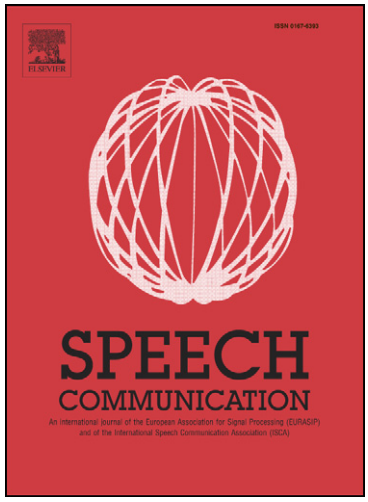
Anger Recognition in Speech Using Acoustic and Linguistic Cues

Tim Polzehl, Alexander Schmitt, Florian Metze, Michael Wagner



Anger Recognition in Speech Using Acoustic and Linguistic Cues

Tim Polzehl, Alexander Schmitt, Florian Metze, Michael Wagner

PII: S0167-6393(11)00067-7
DOI: 10.1016/j.specom.2011.05.002
Reference: SPECOM 1991

To appear in: *Speech Communication*

Received Date: 1 May 2010
Revised Date: 10 February 2011
Accepted Date: 4 May 2011

Please cite this article as: Polzehl, T., Schmitt, A., Metze, F., Wagner, M., Anger Recognition in Speech Using Acoustic and Linguistic Cues, *Speech Communication* (2011), doi: 10.1016/j.specom.2011.05.002

Anger Recognition in Speech Using Acoustic and Linguistic Cues

Tim Polzehl[a], Alexander Schmitt[b], Florian Metze[c], Michael Wagner[d]

[a]*Quality and Usability Lab, Technischen Universität Berlin / Deutsche Telekom Laboratories, Ernst-Reuter-Platz 7, D-10587 Berlin, Germany*
[b]*Dialogue Systems Group / Institute of Information Technology, University of Ulm, Albert-Einstein-Allee 43, D-89081 Ulm, Germany*
[c]*Language Technologies Institute, Carnegie Mellon University, 5000 Forbes Avenue, Pittsburgh, PA 15213, U.S.A.*
[d]*National Centre for Biometric Studies, University of Canberra, ACT 2601, Australia*

Abstract

The present study elaborates on the exploitation of both linguistic and acoustic feature modeling for anger classification. In terms of acoustic modeling we generate statistics from acoustic audio descriptors, e.g. pitch, loudness, spectral characteristics. Ranking our features we see that loudness and MFCC seems most promising for all databases. For the English database also pitch features are important. In terms of linguistic modeling we apply probabilistic and entropy-based models of words and phrases, e.g. Bag-of-Words (*BOW*), Term Frequency (*TF*), Term Frequency - Inverse Document Frequency (*TF.IDF*) and the Self-Referential Information (*SRI*). SRI clearly outperforms vector space models. Modeling phrases slightly improves the scores. After classification of both acoustic and linguistic information on separated levels we fuse information on decision level adding confidences. We compare the obtained scores on three different databases. Two databases are taken from the IVR customer care domain, another database accounts for a WoZ data collection. All corpora are of realistic speech condition. We observe promising results for the IVR databases while the WoZ database shows overall lower scores. In order to provide comparability in between the results we evaluate classification success using the f1 measurement in addition to overall accuracy figures. As a result, acoustic modeling clearly outperforms linguistic modeling. Fusion slightly improves overall scores. With a baseline of approximately 60% accuracy and .40 f1-meaurement by constant majority class voting we obtain an accuracy of 75% with respective .70 f1 for the WoZ database. For the IVR databases we obtain approximately 79% accuracy with respective .78 f1 over a baseline of 60% accurracy with respective .38 f1.

Key words:
emotion detection, anger classification, linguistic and prosodic acoustic modeling, IGR ranking, decision fusion, IVR speech

Introduction

Detecting emotions in vocal human-computer interaction (HCI) is gaining increasing attention in speech research. Moreover, classifying human emotions by means of automated speech analysis is achieving a level of performance, which makes effective and reliable deployment possible. Emotion detection in interactive voice response (IVR) systems can be used to monitor quality of service or to adapt emphatic dialog strategies (Yacoub et al., 2003; Shafran et al., 2003).

Anger recognition in particular can deliver useful information to both the customer and the carrier of IVR platforms. It may indicate potentially problematic turns or slots, which could in turn lead to improvements or refinements of the system. It can further serve as trigger to switch between tailored dialog strategies for emotional conditions to better react to the user's behavior (Metze et al., 2008; Burkhardt et al., 2005a), including the rerouting of customers to a human operator for assistance when problems occur.

There are many ways, in which a person's emotion can be conveyed. However, in the present voice-based scenario, two factors prevail: the choice of words and acoustic variation. When a speaker expresses an emotion while adhering to an inconspicuous intonation pattern, human listeners can nevertheless perceive the emo-

*Corresponding author: Tim Polzehl, Tel: +49-30-8353-58227
Email addresses: tim.polzehl@telekom.de (Tim Polzehl), alexander.schmitt@uni-ulm.de (Alexander Schmitt), fmetze@cs.cmu.edu (Florian Metze), michael.wagner@canberra.edu.au (Michael Wagner)

tional information through the lexical content. On the other hand, words that are not generally emotionally salient can certainly be pronounced in a way, which conveys the speaker's emotion in addition to the mere lexical meaning. Consequently, our task is to capture the diverse acoustic and linguistic cues that are present in the speech signal and to analyze their correlation to the speaker's emotion.

Our linguistic approach analyzes the lexical information contained in the spoken word and its correlation to the emotion of anger. The level of anger connotation of a word can be estimated using various concepts. First, we apply the concept of Emotional Salience (Lee and Narayanan, 2005; Lee et al., 2008), which models posterior probabilities of a class given a word and combines this information with the prior probability of a class. This concept can be extended to include contextual information by modeling the salience of not just one word, but word combinations, i.e. n-grams. Further, we compare these models to traditional models from the related field of information retrieval, i.e. models that estimate term frequencies (TF) or words used (Bag-of-Words, BOW) as explained in Section 3.

Our prosodic approach examines expressive patterns that are based on vocal intonation. Applying large-scale feature extraction, we capture these expressions by calculating a number of low-level acoustic and prosodic features, e.g. pitch, loudness, MFCC, spectral information, formants and intensity. We then derive statistics from these features. Mostly, the statistics encompass moments, extrema, linear regression coefficients and ranges of the respective acoustic contours. In order to gain insight into the importance of our features we rank them according 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. Only the most promising features are retained in the final feature set for acoustic classification.

In a final step, we fuse information from both linguistic and acoustic classification results to obtain a complex estimate of the emotional state of the user.

We compare our features for three different corpora. One database comprises American English IVR recordings (Schmitt et al., 2010), another contains German IVR recordings Burkhardt et al. (2009). Both databases account for mostly adult telephony conversations with customer-care hotlines and contain a high number of different speakers. A third database comprises recordings from a Wizard of Oz (WoZ) scenario conducted with a small number of German children (Steidl et al., 2005).

1. Related Work and Realistic Database Conditions

When comparing existing studies on anger recognition, one has to be aware of the precise conditions of the underlying database design, as many of the results published hitherto are based on acted speech data. Some of these databases include sets of prearranged sentences. Recordings are usually done in studios, minimizing background noise, recording speakers (one at a time) multiple times until a desired degree of expression is reached. Real life speech does not have any of these settings.

As much as 97% accuracy has been reported for the recognition of angry utterances in a 7 class recognition test performed by humans on the TU Berlin EMO-DB (Burkhardt et al., 2005b), which is based on speech produced by German-speaking professional actors. The lexical content is limited to 10 pre-selected sentences, all of which are conditioned to be interpretable in six different emotional and a neutral-speech contexts. The recordings have wideband quality. Experiments on a subset, which featured high emotion recognition rates and high naturalness votes, both by human listeners, resulted in 92% accuracy when Schuller (2006) classified for the emotions and neutral speech automatically.

Comprising mostly read sentences, but also some free text passages, a further anger recognition experiment was carried out on the DES database by Enberg and Hansen (1996). The accuracy for classification into 5 classes in a human anger recognition experiment resulted in 75%. All recordings are of wide band quality as well. Classifying this database automatically, Schuller (2006) reported an accuracy of 81%.

When speakers are not acting, namely when there is no professional performance, we need to rely on the impressions of a number of independent listeners. Since no agreed-upon common opinion exists on how a specific emotion 'sounds', it has become standard practice to take into account the opinion of several raters. To obtain a measurement for consistency of such ratings, an inter-labeler agreement measure is often applied. It is defined as the count of labeler agreements, corrected for chance level and divided by the maximum possible count of such labeler agreements. It should be noted that the maximum agreement also depends on the task, as for example the inter-labeler agreement in a gender recognition task is expected to be higher than that in an anger rating task. We assume that low inter-labeler agreement on the different emotion categories in the training and test data would predict a low automatic classification score, since in cases where humans are uncertain about classification, the classifier would likewise have diffi-

culty in differentiating between the classes. Batliner et al. (2000) further analyzes emotion recognition performance degradations when comparing acted speech data, read speech data and spontaneous speech obtained from a WoZ scenario. Performances on acted speech data were much better in all considered experiments.

Lee and Narayanan (2005) as well as Batliner et al. (2000) used realistic narrow-band IVR speech data from call centers. Both applied binary classification with Batliner et al. (2000) discriminating angry from neutral speech, and Lee and Narayanan (2005) classifying for negative versus non-negative utterances. Given a two class task, it is very important to know the prior probabilities of class distribution. Batliner et al. (2000) reach 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 (2005) reach a gender-dependent accuracy of 81% for female and 82% for male speakers. They measured inter-labeler agreements as 0.45 for male and 0.47 for female speakers, which can be interpreted as moderate agreement. For both gender classes, constant voting for the non-negative class would achieve about 75% accuracy already and - without any classification - would outperform the results obtained by Batliner et al. (2000).

Exploiting acoustic and linguistic information Schuller et al. (2004) and Lee and Narayanan (2005) apply late fusion strategies. Using predominantly acted emotions from the automotive domain Schuller et al. (2004) combines acoustic and linguistic information in order classify into seven emotional states. Extracting few acoustic features, the main difference to the present work lies within the incorporation of linguistic information. He hierarchically clusters individual words into bigger phrase and super-phrase levels using belief networks. Also Lee and Narayanan (2005) uses few acoustic features and combines them with linguistic information using Emotional Salience models by averaging on decision level. He proposes to calculate activations from Emotional Salience scores and calculates accuracies in a gender dependent way.

In order to compare the performance of our linguistic and acoustic models for the different databases, we calculate classification success using two evaluation scores: accuracy and the f1-measure. Given the skewed class distribution, the accuracy measure overestimates if the model of the majority class yields better results than the models for the non-majority classes. As reported above, such an inequality in model performance is not uncommon[1]. We therefore focus on the unit function f1-measurement. It is defined as the (unweighted) average of F-measures from all classes, which in turn account for the harmonic mean of both precision and recall of a given class. However, in order to be comparable to other works, we also show accuracy figures. It should be noted that comparisons between results of studies that use different evaluation measures are thus often biased and, in some cases, may even be invalid.

Publications contributed to the INTERSPEECH 2009 Emotion Challenge (Schuller et al., 2009b) give a good overview of recent developments in terms of classifier diversity and acoustic feature modeling. All participant publications are based on the same training and test corpus definitions, and the results are therefore more comparable than results from single case studies. The present study also includes the benchmark corpus, i.e. the realistic *Aibo* corpus as presented in Section 2. Prevailing classification algorithms applied in the benchmark are Support-Vector-Machines (SVM), Gaussian-Mixture-Models (GMM) as well as the combination of them, i.e. GMM-SVM Super-Vector approach, as presented by Dumouchel et al. (2009). Also, dynamic GMM-HMM approaches, as widely used in speech recognition, were proposed (Vlasenko and Wendemuth, 2009) among other methods. Best scores were generally obtained by fusing the several classifiers at the decision level. All those models are based on acoustic, including prosodic, feature extraction. Polzehl et al. (2009b) proposed to include linguistic knowledge, namely to model information that can be drawn from the words the speakers used. The best systems reached average recalls, which was the primary evaluation criterion, of approximately 70% and an accuracy of 69%, which, after all, represents a small improvement over constant majority class voting only. Overall, as most results from the different systems in the benchmark were very close, the challenge illustrated the difficulty in recognizing emotions from speech. The present study elaborates on the exploitation of both linguistic and acoustic feature modeling and the application of decision fusion.

In the context of humanmachine interaction, analyses of emotional expressions are generally aimed at the design of Embodied Conversational Agents. This predominantly relates to application in automated dialog systems. Related to research on human-computer interaction, also human-human interaction has been analyzed for emotions. Although the matter of interest is identi-

[1] Note, that besides the fact that class skewness affects the accuracy evaluation criterion, it also implies an inequality in amount of training data that can also contribute to performance differences.

cal, a deeper look into the differences reveals that emotionally colored speech is more likely to be encountered in human-human interactions (Devillers et al., 2005). Also intensity and forms can vary. This is due to the level of both, self-restriction while interacting with a system and the system's restriction in interaction context. A comprehensive study on human-human call-center emotion analysis and machine classification can be found in Vidrascu and Devillers (2007); Devillers et al. (2005).

2. Selected Corpora

Nearly all studies on anger recognition are based on single corpora 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 corpora. All of the selected databases account for real life conditions, i.e. they have background noise, recordings include cross- and off-talk, speakers are free in the choice of words and do not enunciate as clearly as trained speakers do.

The German IVR database contains about 21 hours of recordings from a German voice portal. Customers call in to report on problems, e.g. problems with the phone connection. The callers are being preselected by an automated voice dialog before they are passed to an agent. The data can be subdivided into 4683 dialogs, averaging 5.8 turns per dialog. For each turn, three 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 by Burkhardt et al. (2009). Following the extension of Cohen's kappa for multiple labelers by Davies and Fleiss (1982), we obtain a value of $\kappa = 0.52$, which corresponds to moderate inter-labeler agreement (Steidl et al., 2005). Finally, our experimental set contains 1951 Anger turns and 2804 Non-Anger turns which correspond approximately to a 40/60 split of anger/non-anger distribution. The average turn length after removing initial and final pauses is 1.8 seconds. A more detailed description of the corpus can be found in Burkhardt et al. (2009).

The English IVR database originates from a US-American portal designed to solve Internet-related problems jointly with the caller. It helps customers to recover Internet connections, reset lost passwords, cancel appointments with service employees or reset lost e-mail passwords. If the system is unable to help the customer, the call is escalated to a human operator. 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 eliminated because 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 dialogs) contained annoyed or angry utterances. In order to be able to compare results of both corpora we matched the conditions of the English database to the conditions of the German database, i.e. we collapsed *annoyed* and *angry* to *angry* and created test and training sets according to the 40/60 split. The resulting set consists of 1560 Non-Anger and 1012 Anger turns. The inter-labeler agreement results in $\kappa = 0.63$, which also represents moderate agreement. The average turn length after eliminating initial and final pauses is approximately 0.8 seconds. A more detailed description of the corpus can be found in Schmitt et al. (2010).

The German WoZ *AIBO* database consists of children interacting with the AIBO robot dog. 51 children (age 10-13) were recorded in a Wizard-of-Oz scenario. The children were given the task to navigate the robot through a certain course of actions using voice commands. When the robot reacted disobediently, it provoked emotional reactions from the children. The data amounts to 9.2 hours of 16bit/16kHz speech recordings in total. Five labelers annotated the utterances with respect to 10 emotion-related target classes, which were eventually mapped to a binary division between negative (*NEG*), subsuming *touchy*, *angry*, *reprimanding* and *emphatic* labels, and non-negative (*IDL*) utterances, subsuming all other classes, as described in (Steidl et al., 2005). Recordings were split into chunks by syntactic-prosodic criteria. For the present experiments we chose a subset of 26 children[2], which results in 3358 NEG and 6601 IDL chunks corresponding to a 33/66 split. Inter-labeler agreement results in $\kappa = 0.56$. A more detailed description of the corpus can be found in Steidl (2009).

Details of all three corpora are listed in Table 1. While the IVR databases contain different degrees of anger expression in the Anger class, the WoZ database also subsumes other emotion-related states. Thus, more diverse patterns in the WoZ Anger class can be expected. Further, all samples from the selected databases were presented to the labelers chronologically and independently. This way, the history of a turn being part of

[2]This set corresponds to the AIBO *chunk train set* used in the INTERSPEECH 2009 Emotion Challenge (Schuller et al., 2009b).

a dialog course was known to the labelers, i.e. the label decision includes the context of it. The labelers of the IVR databases were familiar with the respective voice portals and linguistic emotion theory. The labelers of the WoZ database were advanced students of linguistics. Rating the turns or chunks, acoustic and linguistic information processing happened simultaneously, i.e. all stimuli were given in audible, not written form. In order to facilitate formal comparisons, we will refer to the NEG and IDL classes in the WoZ database as Anger and Non-Anger classes and consider the given chunks as corresponding to turns.

3. Linguistic Feature Modeling

Linguistic features model the information given by the transcription of the spoken words or by word hypotheses obtained from automatic speech recognition (ASR) of the user's utterances. We investigate the performance of word modeling for anger recognition using four different feature spaces, i.e. Bag-of-Words (BOW), Term Frequency (TF), Term Frequency - Inverse Document Frequency ($TF.IDF$) and the Self-Referential Information (SRI). BOW builds up a feature space by collecting all words from a data set and registering the words contained in an individual document. Our BOW model contains 1 or 0 for present or absent words respectively. As a result, word lists can be very large and feature spaces are sparsely populated. TF refers to a similar method. Instead of marking absence and presence, normalized word counts are registered in a vector space. TF.IDF weights TF by the inverse of the document frequency, i.e. the frequency of documents containing an individual word. BOW, TF and TF.IDF are frequently used in information retrieval tasks, e.g. text classification (Huang et al., 2001).

Departing from the concept of relative entropy between two probability mass functions, we calculate the information of a word with respect to an emotion class. Let $w \in W$ be a word out of a vocabulary, $\epsilon \in E$ an emotion out of all target emotion classes and $P(\epsilon)$ the prior probability of an emotion. The Self-Referential Information about an emotion class, given a posterior probability that a certain word implies a certain emotion, can be estimated by:

$$SRI(\epsilon, w) = log \frac{P(\epsilon|w)}{P(\epsilon)} \tag{1}$$

At the turn level, we sum up the word- and class-specific SRI values and decide for the class of maximum SRI sums. Departing from SRI Lee and Narayanan (2005) calculate the *Emotional Salience* of a word using the mutual information between the probability of a word and the probability of a class. Statistical independence of the two probabilities would result in a mutual information of zero. Consequently the higher the salience, the stronger the correlation with the class labels. Let k be the number of classes, the Emotional Salience is defined as:

$$MI(E, W = w) = \sum_{j=1}^{k} P(\epsilon_j|w) \cdot log \frac{P(\epsilon_j|w)}{P(\epsilon_j)} \tag{2}$$

Table 2 presents 10 examples from the most salient words for the three databases. We also calculate the activation feature proposed by Lee and Narayanan (2005), which weights the SRI word summation with the prior class probabilities at the turn level.

Table 3 shows the classification results of the linguistic features for the three databases obtained by 10-fold cross validation. When the individual test splits of the cross-validation folds contain words of unknown SRI or Emotional Salience, i.e. if they are out-of-vocabulary (OOV), we skip their contribution to the turn summation. If a turn completely consists of OOV words it is classified as belonging to the majority class. In general, we notice an average of 8% OOV for the German databases and 5% for the English, both of with are relatively low figures.

Looking at the classification results, traditional text mining features like BOW, TF and TF.IDF are not suitable for emotion recognition on our databases. Looking at the F-measures of the individual classes we can see, that too many test samples were classified as Non-Anger. Also, unknown test words in the cross-validation splits cannot be assigned to the majority class. SRI and Emotional Salience achieve better F-scores for the Anger class, while at the same time the F-measure of Non-Anger class degrades. Due to weighting, Emotional Salience seems more robust to imbalanced class distributions while SRI is more influenced by the majority class. However, the absolute performance of the features does not exceed the prior class probabilities considerably. For further experiments we retain the most promising features, i.e. SRI and Emotional Salience.

Since SRI and Emotional Salience are dependent on word posteriors, which in turn depend on word counts, we examine the impact differences in database design on our recognition recognition task. Analyzing the impact of differences in number of turns, total number of words, and vocabulary size we generate 10 randomly chosen subsets retaining original class distributions. Original sizes and targeted, subsampled sizes are

Table 1: Database Conditions.

	German IVR	*English IVR*	*German WoZ*
Domain	Mobile	Internet Support	Directing Robot
Number of Dialogs in Total	4682	1911	-
Duration in Total	21h	10h	9.2h
Number of Raters	3	3	5
Speech Quality	Narrow-band	Narrow-band	Wide-band
Deployed Subsets for Anger Recognition			
Number of Speakers	683	417	26
Number of Turns	4515	2328	9959
Number of Words in Total	11812	3709	26157
Average Number of Words per Turn	2.6 ± 3.7	1.6 ± 1.5	2.7 ± 1.7
Vocabulary Size	1179	286	901
Perplexity[a]	233	40	78
Average Utterance Length in Seconds[b]	1.80	0.84	0.87
Average Duration Anger in Seconds	3.27 ± 2.27	1.87 ± 0.61	0.87 ± 0.51
Average Duration Non-Anger in Seconds	2.91 ± 2.16	1.57 ± 0.66	0.87 ± 0.62
Extended Cohen's Kappa	0.52	0.63	0.56

[a] on basis of an uni-gram word model
[b] without initial or final turn pauses

Table 2: Salience of Words.

	German IVR		*English IVR*		*German WoZ*	
Class	*word*	*Salience*	*word*	*Salience*	*word*	*Salience*
Anger	dämlicher *(stupid)*	1.298	wrong	1.321	Schluss *(finish)*	1.567
Anger	teuer *(expensive)*	0.630	operator	0.860	stoppen *(to stop)*	1.040
Anger	doch *(<exasperation>)*	0.578	person	0.795	aufhören *(to end)*	0.756
Anger	warum *(why)*	0.558	please	0.723	faul *(lazy)*	0.513
Anger	falsch *(wrong)*	0.280	support	0.575	endlich *(finally)*	0.325
Non-Anger	korrekt *(right)*	0.751	correct	0.735	brav *(good/obedient)*	0.592
Non-Anger	einfach *(simple)*	0.688	okay	0.350	fein *(fine)*	0.592
Non-Anger	danke *(thanks)*	0.338	right	0.207	schön *(nice)*	0.492
Non-Anger	okay	0.239	ready	0.132	gut *(good)*	0.479
Non-Anger	Bonuspunkte *(bonus points)*	0.140	connected	0.087	okay	0.333

Table 3: Classification Results using Linguistic Information.

	German IVR	*English IVR*	*German WoZ*
f1 | accuracy using BOW	.50 | 54.6%	.47 | 59.2%	.44 | 59.7%
f1 | accuracy using TF	.51 | 53.9%	.47 | 58.0%	.47 | 60.8%
f1 | accuracy using TF.IDF	.51 | 54.6%	.46 | 58.3%	.46 | 61.9%
f1 | accuracy using Emotional Salience	.64 | 66.7%	.58 | 65.3%	.63 | 69.1%
f1 | accuracy using SRI	.63 | 63.9%	.58 | 59.9%	.68 | 68.9%
F_{Anger} | $F_{Non-Anger}$ using BOW	.40 | .64	.21 | .72	.15 | .74
F_{Anger} | $F_{Non-Anger}$ using TF	.39 | .63	.23 | .71	.19 | .74
F_{Anger} | $F_{Non-Anger}$ using TF.IDF	.37 | .64	.20 | .72	.13 | .75
F_{Anger} | $F_{Non-Anger}$ using Emotional Salience	.74 | .53	.76 | .40	.78 | .49
F_{Anger} | $F_{Non-Anger}$ using SRI	.69 | .56	.66 | .51	.74 | .62

given in Table 1. Estimating averages from the random sets we match the conditions between the databases, e.g. by subsampling the German IVR database to match the German WoZ vocabulary size. Each randomly generated subset is further processed by 5-fold cross-validation to calculate SRI and Emotional Salience performances. As a result, none of the matched conditions, i.e. subsampled versions of original databases, results in considerable emotion recognition differences. Hence, SRI and Emotional Salience seem relative robust with regard to these conditions.

When matching the dictionary sizes between the German databases, the IVR database shrinks to 58% of its original number of turns while the perplexity almost stays constant. It shrinks from 233 to 223. As can be seen from Table 1 the databases essentially differ in perplexity, which in other words expresses the amout of confusion when choosing uniformly and independently among all words. Unfortunately, subsampling the database to match the complexity condition does not yield sufficient volumes of data for training. A preliminary classification experiment on the basis of these sets showed an intolerably high standard deviation among the results from cross-validation splits.

Expanding the basic modeling unit from separated words to phrases we include contextual word information by calculating Emotional Salience of phrases, e.g. Metze et al. (2009). Because of a low average word-per-turn rate, we apply phrase modeling including 2 consecutive words only. The resulting vocabularies comprise 4053 entries for German WoZ, 4973 for German IVR and 880 entries for the English IVR database. Table 4 shows the results when all phrases are taken in to account, regardless of their frequency of occurrence in the sets. In order to obtain more robust estimates we set a frequency threshold. Table 4 also shows the scores when admitting only phrases (and words) that occur more than 4 times.

Contextual word inclusion does not improve emotion recognition scores on our corpora. Also, the use of n-gram models directly has been proposed (Steidl, 2009; Shafran and Mohri, 2005) as well as other linguistic features, e.g. part-of-speech (*POS*) or higher semantics representations as reported on by Schuller et al. (2009a). None of respective techniques resulted in significant improvements for the respective corpora reported by the literature. It should be noted that all of the experimental databases in those works were of low average word-per-turn rate as well. Hence, an interpretation of the effect of linguistic context modeling to emotion classification has to be deferred to future experiments including databases with longer utterances. However, Steidl (2009) reports on different word clustering techniques that, after all, seem to be most promising for databases with very short utterances. Since these methods need additional labeling, the respective features will not be available for most applications.

All hitherto presented experiments were conducted using speech transcripts generated from human transcriptionists. We are now focusing on the impact of word hypotheses quality, as obtained from automatic speech recognition (ASR). Metze et al. (2009), Polzehl et al. (2009b) as well as Schuller et al. (2009a) report Word Error Rates (*WER*) of less than 20% and more than 30%, respectively. At the same time, the impact on anger recognition reveals very small.

In the next experiment we therefore focus on the dependency of emotion recognition from text upon automatic transcription quality. We simulate ASR quality by systematically varying the decoding beam of a speech recognizer. Narrowing the beam of the speech recognizer we observed a monotonously rising word error rate[3]. Analyzing the impact on emotion classification we observe, that anger classification stays roughly constant as long as the word accuracy does not fall below a certain threshold. Calculating accuracy scores using Emotional Salience models this threshold was found to be around 40% word accuracy for the Geman WoZ database. Figure 1 shows the graphs of the unweighted average recall and the weighted average recall, i.e. accuracy, of angry utterances. Starting at our best word accuracy of 82%, we observe a small decrease of roughly 5% in emotion recognition when downgrading ASR accuracy to approximately 40%.

As an interesting result, word errors are not strongly harming the emotion recognition system as long as the errors occur consistently. Looking more closely, wrongly recognized words are correlated to emotion classes when building the Emotional Salience model. When the same error occurs in decoding, the erroneous token nevertheless link to the respective emotion class. After all, it can even be more suitable to use automatic, i.e. erroneous transcripts instead of human transcriptions because the resulting models are more compact and show higher coherence in terms of training-testing mismatch.

Focusing on the impact of training-testing mismatch Figure 1 shows a comparison between a emotion recognizer trained on best possible transcripts for training of emotion models and a second m̈atched" recognizer

[3]Note that different speech recognizers can behave in different ways when narrowing the decoding beam since word error rates also depend on insertion and deletion rates.

Table 4: SRI and Emotional Salience of Phrases.

	German IVR	*English IVR*	*German WoZ*
f1 SRI all words	.62	.58	.67
f1 Emotional Salience all words	.62	.57	.63
f1 SRI min 4 words	.62	.57	.67
f1 Emotional Salience min 4 words	.61	.56	.64

trained with aligned transcript quality for both training and testing. The absolute difference results in roughly 2% only. Thus we believe, it is not essential to have high-effort transcription for training in order to achieve good results when testing. As long as ASR errors happen systematically the emotion models will capture the class relevant information, even given erroneous word hypotheses. A more detailed description of the speech recognition systems can be found in Metze et al. (2010).

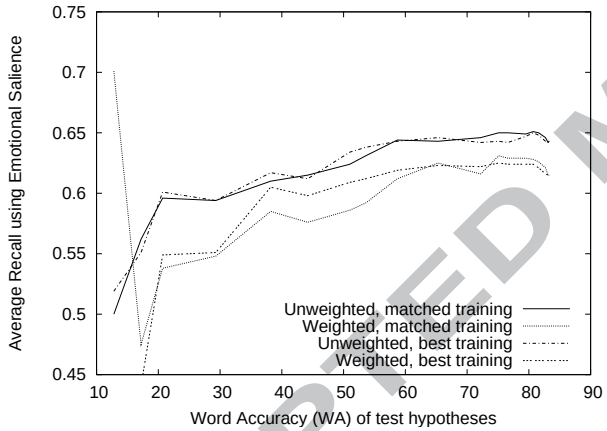


Figure 1: Impact of ASR Word Accuracy on Emotion Classification.

4. Acoustic Modeling

4.1. Feature Definition

Our acoustic feature definitions, including spectral and prosodic features, provide a broad range of information about vocal expression patterns that can be useful when classifying speech metadata. Our approach is structured into two consecutive steps. In the first step an audio descriptor extraction unit processes the raw audio format and provides speech descriptors. In the second step a statistics unit calculates various statistics on both the descriptors and certain sub-segments of them. All descriptors are extracted using 10ms frame shift. For any windowing we used Gaussian windows. The resulting audio descriptors can be sub-divided into 7 groups: *pitch, loudness, MFCC, spectrum, formants, intensity* and *other*.

We extract *pitch* using autocorrelation as described by Boersma and Weenink (2009). Octave confusions between sub-segments of a turn are further processed by a rule-based path finding algorithm. We convert pitch into the semitone domain using the mean pitch as the reference value for a whole turn. We apply piecewise cubic interpolation and smoothing by local regression using weighted linear least squares.

We calculate *loudness* as defined by Fastl and Zwicker (2005). This measure operates on a Bark filtered version of the spectrum and finally integrates the filter coefficients into a single loudness value in sone units per frame.

We further filter the spectrum into the mel domain and apply a discrete cosine transformation (DCT) to obtain 16 mel-frequency cepstral coefficients ($MFCC$).

Other features drawn from the *spectral* representation are the spectral centroid and the 95% roll-off point of spectral energy. Both features capture aspects related to the spectral slope (also called the spectral tilt) and correspond to perceptual impression of sharpness and brightness of sounds (Fastl and Zwicker, 2005). Abrupt changes in the spectrum are captured by calculating the spectral flux.

We extract five $formant$ frequencies and estimate their bandwidths. Taken directly from the speech signal we extract the *intensity* contour in dB. Referred to as *other* features we calculate the Harmonics-to-Noise Ratio (HNR), the correlation between pitch and intensity, the Zero-Crossing-Rate, and the relation of pitched and non-pitched speech segments as individual features.

The statistic unit derives means, moments of first to fourth order, extrema and ranges from the respective contours in the first place. Special statistics, e.g. a linear regression analysis or a Discrete Cosine Transformation (DCT), are then applied to certain descriptors such as pitch, loudness and intensity. Applying the DCT to these contours directly, we model their spectral composition in terms of fast or slow moving contour constituents. Finally, we append delta coefficients of first and second order and calculate their statistics like-

wise. On the basis of a voiced, unvoiced and silence segmentation we calculate ratios of features from these segments both separately and jointly. A more detailed description of the feature setup can be found in Polzehl et al. (2010).

As can be seen from Table 1 our databases consist mostly of very short utterances. We assume that every turn is a short utterance of one prosodic entity. Consequently we calculate our statistics to account for whole utterances. Intuitively, this seems suboptimal for very long utterances. In a comparative study on emotion recognition using phoneme-, word- and sentence-level analysis Vlasenko et al. (2008) concludes, that larger units seem to be beneficial for emotion recognition. He obtains best results when calculating features on sentence-level. Furthermore, other systems also model the course of acoustic contours by dynamic methods, e.g. Vlasenko and Wendemuth (2009), or hybrid methods, e.g. Vlasenko et al. (2007).

All in all, we obtain some 1450 features. Table 5 shows the different audio descriptors groups, the number of features calculated from them and preliminary classification scores. Note that differences in absolute number of features can bias the figures, e.g. if a higher number of features also means a higher amount of relevant information. To determine the amount of relevant information we use a filter-based feature selection method explained below.

4.2. Acoustic Feature Selection and Classification

In order to gain insight into the performance of individual features, we apply 10-fold cross-validation evaluating the entropy-based information-gain-ratio (IGR) (Duda et al., 2000) of individual features. Given a single feature and the observed values it holds, the IGR generally estimates the reduction of uncertainty about a class distribution given the conditional entropy of observations.

Looking for the optimal feature set we incrementally admit a greater number of top-ranked features for classification. In terms of classification we apply Support-Vector-Machines (SVM) as introduced by Vapnik and Cortes (1995). SVMs have been shown to produce excellent results reliably. Moreover, SVMs are proven to yield good results for small data sets as well. The algorithm views data as sets of vectors from two classes in a multi-dimensional space. A separating hyper-plane in the feature space is constructed, which maximizes the margin between data and the hyper-plane. As a result, SVM classification provides a high degree of generalization. We determine the optimal settings for

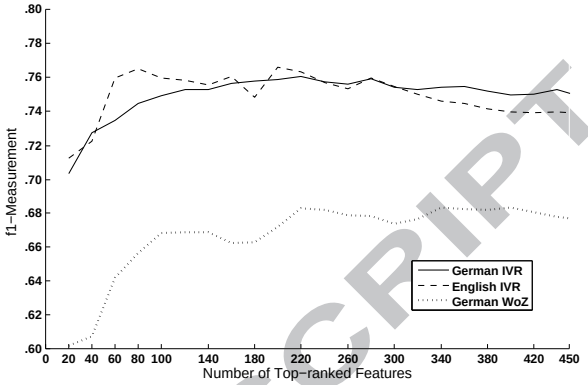


Figure 2: Determination of Optimal Feature Set Size.

the algorithm's complexity parameter by 10-fold cross-validation. SVMs can be extended to non-linear feature spaces by passing the original data points through kernel functions. The choice of the best kernel function can only be done experimentally. We use linear kernel function and expand the kernel function to higher polynomials and radial basis functions (RBF) when fusing the systems. The combination of SVM with an RBF kernel function in turn is very similar to an RBF type of artificial neural network (ANN). We also use ANNs, namely multi-layer perceptrons (MLP), directly for final fusion at the decision level.

In order to compare observations from different distributions we apply $z-normalization$, also known as $z-scores$, $normalscores$ or $standardization$ (Duda et al., 2000) to our features. Eventually, each feature is normalized to zero mean and unit variance before classification. To avoid overestimation we apply 10-fold stratified cross-validation for any classification steps. The optimal feature-set size is 220 for the German databases and 80 for the English database. Table 6 shows the scores obtained.

Comparing the scores for the three databases, we observe good results for the IVR databases. The F-measures of the Anger classes are approximately .7, while the F-measures of Non-Anger classes exceed this figure by .1 absolute. Acoustic modeling of the Anger class of the WoZ corpus seems problematic. The models were not able to capture the emotion-related information needed for classification. As a hypothetic explanation, this could be due to the mapping of the classes $touchy$, $angry$, $reprimanding$ and $emphatic$ into the single cover class of $Anger$. Although the inter-labeler agreement does not indicate human differences in perception between the German databases, the acoustic models might been blurred by the process of sub-

Table 5: Feature Groups and Performance on the Databases.

Feature Group	*Number of Features*	*f1 Performance on German IVR*	*f1 Performance on English IVR*	*f1 Performance on German WoZ*
pitch	240	.68	.73	.63
loudness	171	.68	.71	.67
MFCC	612	.69	.68	.71
spectrals	75	.68	.69	.64
formants	180	.68	.68	.65
intensity	171	.69	.74	.69
other	10	.56	.67	.62

Table 6: Classification Results using Acoustic Information.

Database	*Class*	*Recall*	*Precision*	*F-measure*	*f1-measure*
German IVR	Non-Anger	86.3%	79.0%	.82	.76
	Anger	64.7%	75.8%	.69	
English IVR	Non-Anger	83.6%	80.1%	.82	.77
	Anger	68.8%	74.0%	.71	
German WoZ	Non-Anger	85.1%	77.0%	.81	.68
	Anger	50.0%	63.6%	.56	

summation of classes.

Figure 2 shows the f1 development as the feature space increases. The saw-like shape of the graphs indicates a non-optimal ranking, i.e. some inclusions seem to harm the performances. This is mainly due to heuristic IGR estimation. Regarding the magnitude of the jitter we note that it is only about .02, which after all proves a generally reasonable ranking. The filter seems to predict best for the German IVR database, where the observed jitter is very low. Regarding efficiency, we note that including only 120 features for the German IVR, and 100 features for the German WoZ database results in a loss for f1 of only about .01.

Analyzing the acoustic feature groups in the ranked sets we see that features derived from filtering in the spectral domain, e.g. MFCC and loudness, seem to be most promising for all three databases. They account for more than 50% of all features. However, MFCCs occur more frequently among the top-ranked features when operating on the German IVR database, while loudness features are more frequently among the top ranks when operating on the English IVR database. On the WoZ database, too, loudness is highly ranked more frequently. Pitch features account for approximately 25% of the top sets when trained on the English IVR database, while the number is as small as about 10% when trained on the German top sets.

5. Fusion of Linguistic and Acoustic Features

As previous experiments using early fusion techniques show inferior results, cf. Polzehl et al. (2010), we combine the predictions obtained from the acoustic and the linguistic analysis by decision fusion using SVMs and compare the results to MLP fusion. In the MLP experiments we use back-propagation training in maximal 500 epochs. A validation set size of 10% is used for early training termination. The nodes of the three layers are connected by sigmoid activation functions, the middle layer comprises 4 nodes. We further extent the SVM classifier to non-linear mapping by transforming data using a RBF kernel function. We determine the optimal settings for the algorithm's complexity parameter and the kernel width by 10-fold cross-validation on the train set in a grid-search manner. The final feature space consists of eight features, four from each acoustic and linguistic system. Both systems contribute a predicted class. The acoustic system further generates scores by logistic regression analysis for each prediction. The linguistic system contributes the estimated SRI scores directly. On the basis of these scores we compute normalized confidence estimates for both classifiers by computing the rank for a score in its population and renormalizing this to a range of [0,1]. In our case, the normalized rank corresponds to the discrete probability distribution. In other words, we estimating the amount of confidence in a prediction to belong to a class by con-

sidering likelihoods of predictions of higher and lower values for that class. Table 7 shows the final classification scores obtained by 10-fold cross-validation.

As expected, the fusion of both types of information at the decision level generates slight improvements for all corpora. Looking at the magnitude of the difference, little improvement could be expected due to inferior performance of the linguistic classification. After all, the final scores resemble the scores obtained from the acoustic classification. We note that for the approximate 33/66 split of class distributions for the WoZ corpus, constant classification into the majority class would result in approximately .40 f1, 66% accuracy. For the IVR databases constant majority class voting would result in .38 f1, 60% accuracy. Similarly, to the results of the acoustic classification, the low F-measures for the Anger class models for the WoZ database turns out to be problematic. Another expected result is that non-linear classification yields better results than linear classification. Furthermore, in our experiments SVM-RBF slightly outperforms MLP fusion.

6. Summary and Results

The present study has investigated the exploitation of both linguistic and acoustic feature modeling for anger classification. In terms of acoustic modeling we generate statistics from acoustic speech features, e.g. pitch, loudness and spectral characteristics. Ranking our features, we see that loudness and MFCC seem most promising for all three databases. For the English database pitch features are also important. In terms of linguistic modeling we apply probabilistic and entropy-based models for words and phrases, e.g. BOW, TF.IDF and SRI. SRI clearly outperforms vector space models. Modeling phrases improves the scores slightly.

After classification of both acoustic and linguistic information at separated levels, we fuse the information at the decision level by adding confidences. We compare the scores obtained for three different databases. Two databases are in the IVR customer care domain, and another database is in a Wizard-of-Oz domain. All corpora represent realistic speaking conditions.

We observe promising results for the IVR databases while the WoZ database shows lower scores overall. In order to provide comparability between the results, we evaluate classification success using the f1-measurement in addition to overall accuracy figures. As a result, acoustic modeling clearly outperforms linguistic modeling. Fusion improves overall scores slightly. With a baseline of approximately 60% accuracy and .40 f1 by constant majority class voting we obtain an accuracy of 75.3% with respective .70 f1 for the WoZ database. For the IVR databases we obtain approximately 78.9% and 78.2% accuracy for German and English respectively, while f1-measurements result in .78. Baselines figures of these databases are 60% accuracy with .38 f1.

7. Discussion

For the anger classification tasks in actually deployed IVR domains, which we have presented here, acoustic modeling seems consistently more promising than linguistic modeling. However, there are a number of factors, which could have influenced our results and whose importance would need to be re-assessed for other databases. Although we have produced results on three different databases, all the results explained here must also be interpreted as corpus-dependent. Therefore, the following Sections discuss key factors of influence in terms of similarities and differences between the databases.

7.1. Signal Quality

While the WoZ database has been recorded in wide band quality under controlled conditions, the IVR databases are of narrow-band quality including noise. However, signal quality does not seem to be the predominant factor in our anger recognition experiments since the results for the assumed higher-quality WoZ database do not yield better classification scores. Furthermore, the IVR speech quality can also be subdivided into sub-classes, since 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 usually transmit less compressed speech it could be assumed that there is more information retained in it. Eventually, the impact of the difference in the total amount of information between wide-band and narrow-band speech as well as the differences caused by speech transmission remain to be addressed in further experiments in the future.

7.2. Speech Duration

Another factor in the database design is the average turn length, which also correlates to the number of words in a turn. On the one hand, a longer turn offers more data for both linguistic and acoustic analysis. On the other hand, emotional expression can be very short

Table 7: Classification Performance after Fusion.

Database	*Classifier*	*F-measure Anger*	*F-measure Non-Anger*	*Accuracy*	*f1-measure*
German IVR	SVM-linear	.70	.82	77.7%	.76
	SVM-RBF	.73	.83	79.0%	.78
	MLP	.72	.82	78.2%	.77
English IVR	SVM-linear	.70	.82	77.5%	.76
	SVM-RBF	.73	.82	78.2%	.78
	MLP	.73	.81	78.1%	.77
German WoZ	SVM-linear	.55	.81	73.3%	.68
	SVM-RBF	.57	.83	75.3%	.70
	MLP	.57	.83	75.1%	.70

naturally. Longer passages might include diverse emotional user state changes. Segmentation and labeling issues need to be addressed in future work in this respect. In terms of modeling, statistics are presumably more reliable on the basis of longer speech utterances, as the acoustic system is based on turn-wise statistics such as mean and regression analysis. Also Vlasenko et al. (2008) concludes, that larger units seem to be beneficial for emotion recognition. Also, linguistic phrase modeling can be expected to improve. However, the differences in speech duration in the selected databases do not show significant influence. The German databases are comparable in terms of average words per turn, though the standard deviation is higher for the IVR database. The English database is of lower average. In the present work, differences in overall performance do not appear to be correlated to this condition. As speech duration is mostly limited by database-dependent turn lengths or phrase lengths, most common units of analysis are naturally turns or words. However, also modeling on subword level, e.g. phonemes (Vlasenko and Wendemuth, 2009), phoneme-classes (Bitouk et al., 2010), has been applied. Eventually, the present diversity in the literature suggests that the optimal analysis unit strongly depends on the data.

7.3. Scenario

The captured data originate in all cases from human-computer interaction (HCI) and are not mixed with human-human interaction data. In all three scenarios the speakers believed to be talking to a machine. The basic comparability of the corpora for a HCI anger task can thus be seen as given.

7.4. Language

It should be noted that the findings seem to generalize well across the two languages, i.e. English and German. Both IVR databases yield comparably good results for both the acoustic and linguistic task. Further details, including a more thorough analysis of important features, mono- and multilingual performance evaluation, and a literature survey, can be found in Polzehl et al. (2009a). Comparing mono- and multilingual emotion recognition experiments on English, Slovenian, Spanish, and French acted speech recordings also Hozjan and Kacic (2003) conclude that multilingual emotion recognition can be applied successfully using prosodic features. Working on Chinese, English, Russian, Korean and Japanese acted speech Wang et al. (2009) confirms this finding. He concludes, that although prosodic features show variation with respect to different emotions, the variation with respect to different languages indicates commonness.

7.5. Domain and Vocabulary Size

The domain of the data presumably seems to be of importance. Although the perplexity and the vocabulary size of the English IVR are significantly smaller than for the German IVR, the resulting scores are comparable. Downsampling the databases to match the smaller vocabulary size conditions did not show a significant influence on anger recognition in terms of linguistic modeling. On the one hand, if the vocabulary is limited, speakers may need to resort more to acoustic means in order to express emotions. This could explain the comparatively low performance of the linguistic models for the English IVR domain. On the other hand, the acoustic models are expected to be more distinct in that case. Still, the acoustic model performance is comparable to the models trained on the German IVR database, which has a much larger vocabulary. Future experiments therefore need to focus on acoustic emotion detection when subsampling according to linguistic word paradigm criteria. We further contrasted data

from a hotline domain with data from a navigational instruction domain. Linguistic and acoustic performance proved poor for the navigational domain generally. Further experiments separating the WoZ corpus into domain specific characteristics need to be conducted. Directing commands to a robot may well, for example, cause different speech behavior than inquiring for assistance.

7.6. Speakers

The number of speakers may also have influenced the results. The WoZ corpus contained 26 speakers, while the IVR corpora contained more than 400 speakers. The resulting models from the IVR databases thus might be more robust in terms of speaker-independence. All results presented are obtained by speaker-dependent 10-fold cross-validation[4]. In addition, the possibility that callers exceptionally dialed in more than one time cannot be denied. However, the ratio between the high number of speakers in the IVR cross-validation sets and the average dialog length results in a quasi speaker-independent cross-validation folds setup. On the other side, speaker dependency in the setup of the WoZ experiments will most likely lead to a lower performance, when the system encounters new speakers. Also recording children versus adults could potentially have a major effect. It is not hard to imagine that children may adapt quite differently and with more variation when speaking with a toy robot machine than adults dialing into a customer care hotline. Provided that specific emotional expressions are shaped by experience, adults may have had specific associations when initiating the call that may have led to more trained and more categorical reactions during the whole interaction.

7.7. Class Labels and Agreement

The inter-labeler agreement is similar for both German databases with $\kappa = 0.53$ for the German IVR and $\kappa = 0.56$ for the German WoZ corpus. The slightly higher $\kappa = 0.63$ for the English IVR corpus could indicate a clearer separation of the expressive patterns and could thus explain the good performance on the English IVR. The low performance of the WoZ corpus can perhaps be attributed to the diversity of the anger class labels which comprise *touchy*, *angry*, *reprimanding* and *emphatic*. Here, acoustic profiles may be overlapping, interfering or mutually blurring. Experiments using the original class definitions may well lead to more precise results. Also differences in labeler training could be influential. For the presented databases the labelers listened to the speech utterances before voting. An annotation phase based only on the linguistic content (without listen to the dialogs) could help to obtain better performances with linguistic features. In addition, Steidl (2009) concludes, the mere task of choosing exactly one category forces the labelers to fell hard- and clear-cut decisions, independently of the intensity of the emotional state. The very point of cutting varies intra- and inter-personally. Still, the κ values achieved here are typical for emotion databases, which generally amount to "fair" agreement between labelers only. This points to a general difficulty in determining ground truth for an abstract condition such as "emotion", when only acoustic and linguistic evidence is available, which is a challenge for all work on real-life data on this task, be it performed by humans, or machines.

[4]Since we are not competing with other algorithms in this article, our experiments use all data available for cross-validation.

8. Acknowledgments

The authors wish to thank Prof. Dr. Sebastian Möller, Prof. Dr. Wolfgang Minker, Dr. Felix Burkhardt, Dr. Joachim Stegmann, Dr. David Sündermann, Dr. Roberto Pieraccini, and Dr. Jackson Liscombe for their encouragement and support. The authors also wish to thank all colleagues at their respective laboratories for support and insightful discussions.

References

Batliner, A., Fischer, K., Huber, R., Spilker, J., Nöth, E., 2000. Desperately Seeking Emotions: Actors, Wizards, and Human Beings, in: ISCA Workshop on Speech and Emotion.

Bitouk, D., Verma, R., Nenkova, A., 2010. Class-Level Spectral Features for Emotion Recognition. Speech Communication 52, 613–625.

Boersma, P., Weenink, D., 2009. Praat: Doing Phonetics by Computer.

Burkhardt, F., van Ballegooy, M., Huber, R., 2005a. An Emotion-Aware Voice Portal, in: Proceedings of Electronic Speech Signal Processing ESSP.

Burkhardt, F., Polzehl, T., Stegmann, J., Metze, F., Huber, R., 2009. Detecting real life anger, in: Proc. of International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2009), IEEE, Taipei, Taiwan. pp. 4761–4764.

Burkhardt, F., Rolfes, M., Sendlmeier, W., Weiss, B., 2005b. A Database of German Emotional Speech, in: Proc. of the Annual Conference of the International Speech Communication Association (Interspeech 2005), ISCA.

Davies, M., Fleiss, J., 1982. Measuring Agreement for Multinomial Data.

Devillers, L., Vidrascu, L., Lamel, L., 2005. Challenges in Real Life Emotion Annotation and Machine Learning based Detection. Journal of Neural Networks 18, 407–422.

Duda, R.O., Hart, P.E., Stork, D.G., 2000. Pattern Classification. John Wiley & Sons. 2nd edition.

Dumouchel, P., Dehak, N., Attabi, Y., Dehak, R., Boufaden, N., 2009. Cepstral and long-term features for emotion recognition, in: Proc. of the Annual Conference of the International Speech Communication Association (Interspeech 2009).

Enberg, I.S., Hansen, A.V., 1996. Documentation of the Danish Emotional Speech Database. Technical Report. Aalborg University, Denmark.

Fastl, H., Zwicker, E., 2005. Psychoacoustics: Facts and Models. Springer, Berlin. 3rd edition.

Hozjan, V., Kacic, Z., 2003. Context-Independent Multilingual Emotion Recognition from Speech Signals. International Journal of Speech Technology 6, 11–320.

Huang, X., Acero, A., Hon, H.W., 2001. Spoken Language Processing. Prentice Hall.

Lee, C.M., Narayanan, S.S., 2005. Toward Detecting Emotions in Spoken Dialogs. IEEE Transactions on Speech and Audio Processing 13, 293–303.

Lee, F.M., Li, L.H., Huang, R.Y., 2008. Recognizing Low/High Anger in Speech for Call Centers, in: International Conference on Signal Processing, Robotics and Automation, World Scientific and Engineering Academy and Society (WSEAS). pp. 171–176.

Metze, F., Batliner, A., Eyben, F., Polzehl, T., Schuller, B., Steidl, S., 2010. Emotion recognition using imperfect speech recognition, in: Proc. of the Annual Conference of the International Speech Communication Association (Interspeech 2009), IEEE, Makuhari, Japan. pp. 1–6.

Metze, F., Englert, R., Bub, U., Burkhardt, F., Stegmann, J., 2008. Getting Closer: Tailored HumanComputer Speech Dialog. Universal Access in the Information Society .

Metze, F., Polzehl, T., Wagner, M., 2009. Fusion of acoustic and linguistic speech features for emotion detection, in: Proc. of International Conference on Semantic Computing (ICSC 2009), IEEE, Berleley, CA, USA.

Polzehl, T., Schmitt, a., Metze, F., 2009a. Comparing features for acoustic anger classification in german and english ivr systems, in: Proc. of International Workshop of Spoken Dialogue Systems (IWSDS 2009), University of Ulm, Ulm, Germany.

Polzehl, T., Schmitt, A., Metze, F., 2010. Salient features for anger recognition in german and english ivr portals, in: Spoken Dialogue Systems Technology and Design. Springer, Berlin, Germany, pp. 81–110.

Polzehl, T., Sundaram, S., Ketabdar, H., Wagner, M., Metze, F., 2009b. Emotion classification in children's speech using fusion of acoustic and linguistic features, in: Proceedings of the Annual Conference of the International Speech Communication Association (Interspeech 2009), ISCA, Brighton, England. pp. 340–343.

Schmitt, A., Tschaffon, U., Heinroth, T., Minker, W., 2010. Inter-Labeler Agreement for Anger Detection in Interactive Voice Response Systems, in: 6th International Conference on Intelligent Environments (IE'10).

Schuller, B., 2006. Automatische Emotionserkennung aus Sprachlicher und manueller Interaktion. Dissertation. Technische Universität München. München.

Schuller, B., Metze, F., Steidl, S., Batliner, A., Eyben, F., Polzehl, T., 2009a. Late fusion of individual engines for imrpoved recognition of negative emotion in speech - learning vs. democratic vote, in: International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE.

Schuller, B., Rigoll, G., Lang, M., 2004. Speech Emotion Recognition Combining Acoustic Features and Linguistic Information in a Hybrid Support Vector Machine - Belief Network Architecture, in: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).

Schuller, B., Steidl, S., Batliner, A., 2009b. The interspeech 2009 emotion challenge, in: Proc. of the Annual Conference of the International Speech Communication Association (Interspeech 2009).

Shafran, I., Mohri, M., 2005. A Comparison of Classifiers for Detecting Emotion from Speech, in: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).

Shafran, I., Riley, M., Mohri, M., 2003. Voice Signatures. in: Automatic Speech Recognition and Understanding, 2003. ASRU '03. 2003 IEEE Workshop on, pp. 31–36.

Steidl, S., 2009. Automatic Classification of Emotion-Related User States in Spontaneous Children's Speech. Ph.D. thesis.

Steidl, S., Levit, M., Batliner, A., Nöth, E., Niemann, H., 2005. "Of All Things the Measure is Man" - Classification of Emotions and Inter-Labeler Consistency, in: IEEE (Ed.), International Conference on Acoustics, Speech, and Signal Processing (ICASSP), pp. 317–320.

Vapnik, V., Cortes, C., 1995. Support Vector Networks. Machine Learning 20, 273–297.

Vidrascu, L., Devillers, L., 2007. Five Emotion classes Detection in Real-World Call centre Data: the Use of Various Types of Paralinguistic Features, in: Paraling.

Vlasenko, B., Schuller, B., Mengistu, K., Rigoll, G., Wendemuth, A., 2008. Balancing Spoken Content Adaptation and Unit Length in the Recognition of Emotion and Interest, in: Proc. 9th INTERSPEECH 2008 incorp. 12th Australasian Int. Conf. on Speech Science and Technology (SST 2008), pp. 805–808.

Vlasenko, B., Schuller, B., Wendemuth, A., Rigoll, G., 2007. Combining Frame and Turn-Level Information for Robust Recognition of Emotions within Speech, in: Proc. of the Annual Conference of the International Speech Communication Association (Interspeech 2007), pp. 2225–2228.

Vlasenko, B., Wendemuth, A., 2009. Processing affected speech within human machine interaction, in: Proc. of the Annual Conference of the International Speech Communication Association (Interspeech 2009).

Wang, Y., Li, B., Meng, Q., Li, P., 2009. Emotional Feature Analysis and Recognition in Multilingual Speech Signal, in: Electronic Measurement & Instruments (ICEMI), Beijing.

Yacoub, S., Simske, S., Lin, X., Burns, J., 2003. Recognition of Emotions in Interactive Voice Response Systems, in: Eurospeech, Geneva, pp. 1–4.