Approaching Multi-Lingual Emotion Recognition from Speech - On Language Dependency of Acoustic/Prosodic Features for Anger Detection

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

[1]Deutsche Telekom Laboratories / Quality and Usability Lab, Technische Universität Berlin
[2]Dialogue Systems Group Institute, Information Technology University of Ulm
[3]Language Technologies Institute, Carnegie Mellon University, Pittsburgh

tim.polzehl@telekom.de, alexander.schmitt@uni-ulm.de, fmetze@cs.cmu.edu

Abstract

This paper reports on mono- and cross-lingual performance of different acoustic and/or prosodic features. We analyze the way to define an optimal set of features when building a multilingual emotion classification system, i.e. a system that can handle more than a single input language. Due to our findings that cross-lingual emotion recognition suffers from low recognition rates we analyze our features on both an American English and a German database. Both databases contain speech of real-life users calling into interactive voice response (IVR) platforms. After calculating performance scores when cross-lingual decoding is involved, i.e. when an emotion classification system is confronted with a language it has not been trained on, we further report on different strategies to build a single feature space that is capable of dealing with both languages. We estimate the relative importance of different features for different languages by looking at their distribution, their classification scores and their rank in terms of information gain ratio. Finally, we construct a feature space on the joint data, replacing two formerly separated system by a single on. We obtain a bi-lingual emotion recognition system which performs as well as the monolingual systems on the test data.

Index Terms: emotion recognition, anger detection, IVR speech, IGR, acoustic prosodic features, speech processing

1. Introduction

We analyze the performance of different features for an anger detection task from customers callin into IVR speech dialogue platforms. Detecting emotions in the dialogues can be useful when monitoring quality of service or when designing more natural dialogue adaptation strategies. Especially anger detection can deliver useful information to both the customer and the carrier of IVR platforms. As it indicates potentially problematic turns to the carrier it can further be applied to trigger different dialogue adaptation steps when striving to meet the expectation of customers who are already in disposition with a system.

Dealing with IVR speech utterances one usually deals with very short utterances, many of which are one-word sentences. Modeling the acoustic and/or prosodic characteristics of these utterances accounts for capturing the expressive intonation of expressive speech patterns. This study reports on similiarities and differences in feature performance and the optimal way to build a multi-lingual emotion recornition system.

Previous research reported on acoustic and linguistic anger classification systems that operate on a single language [1, 2, 3, 4, 5]. We now examine the performance of a broad variety of acoustic features on two different languages, i.e. English and German, when integrating the separate systems into a unified system capable of processing both languages. We calculate a ranking of best performing features for both languages individually, join the most promising features and analyze the cross-language performance. As a result we obtain a robust feature subset that gives best results when applied to combined English and German speech input.

Database characteristics are given in Section 2. Section 3 explains our static feature modeling approach. After ranking and classifying our features as described in Section 4 we look at high-ranked features for both sets and analyze their distributions in Section 5. Different experiments to combine the most promising features of both languages are given in the respective sub-sections.

2. Databases

Our databases consist of narrow band telephony speech recorded on IVR platforms dealing with issues of providing support in Internet and telephony related services and troubleshooting. Both of the databases are of real-life conditions, i.e. they do have background noise, people do cross- and off-talk, they are free in choice of words, hesitate and speak in an conversational style.

The German database roughly captures 21 hours recordings. 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 [6]. Taking a subset for experiments from the original set, 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. Following Davies extension of Cohen's Kappa [7] we obtain a value of $\kappa = 0.52$ which corresponds to fair inter labeler agreement. The average turn length after cleaning out initial and final pauses is 0.84 seconds.

The English database originates from a US-American IVR portal. 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 taken 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. Deducing a relevant sub set 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. We measure substantial agreement in Kappa of $\kappa = 0.63$. The average turn length after cleaning out initial and final pauses is 1.8 seconds.

3. Prosodic and Acoustic Modeling

In our prosodic and acoustic feature definition we calculate a broad variety of information about vocal expression patterns that can be useful when classifying speech metadata. Dealing with IVR speech we usually deal with very short utterances. We interpret every turn as 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. Our feature definition consists of two consecutive units: an initial audio descriptor extraction unit followed by a unit that calculates various statistics on both the descriptors and certain subsegments of them.

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.

Pitch features are calculated by autocorrelation. After converting it into the semitone domain we apply piecewise cubic interpolation and smoothing by local regression using weighted linear least squares. We make use of relative thresholds and a rule-based path finding algorithm to prevent octave jumps.

We extract perceptual *loudness* as defined by [8]. 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. Further we 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.

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

Due to narrowband speech quality we extract 5 *formant* frequencies and estimate the resprective 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. 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.

After finishing the extraction of audio descriptors our statistical unit now derives means, moments of first to fourth order, extrema and ranges from the respective discriptors' contours in the first place. Special statistics are then applied to certain contours. Pitch, loudness and intensity are further processed by a DCT in order to model the behavior over time. High correlation with the lower coefficients indicates a rather slowly moving contour while mid-range coefficients would rater correlate to fast moving audio descriptors. Higher order coefficients would correlate to micro-prosodic movements of the respective curves, which corresponds to a kind of shimmer in the power magnitude and jitter in pitch movement.

As some features tend to give meaningful values only when applied to special voice characteristics we do segmentation into voiced, unvoiced and silence segments. We calculate features on basis of this segmentation and alsio append features on their ratio.

In order to exploit the temporal behavior at a certain point in time we append first and second order derivatives to the contours and calculate statistics on them alike.

A more detailed description of the feature definition can be found in [1]. All in all, we obtain about 1450 features. Table 2 gives examples of the final features. Table 1 shows the different audio descriptors and the number of features derived from them. Also the f1 performance measurement on the train set is given when classifying on basis of the respective groups exclusively. The f1-measurement will be discussed in the Section 5.

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 1: Feature Groups and Performance on English and German Databases.

4. Feature Selection and 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 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 Machine with a linear kernel function.

Table 1 shows the f1 measurements for our audio descriptor groups. While on German, all feature groups perform close to equal, intensity, loudness, and pitch perform better than the other features on English.

In order to gain insight into the relevance of individual features we apply a filter-based ranking scheme, i.e. Information-Gain-Ratio (IGR). This entropy-based 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. After estimating the gain of information a normalization by the amount of total information that can be drawn from the span of a feature gives the Information Gain Ratio. We obtain the optimal number of features to include by moving along the top-ranks of all features for a language and incrementally append the next lower ranks into the feature space until the f1 performance reaches a maximum. The optimal numbers of top-ranked features to include into the feature space resulted in 231 for the German database and 264 for the English database. Classification rates are presented in Section 5.1.

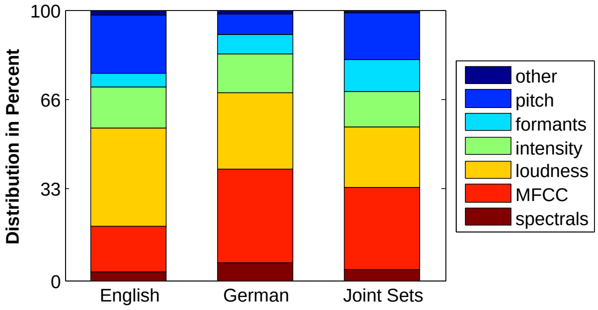

Figure 1: Distribution of Feature Groups in the Optimal Sets

Figure 1 shows the contribution of the different feature groups to the optimal sets. We can clearly see that for the English database a higher proportion of pitch and loudness features are among the top ranks whereas for the German database more MFCC features are amongst top ranks. Further research has shown, that the major difference in MFCC ranks is correlated to the turn length. When subsets of equal length are examined this difference disappears. It was also shown, that differences in pitch and loudness numbers are not vulnerable to turn length. More details can be found in [1].

5. Experiments

We now report on the separate systems and their performance as well as their cross-lingual performance. Note that at these stages the classifiers had not been trained on any other material than the original languages. As a next step we build a shared ranking and analyze which features are promising for both systems. We set up a unified feature space, evaluate its performance and compare it to a system trained and ranked on both databases. Table 3 shows the results of our experiments.

5.1. Mono-Lingual Performance

After ranking our features we obtain f1 classification scores of 78.2% for the English and 74.7% for the German database on the training set. Processing our test sets we obtain f1 figure of 77% on the English database and 77.2% on the German database. Figure 1 shows the distribution of different feature groups in the optimal sets.

5.2. Cross-Lingual Performance

We use the term *cross-lingual* evaluation when a system built on English language decodes a German term or phrase and vice versa. When the system trained on English speech is evaluated on the German test set we recognize a drop of f1 down to 65.2%. That means a loss of 12% in absolute compared to the performance of the mono-lingual German system. If the system trained on the German database is evaluated on the English test set we notice a loss of 6.3% absolute compared to the mono-lingual English system. The f1 of 71.7% shows that although the models built on German speech are more capable of capturing multi-lingual emotional information we still obtain an overwhole decrease of performance for both systems when decoding cross-lingually.

5.3. Ranking Analysis

When looking at the top-ranked features of the separate sets we observe that only 58% of the features are included in both top rankings. These shared features consist of 20% pitch features, 12% formant features, another 12% loudness features, 8% intensity features and 48% features of MFCC origin. However, also the most non-concurrent features, i.e. features that are included in the top ranks of just one database, are of MFCC origin. Consequently, MFCC descriptors are of high importance but serve as basis for different statistics at the same time.

In terms of statistics taking the *maximum* reveals to be highly relevant. Also taking the *mean* seems to be important for both languages. When we examine the actual MFC coefficients that are subject to several statistics we notice a clear predominance of statistics on the first coefficient. Also the lower coefficients ranging until the sixth coefficient seem to be important as well as some scatterly distributed coefficients as the ninth and fourteenth.

Considering the ranks of features drawn from different segments it turns out that statistics calculated on voiced segments exclusively are very frequent among the top ranks, followed by statistics on the whole segments. Statistics on the unvoiced parts are of less importance.

Looking at pitch processing the *minimum* of the first and second order derivatives seems to indicate a reliable feature for both sets. The coefficients from cepstral analysis of pitch movement are frequently found amongst top ranks but, at the same time the actual number of the respective underlying coefficients alter.

For the features calculated on loudness it shows that the *maximum*, the *mean* and most of all the *standard deviation* prove reliability on both corpora. Also features on derivations of the loudness are promising. Coefficients from a discrete cosine transformation (DCT) applied to the loudness directly seem to be of most importance for both databases.

After looking at the distribution of the features we compute an average rank from the former separated ranks. The arithmetic mean of both ranks on the intersection of both sets serves as measure to obtain the new average ranking. Table 2 shows the top 10 ranks.

Average Rank	Feature
1	mean of 1st derivative of loudness
2	min of 10th MFCC on voiced segments
3	10th DCT coeff. of loudness
4	max of 1st deriv. of 14th MFCC on voiced seg
5	kurtosis on 2nd deriv. of intensity
6	10th cepst. coeff. on 2nd deriv. of pitch
7	9th DCT coeff. on 2nd deriv. of loudness
8	std of 1st deriv. of pitch
9	max of 1st deriv. of intensity
10	10th cepst. coeff. on pitch

Table 2: Average Ranking of the Shared Top-Ranked Features from English and German Databases.

5.4. Multi-Lingual Setup

We experiment with two different ways of combining the systems. One way is to take all promising features from both systems into the new feature space. The other is to build a combined data set and rank it as if it was a mono-lingual database.

When combining promising features the unity of the separate top-ranks consists of 375 different features. When learning these features from the German database and evaluating on the German database we get a f1 performance of 75.1% which is a gain of 0.4% absolute on the training database. The fact that the additional features slightly increas the performance is a phenomenon that the IGR filter was not capeable to indicate. This is due to the heuristic filter scheme, leaving out the bias of the classification algorithm. When learning and evaluating the unified features from the English database we get a f1 performance of 78.5% which is again a small gain of 0.3% absolute on the training database. When evaluating on the test set we also notice a slight increase of 0.3% absolute. The increase rises when evaluating the new German models on the German test set. Here we even see a boost of 1.9% in f1 absolute. Consequently, the inclusion of the new features did not jeopardize but increase the system's mono-lingual overall performance.

Now we are interested in the cross-lingual performance. The system trained on English speech obtained a f1 score of 68% on the German test set whereas the system trained on the German set obtained an f1 of 70% when evaluated on the English test set. While improving the scores by 2.8% f1 absolute for the recognition of German emotions we loose 1.7% f1 absolute for the recognition of English emotions. Combining the high-ranked features by unifying them consequently leads a more balanced recognition score for both languages. Calculating the average of both f1 measures we obtain an multilingual average f1 of 69%. However, compared to the monolingual performances of the systems this score seem relatively low. Note that the systems are still trained on the respective original language data exclusively. The low recognition score is also due to the fact that although the most promising features are selected they have not been trained on multi-lingual data.

As a next experiment we now combining the separate databases into a multi-lingual database. When learning the unified feature set from the joint database we obtain 73.1% f1 on the combined test set. Still the feature ranks have been obtained from separate rankings.

As our last experiment we re-rank our features on the basis of the joint data set. Building up the new feature space analogously to the separate sets we obtain an optimal number of 363 top-ranked features. Figure 1 shows the feature group distribution of the joint sets. The corresponding f1 measure is 75.6% on the combined train set. On the test set we achieve 74.5% f1 which is an increase of 6.6% compared to the unified ranking procedure. When the models are evaluated on the original German test set we still obtain 77.7% f1 whereas the evaluation on the original English test set results in 73.2% f1. After all, we show that building an multi-lingual system we are able to keep up the level of performance on the German speech while we lose 3.8% f1 for the English database.

6. Summary and Discussion

In this paper we have analyzed the potential benefits of different features and feature groups for the task of anger recognition from English and German speech. We have shown differences in feature rankings and classification scores in between the data sets. Grouping our features due to both audio descriptor contours and statistics taken from them we have indicated to potentially beneficial groups. MFCC descriptors seem to be of most importance to both languages, followed by loudness and formant information. Also for the English database pitch is given high ranks. When unifying the two rankings for the separate sets we notice a predominance of MFCC statistics in the intersection. Most frequent are coefficients of lower MFCC orders. However, the actual statistics applied to them alter. Most important are the *maximum* and *mean* predominantly calculated on the voiced segments of the utterance. In terms of pitch the derivatives are of most frequent high rank. Coefficients from a DCT are of most importance to the loudness descriptors. Also the *maximum* and the *standard deviation* are high value statistics for loudness.

Setup	English f1	German f1
Mono-Lingual Train Sets	78.2	74.7
Unified Features on Train Sets	78.5	75.1
Mono-Lingual Test Sets	77.0	77.2
Unified Features Test Sets	78.8	77.0
Cross-Lingual Eval. of Mono-Lingual Features on Test Sets	65.2	71.7
Cross-Lingual Eval. of Unified Features on Test Sets	68.0	70.0
Mono-Lingual Eval. of Joint Datbases Sets	73.2	77.7

Table 3: Performance Figures of the Different Database and Feature Combinations.

We have further calculated f1 classification scores for mono- and multi-lingual performance. Table 3 shows the classification scores for the different experiments in feature space setup. After all, we are able to build a multi-lingual emotion recognition system that performs with an f1 of 75.6 on the multi-lingual train set and 74.5 f1 on the multi-lingual test set. While keeping up with the mono-lingual German system performance we notice a slight drop in decoding English emotions when fusing the two systems into one. Further, the applied ranking filter was not always able to indicate to the most beneficial features. Future research using wrapper based approaches will focus on this issue.

7. References

[1] T. Polzehl, A. Schmitt, and F. Metze, "Comparing features for acoustic anger classification in german and english ivr portals," in *International Workshop on Spoken Dialogue Systems (IWSDS)*, Nov. 2009.

[2] T. Polzehl, S. Sundaram, H. Ketabdar, M. Wagner, and F. Metze, "Emotion classification in children's speech using fusion of acoustic and linguistic features," in *Emotion Challenge Benchmark, Interspeech*, 2009.

[3] O. Herm, A. Schmitt, and J. Liscombe, "When calls go wrong: How to detect problematic calls based on log-files and emotions" in *Proc. of the International Conference on Speech and Language Processing (ICSLP)*, sep 2008.

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

[5] I. Shafran and M. Mohri, "A comparison of classifiers for detecting emotion from speech," in *Proc. of ICASSP*, 2005.

[6] F. Burkhardt, T. Polzehl, J. Stegmann, F. Metze, and R. Huber, "Detecting real life anger," in *Proc. of ICASSP*, Apr. 2009.

[7] M. Davies and J. Fleiss, "Measuring agreement for multinomial data," in *Biometrics*, vol. 38, 1982.

[8] H. Fastl and E. Zwicker, *Psychoacoustics: Facts and Models*, 3rd ed. Springer, Berlin, 2005.