Fusion of Acoustic and Linguistic Speech Features for Emotion Detection

Florian Metze[1], Tim Polzehl[2] and Michael Wagner[3]
[1]*SCS/ LTI, Carnegie Mellon University, Pittsburgh, USA*
[2]*Deutsche Telekom Laboratories, Berlin, Germany*
[3]*National Centre for Biometric Studies, University of Canberra, Australia*
fmetze@cs.cmu.edu, tim.polzehl@telekom.de, michael.wagner@canberra.edu.au

Abstract

This paper describes a system that deploys acoustic and linguistic information from speech in order to decide whether the utterance contains negative or non-negative meaning. An earlier version of this system was submitted to the Interspeech-2009 Emotion Challenge evaluation. The speech data consist of short utterances of the children's speech, and the proposed system is designed to detect anger in each given chunk. Various frame-based cepstral, prosodic and acoustic features are extracted automatically and classified by means of a support vector machine. An automatic speech recognizer transcribes the utterances and yields a separate classification, based on the degree of emotional salience of the words. The emotionally salient words are computed on word hypotheses, so that un-transcribed training data is sufficient. Late fusion is applied to make a final decision on anger vs. non-anger of the utterance.

1. Introduction

Speech utterances not only contain the literal meaning of the words spoken, but convey a wealth of additional information to the listener [1]. The language spoken, the speaker's dialect, accent and sociolect as well as the specific choice of grammatical construct, words chosen over synonyms, emphasis, articulation etc. all convey a rich context for the literal message to the native listener of the speaker's language. The human listener is particularly finely tuned to the detection of a range of emotions in the speaker's utterance. A native listener is able to detect certain emotions by recognizing salient words, which are associated with those emotions. However, a speaker's emotions are also accessible to some extent to non-native listeners who are able to utilize acoustic cues to distinguish emotions in speakers whom they otherwise do not understand.

This dual association of human emotion with the linguistic content of an utterance and with some of its acoustic characteristics has motivated us to explore a combination of the acoustic and the linguistic features of utterances, in order to detect an angry disposition of the speaker. The acoustic information comprises a wide range of spectral, prosodic and glottal-source features extracted from the speech signal automatically, while the linguistic information is based on the words of the utterance either as detected by automatic speech recognition or as transcribed by human listeners, and their salience for anger as given by a labeled data corpus.

The data corpus contains approximately 18,000 utterances of children who converse with an Aibo robot dog in German. The "dog" often behaves erratically and flouts the children's commands, causing the children to become annoyed and angry with it, which is reflected in many of the utterances. Utterances were tagged with emotions as they were perceived by human labelers.

This paper builds on an earlier paper [16], which was part of the Interspeech 2009 Emotion Recognition Challenge [2], which discusses the database and related work in more detail. The acoustic subsystem of the earlier paper is used as a baseline, which is combined with several automatic speech recognizers using different fusion methods. The dispositions we are trying to distinguish are therefore a combination of negative emotions, particularly anger, and all non-negative dispositions.

This paper is structured as follows: Section 2 describes the overall system design, Section 3 gives a short introduction to the AIBO database, Section 4 describes the features used for classification, separated into linguistic (text-based) features, and acoustic and prosodic features. While the classification based on text is covered integrated in Section 4, Section 5 is dedicated to the classification of the acoustic and prosodic features. The combination of both types of

features is treated in Section 6, while Section 7 summarizes or findings.

2. Overall system design

We have developed an end-to-end system using mostly publicly available sources and toolkits. The overall system design comprises one subsystem, which evaluates a large number of acoustic features it extracts from a given chunk, and a second subsystem, which applies automatic speech recognition and then evaluates the spoken words it recognizes in the chunk.

The acoustic subsystem extracts a large number of acoustic features from the chunk automatically. These comprise frame-based intensity, fundamental-frequency and cepstral features, chunk-based statistical measures of those features, and features based on the shape of the glottal excitation waveform of a central vowel. For testing, an "anger" score is calculated for a chunk by means of a support vector machine with a radial-basis-function kernel. More precisely, the "anger" score applies to a range of negative emotions, which are labeled with the tag "NEG" in the Aibo corpus, while a "non-anger" score applies to diverse non-negative emotions, which is labeled with the tag "IDL". Feature selection and reduction is applied before performing classification and evaluation.

The linguistic subsystem performs a word recognition task on each chunk. The anger-non-anger decision is based on the a-posteriori anger score of the words learnt during training by applying the concept of emotional salience of the recognized words with respect to the anger or non-anger classes, respectively. Each subsystem yields its own decision and confidences. An overall fusion algorithm combines these into the final decision.

The optimality criteria that our systems are developed on are also taken from the EmoChallenge. The system is built to optimize first the un-weighted average recall of emotion classes, and second the weighted average recall (or accuracy) in the classification. For more discussion, see [2].

3. Database

The AIBO database [4] corpus contains short utterances, or "chunks", from a speech database of children who converse with an Aibo robot dog in German. There are 9957 chunks for training and 8257 for testing. Each chunk in the database is tagged with a perceived emotion by a set of human labelers. For development, we subdivided the training database into 10 partitions (round-robin), such that evaluation data are not seen by the system during the training phase.

It consists of spontaneous German speech that is emotionally colored. While 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 reaction from the children. The data was collected at two separate schools and amounts to 9.2h of 16bit/ 16kHz speech recordings in total. 5 labelers annotated the utterances with respect to 10 target classes and neutral, which were eventually mapped to a binary division between negative ("NEG") and non-negative ("IDL") utterances. Turns were split into chunks by syntactic-prosodic criteria. Roughly two thirds of chunks were labeled as non-negative, so all results presented here or in [2,16] represent only a small, but significant, improvement over chance level. Note that the train and test set were completely independent since the speakers were from two different schools.

4. Feature extraction

We considered two distinct feature sources for our experiments on this database. The linguistic source is drawn from the actual words the children use to direct the robot. Using the concept of emotional salience [3], it is possible to assign an emotion to an utterance, given a word hypothesis or transcription. Prosodic and acoustic information provide another useful source for characterizing speech utterances. We extracted measurements of intensity and duration, perceptual loudness and fundamental frequency (F_0), formants, cepstra, and voice-source characteristics obtained by inverse filtering, and describe our respective classifier in Section 5. In Section 6, both information sources are fused using their associated confidence scores.

4.1 Linguistic features and classification

The EmoChallenge training database [2,4] provides transcriptions of the training chunks, and previous work [3, 5] show that transcriptions can be used as features for classification of emotional content. In order to generate transcriptions also for the test data, we developed an ASR (automatic speech recognition) system for the EmoChallenge speech. Results will be reported for classification using transcripts (on the training set), ASR hypotheses (on the training and test sets) with varying quality, using ASR confidences, and in combination with original acoustic features.

Our baseline ASR system was trained on about 14h of close-talking, clean 16kHz "background" speech, recorded from adults reading German newspaper texts,

using the Janus [6] toolkit and the Ibis [7] decoder. The acoustic model uses 2000 context-dependent, speaker-independent acoustic models. These were trained using Maximum Likelihood and employ 32 Gaussians with diagonal covariance matrices each in a 42-dimensional MFCC-based feature space after LDA, also using VTLN and speaker-based CMN/ CVN. The baseline language model was also trained using tri-grams on German Broadcast News type text data and transcripts, using a 60k vocabulary.

To adapt this system to the EmoChallenge, we reduced the vocabulary of the original system to 5k words, 4.5k of them unique. This includes 300 new domain-specific words appearing at least two times, including those marked with a "*" (i.e. non-standard speech), as long as they appeared to be emotionally salient, for which pronunciations were generated manually.

We merged the ML update statistics on the "background" database with matching statistics collected on the EmoChallenge data, using fixed weights, to derive MAP adapted acoustic models. For development on the training data, we followed the "leave-one-out" (LOO) principle. We used 10-fold cross-validation and computed "pseudo-speaker-specific" models, leaving this particular pseudo-speaker out in the model update used for testing this speaker, to match conditions on unseen test data as much as possible. The language model (LM) was also adapted to the target domain using a context independent, LOO-aware interpolation [8] of 3-gram background and in-domain LMs for development. Averaged perplexity on the training data is 55.

Table 1. *Configurations of the acoustic model and performance measures in %. WA=word accuracy, DEL=deletions, INS=insertions, UAR=unweighted average recall, WAR=weighted average recall.*

System	WA	DEL	INS	UAR	WAR
V0	0.0			71.2	70.5
V4	82.9	5.0	3.1	70.2	69.9
V5	66.3	1.9	16.9	69.1	70.1
V6	67.9	7.1	7.2	67.6	67.0
V7	73.5	2.0	11.5	70.0	70.1

During tests, the baseline acoustic model was adapted to the test speaker incrementally using unsupervised constrained MLLR in the feature space, and VTLN. We call this system "V4". For tests on the unseen evaluation test data, we loaded an acoustic and language model trained on the full training set, using parameters optimized under the LOO paradigm. Speaker adaptation was performed using automatically determined speaker clusters, which in tests on the training data was found to give virtually the same word accuracy (WA), as when using the known speakers.

As references on the test data are not available at the moment, we scored four different configurations of the system on the development data (using LOO). The results are shown in the left three columns of Table 1. Different configurations of the baseline acoustic model were tested: V5 and V6 do not use unsupervised incremental adaptation to speaker, as the baseline V4 does. The language model weights were varied for each of V5, V6, and V7 to change deletions and insertions. "V0" denotes use of the references. Word accuracy numbers should be regarded as being indicative only, as data normalization was not performed with extraordinary care.

To classify a chunk, we used the emotional salience as proposed by [3], computed either on references or hypotheses. The right two columns in Table 1 show the baseline performance of the different systems. We observe that high word accuracy alone is no guarantee for good performance, as V7 performs nearly as well as V4, virtue of a low deletion rate. V5 also performs significantly better than V6, despite the nearly equivalent word accuracy.

For developing on the training texts for the salience model (transcripts and hypotheses), we employed a 10-fold cross validation strategy, and applied our final model with parameters computed on all the training data to the test data. Each configuration was tested on a model trained using that same configuration, ie. V4 was tested on a model trained using hypotheses generated by V4. We observe a loss of approximately 0.5% UAR/ WAR absolute when testing speech recognition output (V4-V7) using a model trained on manual transcriptions (V0), instead of its own transcripts. We assume this is because V4-V7 share the same vocabulary, while V0 uses a larger vocabulary, but did not investigate this effect further at this time. Pauses and noises were also removed from recognizer output before building models for emotional salience.

In our next experiment, we extended the emotional salience model to include bi- and tri-grams as additional "words", in order to investigate longer-term dependencies in the structure of speech. We also included word posteriors from the recognizer output in the computation, in an effort to mitigate the influence of recognition errors. To achieve this, we modified Equation (6) in [3] to read

$$w_{mk} := \gamma_n \; i(v_n, e_k)$$

during training, using the "gamma" word confidence measure from [9]. The resulting system is called "V19". Table 2 shows the recalls achieved by the model trained and tested on transcriptions (V0), speech recognition output (V4), and speech recognition output with word confidences attached (V38). Note also that we did not need to set frequency cut-offs or balance the training data during training. In the following experiments, V38 will therefore rely on tri-grams, while V0 and V4 will be based on bi-grams.

Table 2. *UAR (left, in %) and WAR (right) for various system configurations and 1-gram, 2-gram, 3-gram emotional salience features.*

	1-gram		**2-gram**		**3-gram**	
V0	71.2	70.5	71.7	70.5	71.6	70.4
V4	70.2	69.9	70.7	70.3	70.6	70.1
V38	70.3	69.3	70.8	69.7	71.2	70.3

We observe that the model including confidences ("V38") can make use of 3-grams, while the other two systems perform best for bi-grams, and there is a 0.2% to 0.5% performance gap to the model trained on references only. In practice, this loss could most likely be compensated for by using more data. For the subsequent fusion experiments, we took the highest salience of any word in the hypothesis as an un-normalized confidence score, as this feature gave best results across a number of tests.

Table 3. *Most salient words for the uni- (left) and tri-gram (right) cases, with a minimum count of 10. "Neid" (meaning "grudge") seems to be a frequent mis-recognition of "nein" ("no"), when said in anger.*

sal	**word**	**emo**	**sal**	**word**	**emo**
.41	super	IDL	.41	gut Aibo	IDL
	…			…	
.52	stopp	NEG	.85	Aibo nicht	NEG
.55	halt	NEG	.86	links gehen Aibo	NEG
.56	Aibolein	NEG	.90	stoppen	NEG
.58	pfui	NEG	.90	Aibo links	NEG
.60	Neid	NEG	.92	Aibo rechts	NEG
.71	hoch	NEG	.93	links Aibo links	NEG
.87	stoppen	NEG	.99	Aibo lauf geradeaus	NEG

Table 3 shows the most emotionally salient words in our experiments for the V38 system. It is clear that the linguistic features can contribute a lot to the detection of anger, as negative n-grams appear with a high salience ("super" and "gut Aibo" are the most salient non-negative expressions).

On the EmoChallenge test set the linguistic system improved from 62.5%/ 58.6% UAR/WAR to 64.8%/ 60.6%, i.e. by about 2% absolute. While we could not obtain further individual scorings, we believe the largest part of this gain is due to the use of tri-grams in the salience model, in combination with confidence scores, less to the use of a speaker adaptive speech recognition system using unsupervised speaker clusters.

4.2 Acoustic and prosodic features

Regarding the group of perceptually motivated acoustic measurements we extracted pitch and perceptive loudness as defined by [10]. For pitch detection we applied Boersma's PDA algorithm [11], using a low voicing threshold to avoid the loss of too many perceptually voiced segments. We also used a high jumping cost for octave confusions and converted the pitch into the semitone domain using the chunks' mean pitch as reference value. Remaining octave confusions between sub-segments of a chunk had to be corrected by a rule-based path finding algorithm. We smoothed the resulting contour using weighted linear regression and interpolated it using piecewise cubic interpolation.

We drew features from the contour by applying statistics like mean, maximum, standard deviation, skewness and kurtosis. We also applied a discrete cosine transformation (DCT) to the pitch directly, in order to obtain a spectrum of pitch movement.

To obtain a perceptive measurement of signal power we calculated the perceptive loudness [10]. This measurement operates on a Bark filtered version of the speech signal and finally gives an estimate of the perceived loudness in sone units. The resulting loudness values are then given to the same statistics as above. Signal intensity values are also subjected to the same statistics as explained above. In addition, we included a measurement of correlation between pitch and intensity as an independent feature.

Another often used acoustic characteristic of emotional speech is the Mel frequency cepstral coefficients (MFCC). Although they were optimized to perform in speech recognition tasks they often give excellent performance in emotion detection tasks as well. We therefore calculated the average, standard deviation and minima and maxima for 15 MFCCs.

Also derived from spectral analysis are the features of position and bandwidth of the formants. Due to the fact that the database contains children's speech and given the recording frequency of 16kHz we looked for

6 formants to be determined. For each formant contour we defined the same features as for MFCCs.

We included three further contours: the spectral flux, the spectral centroid and the spectral roll off point, the latter of which we set to 95% spectral slice energy. In order to align to human perception we weighted the power spectrum with an dB(A) perception curve before calculating statistics.

In order to capture voice quality we included spectral characteristics of the glottal source, which were obtained by inverse filtering of a prominent pitch period and taking a pitch-synchronous discrete Fourier transform [12]. To get this period we chose the maximum-energy voiced frame in a chunk, where voicing is defined by a harmonics-to-noise ratio (HNR) value of greater than 0.45 within the expected F_0 range of 150-600Hz. A single fundamental period is then extracted by first finding the start sample of the target fundamental period and then the precise duration of that period by determining the maximum autocorrelation of a sequence of samples surrounding the expected start of the next period for lags between $0.95\tau_0$ and $1.05\tau_0$. While the beginning of this period does not correspond to the time of glottal opening, we found that the maximum signal value in the frame provides the most reliable point for the determination of the precise duration of the glottal period.

In order to obtain the glottal excitation waveform, the extracted signal values are subjected to linear prediction analysis and inverse filtering [13]. Throughout the analysis, the extracted signal period is treated as a single period of a periodical signal, such that difference values and autocorrelation functions are evaluated on indices modulo the length of the period. The signal is preemphasized with α=15/16 and a predictor of order 17 is determined by means of the autocorrelation method in order to model the expected number poles for the children's speech. The signal is then inverse-filtered with the resulting LP filter, yielding the required approximation of the glottal excitation waveform. The final step of the determination of the glottal excitation function is a normalization of the phase of the function. The shape of the glottal excitation function is then described by the magnitudes $|X(k)|$ of its discrete Fourier transform (DFT) for $0 \leq k \leq 15$.

After calculating the HNR contour from the autocorrelation lag domain we defined its mean, maximum and standard deviation to be individual features. Drawn directly from the time signal we analyzed the zero-crossing-rate (ZCR) and the offset of the overall elongation.

As some features tend to only give meaningful values when they are applied to special voice characteristics we decided to group each chunk into voiced, unvoiced and silenced regions. We applied a modified version of Rabiner & Sambur's algorithm for isolated-speech detection [14]. Combining this algorithm with our pitch detection we produced a voiced/ unvoiced/ silence grid for each chunk. Considering the problem of relative distance to the microphone that was used during recordings we set up a number of relative features that account for the ratio of features from voiced and unvoiced speech segments. We thus calculated a mean relative perceptive loudness and a mean relative perceptive intensity measurement for all chunks.

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

All in all, we obtained some 1500 features, which partly consist of frequently used features but also introduce new experimentally designed features into the analysis. All features were calculated on a 10ms frame shift rate. Table 4 shows the different feature information sources and the number of features calculated from them.

Table 4. *Information sources, number of features calculated, and (unweighted) average recall.*

Feature Source	**Number of Features**	**Average Recall**
ZCR, elongation, duration, correlation	10	61.5%
Intensity	171	68.9%
MFCC	576	71.1%
Loudness	171	67.6%
Formants	216	65.4%
Spectrum	135	63.6%
Pitch	236	62.6%
Linguistic features	11	49.9%
Inverse filtering	33	64.3%

5. Classification of acoustic features

5.1 Data preparation

All our baseline classification performance was estimated by averaging the results of 10-fold cross validation (LOO). Defining a training set we first split the given set randomly into 10 mutually exclusive parts. In the present case, since the number of IDL utterances were approximately twice the number of NEG utterances, we first equalized the number of samples in each class. To equalize, the IDL samples in each fold were randomly split into two equal sub-parts. The NEG samples in that fold were then

combined with each of the two sub-parts. The average result from the two sub-parts was taken to be the performance estimate for the fold. This procedure aims to more clearly determine the effectiveness of the features and classifiers used in this work. Since no artificial samples were synthesized, we believe this procedure leads to a very conservative and unbiased performance estimate.

5.2 Pretest and classifier determination

The acoustic features described in Section 4 are classified using a Support Vector Machine (SVM). SVMs view data as two sets of vectors in a multi-dimensional space, and construct a separating hyperplane in that space. We initially used an SVM with a linear kernel function for the experiments. However, before applying the features to the SVM, the dimensionality of feature vectors were reduced by applying different dimensionality reduction techniques which are described in the following section.

5.3 Feature selection

To get a first insight into the performance of our features we evaluated them separately in accordance to the groups presented in Table 4. MFCCs performed best in our experiments. Measurements of power such as intensity and perceptive loudness were also performing reasonably. Note that this list gives only a very broad picture of performance since it divides into conceptual feature groups rather than providing single-feature performance assessment. Also the number of features can bias the performance comparison between the groups. Table 4 also presents the number of extracted features along with their average recall, i.e. the number of chunks of a class retrieved divided by the number of chunks of that class in the database. The target measurement presented is the average recall, i.e. averaged over the Anger and non-Anger class.

In order to determine the most promising features for our task individually, we applied an Information Gain (IG) filter. This entropy-based filter estimates the goodness of a single attribute by evaluating its information contribution (gain) of information with respect to the required mean information that leads to a successful classification. To compensate between attributes that show a large difference in variation, i.e. also show large differences in information gain, we calculated the IG-Ratio (IGR) and ranked our features accordingly. Table 5 shows the top 20 ranked features. Results are similar to the results from conceptual feature grouping, i.e. spectral and power-related features are given highest ranks.

Table 5. *Top 20 rankings of the acoustic features.*

Rank	Feature
1	mfcc_max_0coeff_wholeUtterance
2	mfcc_max_0coeff_voicedSegments
3	intensity_mean_voicedSegments
4	mfcc_mean_0coeff_voicedSegments
5	intensity_max
6	intensity_median_voicedSegments
7	spectralMagnitude_13_from_inverseFiltering
8	mfcc_mean_1coeff_voicedSegments
9	loudness_Delta_max
10	loudness_Delta_median_voicedSegments
11	spect._Delta_range_centroid_unvSegments
12	spectrum_mean_flux_wholeUtterance
13	spectrum_std_flux_unvoicedSegments
14	spectrum_mean_flux_unvoicedSegments
15	spectralMagnitude_6_from_inverseFiltering
16	mfcc_mean_0coeff_wholeUtterance
17	spectrum_max_flux_unvoicedSegments
18	loudness_Delta_DCT_1coeff
19	loudness_DCT_2coeff
20	spectrum_std_flux_unvoicedSegments

After ranking the features we searched for an optimal number of features for inclusion. We determined an optimum at 320 features using cross-validation as explained above. Figure 1 shows the resulting graph of unweighted average recall against numbers of features passed to the classifier.

5.4 Optimal classification

In the final classification process we extended the linear SVM to non-linear classification. We evaluated the use of polynomial kernels of different orders experimentally and applied a RBF kernel. The combination of SVM with an RBF kernel function in turn is very similar to an RBF type of Neural Network. We started a grid search to determine the optimal settings of the SVM and the kernel for the training data. Best scores were obtained with an RBF kernel when applying a widened margin constant for the determination of the hyperplane.

Using acoustic/ prosodic information only this setup resulted in an UAR of 75.3% with corresponding accuracy (WAR) of 74.4% on the training data. Our final predictions on the test data as submitted to the EmoChallenge [2] resulted in an UAR of 65.4% and a WAR of 72.4%.

For all predictions using acoustic features, we took the score for the more likely class as output by the SVM as an (un-normalized) measure of confidence.

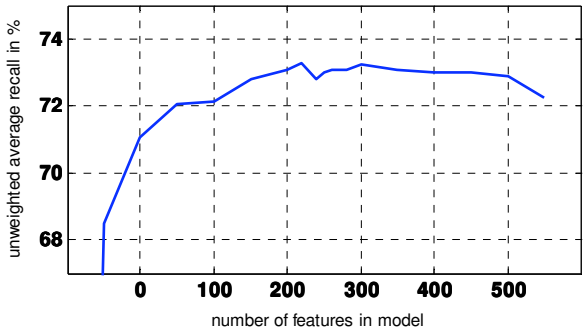

Figure 1. *Effect of the number of features included on average unweighted recall.*

6. System combination

Early experiments, which included linguistic features computed from references as proposed by [3] in the (acoustic) feature selection process and classification, did not improve recognition rates. We therefore developed and optimized separate classifiers on acoustic/ prosodic and linguistic/ textual features and employed a late-fusion strategy. To arrive at a joint decision, we normalized the confidence scores for both classifiers by computing the rank for a confidence score in its population and re-normalizing this to the [0,1] range. We then selected the output with higher normalized confidence to be the output of the combined system, after an additional, empirically determined constant weighting factor was applied to the confidence scores, to compensate for the different baseline performance of the two classifiers. Overall, our confidence scores however are not very reliable, as their distributions generally have only weak positive normalized cross entropy (NCE) [15], even after further processing.

Table 6 compares the performance of various system combinations on the development data. Our system using ASR hypotheses and word confidence measures (V38) beats a system relying on manually transcribed data (V0). The differences between systems are even more pronounced, and advantageous to the systems not relying on transcriptions, than in the pure linguistics case (cf. Table 2).

The confusion matrix of our current best system on the test data is shown in Table 7. It shows balanced errors, and results in an unweighted average recall of 68.1%, and an accuracy of 73.3%, while our baseline system without speaker adaptation, tri-gram features, and confidence measures, but using the same acoustic features, reaches 67.6% and 72.7%. We also experimented with a fusion scheme based on Multi-Layer Perceptrons, but this does not consistently beat the simple confidence-based late-fusion strategy in our experience so far.

Table 6. *Performance (in %) of combined acoustic and different linguistic systems on development data.*

	V0	**V4**	**V38**
UAR	76.2	75.8	76.5
WAR	75.9	76.0	76.2

Table 7. *Confusion matrix on the test set for anger class (NEG) and idle class (IDL), from the best system combining acoustic and linguistic information without resorting to word transcripts during training or test.*

	NEG	**IDL**	**Sum**
NEG	1366	1099	2465
IDL	1110	4682	5792

7. Discussion and conclusion

This paper presents a system to detect angry vs. non-angry utterances of children who are engaged in dialog with an Aibo robot dog. The overall system design comprises two subsystems, one which evaluates a large number of acoustic features it extracts from a given chunk, and a second subsystem, which evaluates the spoken words it recognizes in the chunk. Both sub-systems need labeled training data, but no word transcriptions.

Starting from an evaluation system, this work contributes a comparison of emotional salience features computed on references and entirely on hypotheses. We extend this concept using n-grams and confidence measures and, in combination with acoustic/ prosodic meta-data, reach better performance than a system relying on manually transcribed text. Our system can fully automatically improve automatic categorization of user attitude towards a user interface, based on audio information. This semantic annotation enriches analysis of users' utterances, and it is our hope that an integrated approach for "rich transcription" will improve man-machine interaction in the future.

Our acoustic subsystem extracts a large number of acoustic features from the chunk automatically. These basically comprise frame-based intensity, fundamental-frequency and cepstral features, chunk-based statistical measures of those features, and features based on the shape of the glottal excitation waveform of a central

vowel. We applied feature selection due to the Information Gain Ratio criterion. As a result spectral features and power-related features are given highest ranks. After determination of an optimal number of features to be passed to classification we obtained best classification results using a Support-Vector-Machine extended by a Radial-Basis-Function kernel implementation.

The linguistic subsystem performs a word recognition task on each chunk. The anger-non-anger decision is based on the a-posteriori anger score of the words learnt during training by applying the concept of emotional salience. We improved our scores by applying a speaker-adaptive system that estimated CMN/ CVN, VTLN and constrained MLLR incrementally over a whole speaker.

A decision fusion algorithm combines the scores of the two subsystems by evaluating decisions and normalized confidence scores of both systems. The system performs with a weighted average recall of 76.2% and an unweighted average recall of 76.5% on the development data. Applied to the test data we obtain a weighted average recall of 73.3% with a respective unweighted average recall of 68.1%.

Future work will compare the linguistic classifier presented in this work with other concepts discussed in the literature, and further analyze the influence of factors such as word error rate, confidence measures, and mis-match between training and test conditions with respect to possibility of adaptation, vocabulary issues, etc.

Acknowledgements

The authors would like to thank Shiva Sundaram and Hamed Ketabdar for their contributions to an earlier version of this system. They would also like to thank Qin Jin for providing the unsupervised "children" speaker clusters, the University of Canberra for supporting Michael Wagner's study leave at TU Berlin, and the support staff at CMU, T-Labs and TU Berlin for their generous assistance. We are also indebted to the authors of [2] for allowing us to continue working on the AIBO database after the official end of the EmoChallenge.

References

[1] Austin, J.L., *How to do things with words*, Harvard University Press, Cambridge, Mass, 1962.

[2] Schuller, B., Steidl, S. & Batliner, A., "The Interspeech 2009 Emotion Challenge", in Proc. InterSpeech. Brighton, UK: ISCA, Sep. 2009.

[3] Lee, C. M., & Narayanan, S. S., "Toward Detecting Emotions in Spoken Dialogs", IEEE Transactions on Speech and Audio Processing, Vol. 13, No. 2, pp. 293–303, 2005.

[4] Steidl, S., "Automatic Classification of Emotion-Related User States in Spontaneous Children's Speech", Logos-Verlag, 2009.

[5] Schuller, B., Batliner, A., Steidl, S. & Seppi, D.: "Emotion Recognition from Speech: Putting ASR in the Loop", Proc. ICASSP 2009. IEEE. Taipei, Taiwan. IEEE.

[6] Finke, M., Geutner, P., Hild, H., Kemp, T., Ries, K. & Westphal, M., "The Karlsruhe-Verbmobil Speech Recognition Engine", In Proc. ICASSP-97, Munich, Germany. IEEE.

[7] Soltau, H., Metze, F. & Fügen, C.: "A one pass-decoder based on polymorphic linguistic context assignment," in Proc. 2001 Automatic Speech Recognition and Understanding Workshop (ASRU '01), December, 2001.

[8] Stolcke, A., "SRILM -- An Extensible Language Modeling Toolkit", Proc. IC on Spoken Language Processing, vol. 2, pp. 901-904, Denver, 2002.

[9] Kemp, T. & Schaaf, T.: "Estimating confidence using word lattices", In Proc. EUROSPEECH-1997, 827-830. Rhodes, Greece. ISCA.

[10] Zwicker, E., Fastl, H., *Facts and Models*, 2nd ed., Springer Verlag, Berlin, 1999.

[11] Boersma P., "Accurate Short-Term Analysis of the Fundamental Frequency and the Harmonics-To-Noise Ratio of a Sampled Sound", in Proc. of the Institute of Phonetic Sciences, Vol. 17, Amsterdam, 1993.

[12] Markel, J. & Gray, A., *Linear Prediction of Speech Signals*, Springer Verlag, Berlin, 1975.

[13] Wagner M., "Speaker Verification Using the Shape of the Glottal Excitation Function for Vowels", Proc 11th Australasian Int Conf on Speech Science & Technology, pp 233-238, 2006.

[14] Rabiner, L.R. & Schafer, R.W. "Digital Processing of Speech Signals", J. Acoust. Soc. Am. Volume 67, Issue 4, pp. 1406-1407, April 1980.

[15] NIST, "A tutorial introduction to the ideas behind normalized cross-entropy and the information-theoretic idea of entropy", Tech. Rep., 2004, Avail. at http://www.nist.gov/speech/tests/rt/rt2004/fall/docs/NCE.pdf.

[16] Polzehl, T., Sundaram, S., Ketabdar, H., Wagner, M. & Metze, F., "Emotion classification in children's speech using fusion of acoustic and linguistic features," in Proc. InterSpeech. Brighton, UK: ISCA, Sep. 2009.