INTERSPEECH 2009 BRIGHTON

Emotion Classification in Children's Speech Using Fusion of Acoustic and Linguistic Features

Tim Polzehl [1], *Shiva Sundaram* [1], *Hamed Ketabdar* [1], *Michael Wagner* [1,2], *and Florian Metze* [3]

[1] Quality and Usability Lab, Technische Universität Berlin, Germany
[2] National Centre for Biometric Studies, University of Canberra, Australia
[3] interACT, Carnegie Mellon University, Pittsburgh, USA

{tim.polzehl|shiva.sundaram|hamed.ketabdar}@telekom.de,
michael.wagner@canberra.edu.au, fmetze@cs.cmu.edu

Abstract

This paper describes a system to detect angry vs. non-angry utterances of children who are engaged in dialog with an Aibo robot dog. The system was submitted to the Interspeech2009 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. Frame-based cepstral features, prosodic and acoustic features as well as glottal excitation features are extracted automatically, reduced in dimensionality and classified by means of an artificial neural network and a support vector machine. An automatic speech recognizer transcribes the words in an utterance and yields a separate classification based on the degree of emotional salience of the words. Late fusion is applied to make a final decision on anger vs. non-anger of the utterance. Preliminary results show 75.9% unweighted average recall on the training data and 67.6% on the test set.

Index Terms: speech processing, meta-data extraction, emotion recognition, evaluation

1. Introduction

The task of detecting emotions in speech utterances has become an active field in human communication research in the last decade. Its difficulty lies within the design of pattern learning processes capable of interpreting human speech behavior like humans would do. This paper describes a submission to the open-performance sub-challenge of the Interspeech-2009 Emotion Recognition Challenge [1].

We have developed an end-to-end system using 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 evaluates the spoken words it recognizes in the chunk.

The 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. For testing an anger score is calculated for a chunk by means of a support vector machine with a radial-basis-function kernel. Feature selection and reduction techniques are 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. Each subsystem yields its own decision and confidences. A decision fusion algorithm combines the scores of the two subsystems into the final decision.

The provided training data comprise 9957 short utterances, or "chunks", from a speech database of children who converse with an Aibo robot dog in German.

2. Feature Extraction

In general, we considered different feature sources. One, the linguistic source, is drawn from the actual words the children use to direct the robot. Prosodic and acoustic information provide another useful source for characterizing speech utterances. We extracted measurements of intensity and duration, perceptual loudness and fundamental frequency (F0), formants, cepstra, and voice-source characteristics obtained by inverse filtering. Feature statistics like means etc., as well as relations between voiced, unvoiced and silent parts of the chunk were added to the feature vector with the overall result of one fixed-length static feature vector per chunk. Finally to capture temporal behavior, we appended the static vector with the difference (delta) vector.

2.1. Linguistic Features

The Emotion Challenge training database provides transcriptions of the chunks, and previous work [2,3] shows 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 automatic speech recognition (ASR) system for the challenge.

Our baseline ASR system was trained on about 14h of close-talking "background" speech, recorded from adults reading newspaper texts, using the Janus/ Ibis toolkit [4]. The acoustic model uses 2000 context-dependent, speaker-independent acoustic models. These were trained using 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 (LM) was also trained on German Broadcast News type text data.

To adapt this system to the Emotion Challenge, we reduced the vocabulary and language model of the original system to about 5k words and added domain-specific words that appear at least 2 times in the corpus. We then merged the respective Maximum Likelihood (ML) update statistics, using fixed weights, to derive new acoustic models. For development on the training data, we computed speaker-specific models and evaluated them in a LOSO (leave-one-speaker-out) method. The language model was adapted to the

target domain using a context independent interpolation [5] of 3-gram background and in-domain LMs, which were speaker-specific when used on the training data.

To classify a chunk, we used the emotional salience as proposed by [3], computed either on references or hypotheses. The scores of the emotional salience class models were used in 2 different approaches, namely feature and decision fusion. The results of fusion at decision level are given in Section 4. In terms of feature fusion we defined statistics on class-dependent emotional salience word scores similar to the approach in [2], but could not improve the acoustic-online baseline in our combination experiments.

As the test data did not provide speaker labels, we did not use a speaker-adaptive ASR system for testing. We however experimented with a speaker-adaptive system that estimated CMN/ CVN, VTLN and constrained MLLR incrementally over a whole speaker. Results show robust estimates for these parameters. Table 1 shows the respective recalls.

Table 1. *Weighted (W) and unweighted (U) average recalls (AR) achieved on development (dev) and test (eval) data for the baseline (base) and speaker-adaptive (adapt.) systems in percent*

(%)	base dev	base eval	adapt. dev	adapt. eval
UAR	68.8	62.4	71.2	67.6
WAR	67.0	58.8	70.3	72.7

Using our implementation, we achieved an UAR of 71.5% and a WAR of 70.4% on the development data using the transcripts. The performance loss incurred through the use of ASR is very low. The most emotionally salient words were words like “Aibolein”, “pfui”, “stopp”, etc..

As word error rate (WER) was not the primary target, no normalization was performed for scoring. The speaker-independent system however runs at <30% WER on the development data (LOSO method for training), while the speaker-adaptive system runs at <20% WER.

2.2. Acoustic and Prosodic Features

Previous work [12] showed that acoustic or prosodic characteristics show different performance due to different database and task design. We therefore extracted a broad variety of information from the audio.

Regarding the group of perceptually motivated acoustic measurements we extracted perceived loudness in sone [6], intensity in dB and pitch in semitones covering a range of 150 to 600Hz using a method based on [7]. Due to corpus design we adjusted the parameter values so as to mitigate loss of too many perceptually weak voiced segments as unvoiced and too many pitch octave jumps. The correlation between pitch and intensity was included as an independent feature. Contours were smoothed using weighted linear regression and interpolated using piecewise cubic interpolation. Common statistics like the mean, maximum and standard deviation, and higher-order statistics like skewness and kurtosis were calculated. We also applied a discrete cosine transformation (DCT) to the pitch contour, capturing the contour shape over the whole chunk. Discrete Fourier transforms into the spectral and cepstral domains were also calculated.

Frame-based acoustics were captured in form of 15 MFCCs and the frequencies and bandwidths of 6 formants, from which we calculated the average, standard deviation, minimum and maximum for each chunk. We also included contours of the spectral flux, the spectral centroid and the spectral roll-off point, where the power spectrum was weighted with a perception curve before calculating statistics. In order to capture voice quality we included spectral characteristics of the glottal source, obtained by inverse filtering [8] of a prominent pitch period in the chunk and taking a pitch-synchronous discrete Fourier transform (DFT) [9]. Another feature related to voice quality is the harmonics-to-noise ratio (HNR). After calculating the HNR contour from the autocorrelation lag domain [7] we added its mean, maximum and standard deviation to the features. Also added were 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 specific voice characteristics each chunk was partitioned into voiced, unvoiced and silent regions using a modified version of [10]. 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 capture the temporal behavior we appended first- and second-order derivatives to the contours and their statistics alike.

All in all we extracted some 1500 features, some frequently used in the literature and some rather experimental and novel. Table 2 shows the different feature information sources and the number of features calculated from them.

Table 2. *Acoustic information sources, number of features calculated and (unweighted) average recall on training set*

Feature Source	Number of Features	Average Recall
ZCR, Elongation, Duration, Correlation	10	61.47%
Intensity	171	68.86%
MFCC	576	71.10%
Loudness	171	67.63%
Formants	216	65.38%
Spectrum	135	63.65%
Pitch	236	62.62%
Inverse Filtering	33	64.27%

3. Classification

3.1. Data Preparation

All our baseline classification performance was estimated by averaging the results of 10-fold cross validation. 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.

3.2. Pretest and Classifier Determination

The acoustic features described in Section 2 are passed to a statistical classifier, in order to build statistical reference models for different classes. Several types of classifiers have been proposed in literature for dealing with the problem of emotion classification mainly generative models such as Gaussian Mixture Models (GMMs) and discriminative models such as Artificial Neural Networks (ANNs) or Support Vector Machines (SVMs). Generative models learn to cover the feature subspace belonging to a certain class. In contrast, discriminative models learn boundaries between different class feature subspaces in a discriminative way. Discriminative models have shown to be superior in terms of performance for the task of emotion classification. ANNs and SVMs are the most popular discriminative classifiers used in the literature.

We initially used a Multi-Layer Perceptron (MLP) as classifier, with feature vectors presented to the input layer and emotion labels to the output layer during training. During testing, the MLP estimates the posterior probability of each emotion class at the output according to the features presented at input. These posterior score values are then used for the decision.

We continued our experiments with a Support Vector Machine (SVM) as classifier. 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.

According to our experiments, the SVM proved superior. In the following we present our experiments and results on using SVM classifier with features described in Section 3.

3.3. 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 2. 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 2 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.

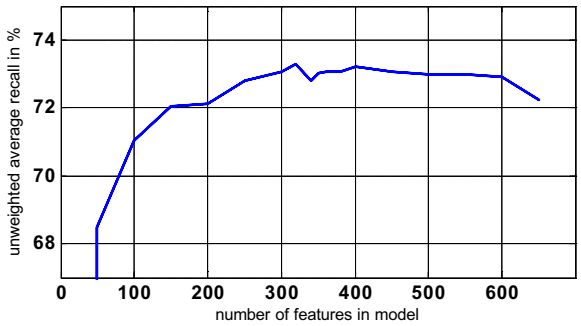
In order to determine the most promising features for our task individually, we applied an Information Gain (IG) filter. This entropy-based filter estimate the goodness of a single attribute by evaluating its information contribution (gain) of information with respect to the required mean information 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 3 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 3. *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	spectrum_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.

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



3.4. 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 unweighted average recall of 75.3% with corresponding accuracy of 74.4% on the training data. Our final predictions on the test data resulted in an unweighted

average recall of 65.39% and a weighted average recall of 72.35%. Our final predictions on the test data using linguistic information only resulted in an unweighted average recall of 67.6% and a weighted average recall of 72.7%.

While incurring a relative small loss in accuracy for both linguistic and prosodic/acoustic systems we loose a relatively high percentage in unweighted average recall when applying our models to the test set.

4. System Combination

Early experiments, which included linguistic features computed from references as proposed by [2] 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 computed normalized 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 constant weighting factor was applied to the confidence scores, to compensate for the different baseline performance of the two classifiers. Overall, confidence scores are not very reliable, as their distributions generally have non-positive normalized cross entropy (NCE) [11], even after further processing.

Finally, after applying late fusion of acoustic/prosodic and linguistic information we obtained an unweighted average recall of 75.9% and a weighted average recall, or accuracy, of 76.0% on the development data. The confusion matrix of the evaluation on the test data is given in Table 4, the corresponding weighted average recall resulted in 72.67 %, the unweighted average recall in 67.55 %.

Table 4. *Confusion matrix on the test set for anger class (NEG) and idle class (IDL)*

	NEG	IDL	Sum
NEG	1352	1113	2465
IDL	1144	4648	5792

5. 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 one subsystem, 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.

The 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 76.0% and an unweighted average recall of 75.9% on the development data. Applied to the test data we obtain a weighted average recall of 72.67% with a respective unweighted average recall in 67.55%.

6. Acknowledgements

The authors would like to thank the organizers for providing this wonderful opportunity and their support staff at T-Labs and TU Berlin for helping out in all situations.

7. References

[1] Schuller, B., Steidl, S. & Batliner, A., “The Interspeech 2009 Emotion Challenge”, Proc. Interspeech, 2009.

[2] Steidl, S., “Automatic Classification of Emotion-Related User States in Spontaneous Children's Speech”, Logos-Verlag, 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] Soltau, H., Metze, F., Fugen, C., Waibel, A., “A one-pass decoder based on polymorphic linguistic context assignment”, IEEE Workshop on Automatic Speech Recognition und Understanding, ASRU, 2001.

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

[6] Zwicker, E., Fastl, H., “Facts and Models”, Springer Verlag, 1999 Second Updated Edition.

[7] 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.

[8] Markel, J. & Gray, A., “Linear Prediction of Speech Signals”, Springer Verlag, Berlin, 1975.

[9] 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.

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

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

[12] Burkhardt, F., Polzehl, T., Stegmann, J., Metze, F., Huber, R., “Detecting Real Live Anger“, Proc. Acoustics, Speech and Signal Processing, ICASSP, Taiwan, 2009.