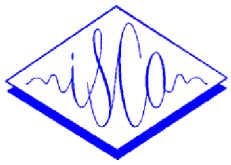

Speech Prosody 2010
Chicago, IL, USA
May 10-14, 2010

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

Tim Polzehl[1], *Alexander Schmitt*[2], *and 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

In this paper, we describe experiments on automatic Emotion Recognition using comparable speech corpora collected from real-life American English and German Interactive Voice Response systems. We compute the optimal set of acoustic and prosodic features for mono-, cross- and multi-lingual anger recognition, and analyze the differences. When an emotion recognition system is confronted with a language it has not been trained on we normally observe a severe system degradation. Analyzing this loss we report on strategies to combine the feature spaces with and without combining and retraining the mono-lingual systems. We report classification scores and feature sets for various cases, and estimate the relative importance of features on both databases. We compare the feature distribution and feature ranks by evaluating information gain ratio. After final system integration, we obtain a single bi-lingual anger recognition system which performs just as well as two separate mono-lingual systems on the test data.

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

1. Introduction

Being able to automatically detect user emotions in speech dialog systems can be useful for a variety of purposes, including monitoring quality of service, or designing more natural dialogs and adaptation strategies. Anger recognition in particular can deliver useful information to the operator of an Interactive Voice Response (IVR) platform, and help to improve the customer experience. Because it can detect potentially "problematic" dialogs or turns, anger recognition can be applied to trigger dialog adaptation steps in order to accommodate the expectation of customers who are already in disposition with the system. At the same time it generates expectations about costs of calls and thus helps to serve callers at a given cost in automated care.

In this work, we use acoustic and prosodic signal characteristics to model expressive speech behavior, such as patterns of intensities, intonations contours or rhythmical characteristics. Previous research reported on acoustic and linguistic anger classification systems that operate on a single language [1, 2, 3, 4]. We now examine a large set of acoustic features on two different languages, American English and German. We calculate a ranking of best performing features for both languages individually, merge the most promising features and analyze the cross-language performance. We obtain a robust subset of features, which gives best results when tested on the combined English and German test sets.

2. Databases

Our databases consist of narrow band telephony speech recorded on IVR platforms in the US and Germany. Both platforms provide self-support for Internet and telephony related services and troubleshooting. The databases include background noise, people use cross- and off-talk, and speak in a conversational style, although we generally observe very short utterances, and many one-word sentences, mainly when navigating menus and answering system questions.

The German database contains 21 hours of 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 [5]. 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 [6] 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 comprises 10 hours of recordings, split into 1911 dialogs. Three labelers divided the corpus into *angry*, *annoyed* and *non-angry* utterances. The final label was generated by majority voting. The resulting distribution comprises 90.2% neutral, 5.1% garbage, 3.4% annoyed and 0.7% angry utterances. 0.6% of the samples in the corpus were removed since all three raters had given different ratings. While the number of angry and annoyed utterances seems very low, 429 calls (i.e. 22.4% of all dialogs) contained annoyed or angry utterances. Deducing a sub that can be compared to the German datrabase we collapsed "annoyed" and "angry" to "angry" and created test and training sets that also have a 40/60 split of anger/non-anger class. 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 moderate agreement, Kappa $\kappa = 0.63$. 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 meta-data. Dealing with IVR speech we usually deal with very short utterances. We therefore 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.

3.1. Audio Descriptors

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 pitch into the semitone domain we apply piecewise cubic interpolation and smoothing by local regression using weighted linear least squares. We use relative thresholds and a rule-based path finding algorithm to prevent octave jumps.

We extract perceptual *loudness* as defined by [7]. 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 features from the spectrum 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 narrow band speech quality we extract 5 *formant* frequencies and estimate the respective bandwidths.

Taken directly from the speech signal we extract the contour of *intensity* in decibel.

Referred to as *other* features we calculate the Harmonics-to-Noise Ratio (HNR), the Zero-Crossing-Rate and features related to speech rhythm. Taken from the relation of voiced to unvoiced speech segments we calculate durational features such as pause lengths and average expansion of voiced segments.

3.2. Statistic Feature Definition

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 descriptors' 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.

Some features give meaningful values when applied to special voice characteristics only. We segmentation into voiced, unvoiced and silence segments. We calculate features on basis of this segmentation and also 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 ($\Delta, \Delta\Delta$) to the contours and calculate statistics on them alike.

A more detailed description of these features 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 descriptor groups 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 Section 4.

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 Figures on English and German Databases

4. Feature Selection and Classification

We calculate classification success using the f1 measurement which is defined as the arithmetic mean of all class-specific F-measures. The F-measure describes the harmonic mean of precision and recall of a given class. Note that an accuracy measurement would allow for false bias since it is influenced by the majority class to a greater extent than by 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.

All results were estimated by applying 10-fold speaker independent cross validation on the training set, i.e. only speakers that are not used for training are in the test set of a fold. We also keep an separate holdout set (global 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 all feature groups seem to perform equally well on the German database, intensity, loudness, and pitch perform better than other feature groups for English speech.

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.

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 in top ranks.

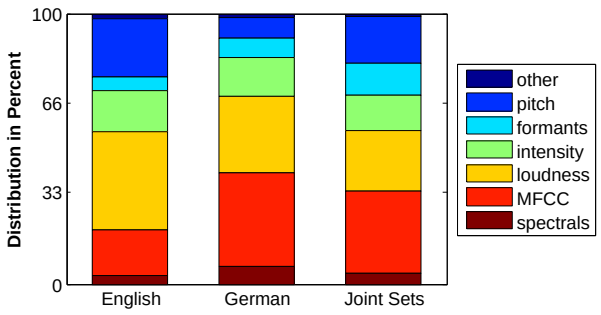

Figure 1: Distribution of Feature Groups in the Optimal Sets

Average Rank	Feature
1	mean of Δ of loudness
2	min of 10th MFCC on voiced segments
3	10th DCT coeff. of loudness
4	max of Δ of 14th MFCC on voiced segments
5	kurtosis on $\Delta\Delta$ of intensity
6	10th cepstral coeff. on $\Delta\Delta$ of pitch
7	9th DCT coeff. on $\Delta\Delta$ of loudness
8	std of Δ of pitch
9	max of Δ of intensity
10	10th cepst. coeff. on pitch

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

After ranking our features we obtain 78.2% f1 for the English and 74.7% for the German database on the training set. On the test sets we obtain 77% f1 for English and 77.2% for German.

5. Multi-Lingual Experiments

We now report on cross-lingual system performance, i.e. when a system built on English language decodes a German phrase and vice versa. Note that at these steps the classifiers had not been trained on any other material than the respective original language data. As a next step we build a unified 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 re-trained and re-ranked on combined databases. Table 3 shows the results of our experiments.

5.1. Cross-Lingual Performance

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 quite capable of capturing multi-lingual emotional information we still obtain an overall decrease of performance for both systems when decoding cross-lingually.

5.2. 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 features that are included in the top ranks of just one database are of MFCC origin. Consequently, MFCCs are of high importance but serve as bases for different statistics.

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 scattered distributed coefficients like 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 be 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* are frequently high-ranked. 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 relative importance for both databases.

We now compute an average rank from the former separated ranks. The arithmetic mean of both ranks serves as measure to obtain the new unified average ranking. Table 2 shows the top 10 ranks of the unified feature set.

5.3. Multi-Lingual Setup

In this step we experiment with two different methods 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 re-rank all features 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. Note that the additional features slightly increase the performance. This is a phenomenon that the IGR filter was not capable to indicate. This is due to the heuristic filter scheme and the ignorance towards 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 using the unified features trained on English speech obtained a f1 score of 68% on the German test set while the system trained on the German database 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 anger we loose 1.7% f1 absolute for the recognition of English anger. Combining the high-ranked features by unifying them eventually leads to a equalization of recognition scores for both languages. Calculating the average of both f1 measures we obtain a multi-lingual average f1 of 69%. However, compared to the mono-lingual 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 can also be 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 combine the separate databases for multi-lingual training. When learning the unified feature set from the combined database we obtain 73.1% f1 on the combined test set. This is a gain of 4.1% f1. Still, the feature ranks have been obtained from unification of formerly mono-lingual rankings.

Finally, we re-rank our features on the basis of the combined data set. Building the new feature space in analogy to the separate sets, we obtain an optimal number of 363 top-ranked features. Figure 1 shows the feature group distribution of the combined set. 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 method. When the models are evaluated on the original German test set we still obtain 77.7% f1 while the evaluation on the original English test set results in 73.2% f1. After all, we show that building a unified multi-lingual system we are able to keep up the level of performance on the German database while losing 3.8% f1 on the English database.

6. Summary and Discussion

In this paper we analyzed different combinations of acoustic and prosodic features for multi-lingual anger recognition. We have shown differences in feature rankings and classification scores in between English and German IVR speech data sets. When merging the two mono-lingual feature rankings for bi-lingual classification, MFCC statistics dominate the unified sets. Within these, the *maximum* and the *mean* calculated on the *voiced* segments are most salient. For pitch, the derivatives are most important. DCT coefficients are most important in the loudness descriptors. Under constrained resources, systems should therefore begin by evaluating these descriptors. Related research has shown, that the difference in MFCC ranks is correlated to the turn length while the difference in pitch and loudness numbers is not [1]. Consequently, time normalization needs to be addressed in future research.

The present work investigated for the first time a multi-lingual emotion recognition system on real-life data. We find the performance of a multi-lingual anger recognition system to be very similar to the mono-lingual systems, so that operational systems for multi-lingual emotion detection seem a possibility, at least for simpler cases, such as anger recognition. We plan to further investigate patterns behind the relative importance of features, in order to be able to eventually attribute the language-dependent importance of certain vocal patterns directly.

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. Eventually, 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 anger when fusing the two systems into one. Similar results were found in [8] when the author compares cross-lingual human anger recognition results. Note, that also differences in call transmission channels applying different compression and encoding schemes could influence the results. However, it is hard to conclude from the signal quality to the impact on our anger recognition task, as more or less information transmitted does not automatically mean more or less relevance to anger recognition. In addition, more subtle differences in IVR design and dialog domain can be influential. Also, the training of labelers can be of impact. Note, that as the English database offers a higher value of inter labeler agreement we would expect a better classification score for it. Finally, the general results of our experiments indicate a reasonable and reliable multi-lingual anger recognition.

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 Databases Sets	73.2	77.7

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

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)*, 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 *International Conference on Speech and Language Processing (ICSLP)*, 2008.

[4] I. Shafran and M. Mohri, "A Comparison of Classifiers for Detecting Emotion from Speech," in *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2005.

[5] F. Burkhardt, T. Polzehl, J. Stegmann, F. Metze, and R. Huber, "Detecting Real Life Anger," in *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2009.

[6] M. Davies and J. Fleiss, "Measuring Agreement for Multinomial Data," in *Biometrics*, vol. 38, 1982.

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

[8] N. Abelin, "Cross-Cultural Multimodal Interpretation of Emotional Expressions - An Experimental Study of Spanish and Swedish," in *Speech Prosody*. ISCA, 2004.