linear kernel function and applied standardization before classification, i.e. every feature was mean-substracted and scaled to a standard deviation equaling one.

Table 2 already gives the f1 measurements for classification of features coming from single audio descriptors. To determine the optimal number of top-ranked features to be included into the feature space we move along the IGR ranking and incrementally include a fixed number of top-ranked features. Figure 2 shows the development of f1 measurement by incremental expansion.

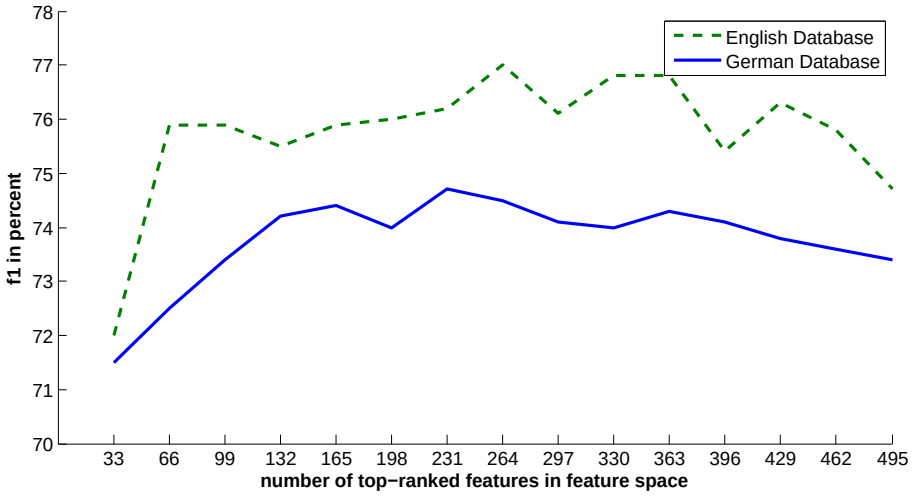


Fig. 2. Determination of optimum feature number by incremental expansion of the feature set according to the top ranked features.

The optimal number of top-ranked features to include into the feature space resulted in 231 for the German database and 264 for the English database. Looking at figure 1 once more we can clearly see that on basis of the English database there is a higher number of pitch and loudness features in the top 250 feature space whereas in the German database more MFCC features can be found. Note that the saw-like shape of graphs in Figure 2 indicate a non-optimal ranking since some feature inclusions seem to harm the performance. However, the magnitude of the jitter here is as low as approx. 1% which after all proves a generally reasonable ranking. If desired, one could also stop at the first local maxima of the f1 curves resulting in a reduced feature set of 66 features for the English and 165 features for the German database without loosing more than 1% f1.

In a final step we adjusted the complexity of our classification algorithm which results in a best score of 78.2 f1 for the English and 74.7 f1 for the German database. Previous studies on both corpora yielded a much lower performance compared to our new findings. The former system described in [14] with the English database reached 72.6% f1 while the system described in [4] developed for the German database reached 70% f1. The performance gain on the

training set of respectively 5.6% and 4.7% f1 in our study can be attributed to the employment of the enhanced feature sets and the feature selection by IGR filtering.

Applied to the holdout sets we obtain figures presented in Table 4. For both sets the features captured more of Non-Anger information that of Anger information. Consequently the F-meassure of Anger class is always lower as for the Non-Anger class. We also see a better recall of English Anger. At the same time we see a better precision in German classification. After all, the overall performance of the final systems proved equivalent.

Database	**Class**	**Recall**	**Precison**	**F-Measure**	**f1-Measure**
German	Non-Anger	88.9%	84.9%	86.7%	77.2%
	Anger	63.7%	72.0%	67.6%	
English	Non-Anger	82.3%	86.0%	84.1%	77.0%
	Anger	72.8%	67.0%	69.8%	

Table 4. System Performance Figures on Test Sets.

7 Results and Discussion

We have shown that detecting angry utterances from IVR speech by acoustic measurements in English language is similar to detecting those utterances in German language. We have set up a large variety of acoustic and prosodic features. After applying IGR filter based ranking we compared the distribution of the features for our languages. Working with both languages we determine an absolute optimum when including 231 (German database) and 264 (English database) top-ranked features into the feature space. With respect of the maximum feature set size of 1450 these numbers are very close. Also the actual features included in the optimal size show accordance. Features derived from filtering in the spectral domain, e.g. mfcc, loudness, seem most promising for both databases, accounting for more than 50% of all features. However, MFCCs occur more frequently under the top-rankled features when operating in the German database, while operating on the English database loudness features are more frequently among top ranks. Another difference lies within the impact of pitch features. Although they are not among the top 50 features they become more and more important when including up to 300 features. They account for roughly 25% when trained on the English database while the number is as small as roughly 10% when trained on the German corpus.

One hypothesis for explaining the differences could be that callers may have dialed in via different transmission channels using different encoding paradigms. While the English database mostly comprises calls that were routed through land line connections the German database accounts for a greater share of mobile

telephony transmission channels. Because fixed line connections transmit usually less compressed speech it can be assumed that there is more information retained in it. However, it is hard to conclude from the signal quality to the impact on our emotion detection task. More information transmitted does not automatically mean more relevance to anger classification.

Another hypothesis for explaining the differences in the results could be the discrepancy in average turn length. The turn length can have a huge effect on statistics when applying a static feature length classification strategy. To estimate the impact of the average turn length we subsampled the German database to match the English average turn length. We processed the subsamples analog to the original database. As a result we obtain major differences in the ranking list when operating on the shorter subset. While MFCC features account for roughly 35% in the original German database the number drops to 22% on the subset. Accordingly, this figure becomes closer to the figure of 18% when working on the English corpus. Consequently we can hypothesize that the longer the turn the more important the MFCC features become. A possible explanation could be the increasing independence of the MFCC to the spoken context when drawing features on turn length. Though 70% of the MFCC features on the original set are also among the top ranked features on the subset the differences seem to be concentrated on the features drawn from the voiced parts. Also the higher coefficients seem to be affected from replacement. Future work will need to focus on these results.

On the other hand, loudness and pitch features tend to remain on the original ranks when manipulating the average turn length. After all we still observe a large difference between the German and the English database when looking at pitch features. Subsampling did not have any significant effect, consequently this difference is not correlated with the average turn length. On basis of this findings we can further hypothesize that there might exist a larger difference in pitch usage in between German and English language at a linguistic level.

Finally the procedures of training the labelers and the more precise differences in IVR design and dialogue domain could be considered as possible factors of influence as well. Also ,as the English database offers a higher value of inter labeler agreement we would expect a better classification score for it. After all, though the classification results on the training sets mirrow this difference they seem very balanced when classifying on the test sets. However, a difference in performance between test and train sets which accounts for less than 4% seems to indicate reasonable and reliable results for our anger detection system on both corpora.

References

1. Anton Batliner, K. Fischer, Richard Huber, J. Spilker, and Elmar Nöth. Desperately seeking emotions: Actors, wizards, and human beings. In *Proc. ISCA Workshop on Speech and Emotion*, 2000.
2. Paul Boersma and David Weenink. Praat: doing phonetics by computer (version 5.1.04), April 2009.

3. Felix Burkhardt, van Ballegooy M., and R. Huber. An emotion-aware voice portal. In *Proceedings of Electronic Speech Signal Processing ESSP*, 2005.
4. Felix Burkhardt, Tim Polzehl, Joachim Stegmann, Florian Metze, and Richard Huber. Detecting real life anger. In *Proc. of ICASSP*, April 2009.
5. Felix Burkhardt, M. Rolfes, W. Sendlmeier, and Benjamin Weiss. A database of german emotional speech. In *Proc. of Interspeech 2005*. ISCA, 2005.
6. M. Davies and J.L. Fleiss. Measuring agreement for multinomial data. volume 38, 1982.
7. Richard O. Duda, Peter E. Hart, and David G. Stork. *Pattern Classification*. 2nd edition, 2000.
8. I. S. Enberg and A. V. Hansen. Documentation of the danish emotional speech database. Technical report, Aalborg University, Denmark, 1996.
9. Hugo Fastl and Eberhardt Zwicker. *Psychoacoustics: Facts and Models*. Springer, Berlin, 3rd edition, 2005.
10. Chul Min Lee and S. S. Narayanan. Toward detecting emotions in spoken dialogs. *Speech and Audio Processing, IEEE Transactions on*, 13(2):293–303, March 2005.
11. Florian Metze, Roman Englert, Udo Bub, Felix Burkhardt, and Joachim Stegmann. Getting closer: tailored humancomputer speech dialog. *Universal Access in the Information Society*, 2008.
12. Tim Polzehl, Shiva Sundaram, Hamed Ketabdar, Michael Wagner, and Florian Metze. Emotion classification in children's speech using fusion of acoustic and linguistic features. In *Emotion Challenge Benchmark, Interspeech*, 2009.
13. Lawrence Rabiner and M. R. Sambur. An algorithm for determining the endpoints of isolated utterances. *The Bell System Technical Journal*, 56:297–315, February 1975.
14. Alexander Schmitt, Tobias Heinroth, and Jackson Liscombe. On nomatchs, noinputs and bargeins: Do non-acoustic features support anger detection? In *Proceedings of the 10th Annual SIGDIAL Meeting on Discourse and Dialogue, SigDial Conference 2009*, London, UK, September 2009. Association for Computational Linguistics.
15. Björn Schuller. *Automatische Emotionserkennung aus sprachlicher und manueller Interaktion*. Dissertation, Technische Universität München, München, 2006.
16. Björn Schuller, Stefan Steidl, and Anton Batliner. The interspeech 2009 emotion challenge. In *Proc. of the International Conference on Speech and Language Processing (ICSLP)*, sep 2009.
17. I. Shafran, M. Riley, and M. Mohri. Voice signatures. In *Automatic Speech Recognition and Understanding, 2003. ASRU '03. 2003 IEEE Workshop on*, pages 31–36, Nov.-3 Dec. 2003.
18. C. E. Shannon. A mathematical theory of communication. *Bell system technical journal*, 27, 1948.
19. Sherif Yacoub, Steven Simske, Xiaofan Lin, and John Burns. Recognition of emotions in interactive voice response systems. In *Proc. Eurospeech, Geneva*, pages 1–4, 2003.