Emotion Recognition using Imperfect Speech Recognition

Florian Metze[⋆]*, Anton Batliner*[‡]*, Florian Eyben*[†]*,*
Tim Polzehl[*]*, Björn Schuller*[†]*, and Stefan Steidl*[‡]

[⋆]Language Technologies Institute, Carnegie Mellon University; Pittsburgh, PA; USA
[†]Institute for Human-Machine Communication, Technische Universität München; Germany
[*]Quality and Usability Lab, Technische Universität Berlin; Germany
[‡]Lehrstuhl für Mustererkennung, Friedrich-Alexander-Universität, Erlangen-Nürnberg; Germany
fmetze@cs.cmu.edu, {schuller|eyben}@tum.de, tim.polzehl@telekom.de,
{batliner|steidl}@cs.fau.de

Abstract

This paper investigates the use of speech-to-text methods for assigning an emotion class to a given speech utterance. Previous work shows that an emotion extracted from text can convey complementary evidence to the information extracted by classifiers based on spectral, or other non-linguistic features. As speech-to-text usually presents significantly more computational effort, in this study we investigate the degree of speech-to-text accuracy needed for reliable detection of emotions from an automatically generated transcription of an utterance. We evaluate the use of hypotheses in both training and testing, and compare several classification approaches on the same task. Our results show that emotion recognition performance stays roughly constant as long as word accuracy doesn't fall below a reasonable value, making the use of speech-to-text viable for training of emotion classifiers based on linguistics.

Index Terms: speech-to-text, emotion detection, meta-data extraction, rich transcription, children's speech

1. Introduction

Speech utterances not only contain the literal meaning (semantics) 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 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. Linguistically based approaches obviously require transcriptions for the training and the test data. Two models have been used to derive an emotion from a transcription of an utterance, one based on emotional salience [2], and one based on several bag of word models [3]. The quality of the result (i. e. the accuracy of the classification of emotions) depends on the quality of the model, which will ideally be computed on automatic hypotheses rather than manual transcriptions of the training data, and the quality of the hypotheses generated on the test data. In this paper, we will investigate the resulting trade-offs, i. e. how does the speech-to-text error rate affect the quality of emotion recognition.

This paper builds on earlier work [4, 3], which was partly used in the 2-class (negative and neutral) sub-challenge of the INTERSPEECH 2009 Emotion Recognition Challenge on the FAU Aibo Emotion Corpus [5].

2. Database and Task

The FAU Aibo Emotion Corpus [6] comprises recordings of German children's interactions with SONY's pet robot Aibo; the speech data are spontaneous and emotionally coloured.

The children were led to believe that the Aibo was responding to their commands, whereas the robot was actually controlled by a human operator. The wizard caused the Aibo to perform a fixed, predetermined sequence of actions; sometimes the Aibo behaved disobediently, thereby provoking emotional reactions. The data was collected at two different schools, called MONT and OHM, from 51 children (age 10–13, 21 male, 30 female); it contains about 8.9 hours of speech without pauses.

Speech was transmitted with a high quality wireless head set and recorded with a DAT-recorder (16 bit, 48 kHz downsampled to 16 kHz). The recordings were segmented automatically into 'turns' using a pause threshold of 1 s.

Turns were then transcribed manually at the word level, and time-aligned. Five labellers (advanced students of linguistics) listened to the turns in sequential order and annotated each word independently from each other as neutral (default) or as belonging to one of ten other classes. This procedure was iterative and supervised by an expert. Data was labelled on the word level, as many utterances are short commands only, and rather long pauses can occur between words due to Aibo's (or the wizard's) reaction time; the emotional/ emotion-related state of the child can change also within turns.

A label was attributed to a word, if three or more labellers agreed on it. The following labels were used, with frequency of occurence: *joyful* (101), *surprised* (0), *emphatic* (2 528), *helpless* (3), *touchy*, i. e. irritated (225), *angry* (84), *motherese* (1 260), *bored* (11), *reprimanding* (310), *rest*, i. e. non-neutral, but not belonging to the other categories (3), *neutral* (39 169); 4 707 words had no majority vote, they also received a neutral label. All in all, there were 48 401 words.

#	NEG	IDL	$\sum$
train	3 358	6 601	9 959
test	2 465	5 792	8 257
$\sum$	5 823	12 393	18 216

Table 1: *Number of instances for the two classes NEG and IDL, at the chunk level.*

Classification experiments on a subset of the corpus [6, Table 7.22, p. 178] showed that manually defined 'chunks', based on syntactic-prosodic criteria developed in [7], cf. [6, Chap. 5.3.5], give the best compromise between length of unit and homogeneity of units, so that these were used here. In contrast to other recent publications, the whole corpus consisting of 18 216 chunks is used under the very same conditions as for the INTERSPEECH 2009 Emotion Challenge [5]. School OHM (9 959 chunks) was used for training, while school MONT (8 257 chunks) was used for testing.

In this paper, we concentrate on the two-class problem consisting of the main classes NEGative (subsuming *angry*, *touchy*, *reprimanding*, and *emphatic*) and IDLe (consisting of all non-negative states); *emphatic* has to be conceived as a pre-stage of anger because on the valence dimension, it lies between neutral and anger, cf. [6, p. 100].

A heuristic approach similar to [6, Chap. 5.3.8] is used to map the raw labels of the five labellers on the word level onto one label for the whole chunk: if 50 % of these raw labels are NEG, then the whole chunk is labelled as NEG. The whole chunk is also considered to be NEG, if at least one third of all raw labels is NEG, and the remaining raw labels are mostly *neutral*, i. e. at least 90 % of all raw labels are either negative (*angry*, *touchy*, *reprimanding*, *emphatic*) or *neutral*. Still, the class distribution is quite unbalanced (see Table 1).

Our systems were developed for the optimality criterion also used in the INTERSPEECH 2009 Emotion Challenge: the systems are tuned to optimize first the un-weighted average recall of emotion classes (UAR), and second the weighted average recall (WAR or accuracy) of the classification. UAR is defined as the unweighted average of the class-specific recalls achieved by the system, while for the WAR calculation the class-specific recalls are weighted by the prior probabilities of the respective classes. It requires that all classes be recognized with (equally) good recall, which boosts the relative importance of minority classes for imbalanced data.

3. Emotion Recognition

This paper compares two different approaches to the identification of emotions in utterances by linguistic analysis: emotional salience and bag-of-word based classifiers, computed over the full recognition vocabulary.

3.1. Emotional Salience Classifier

Emotional Salience [2] seeks to identify the words which are saliently associated with a specific emotional category, i. e. words which are frequent in one category, but not in others. In [4] we had extended this approach to include word confidence scores, and higher-order n-grams, but in order to allow for fair comparison with the bag-of-words (BOW) approach described below, in this work we only compare systems based on bi-grams for emotional salience, but not confidences.

To provide insight into which words carry meaning, and how this is different between references and hypotheses, Table 2 shows the most salient uni-grams. Most salient words have a negative connotation, and training on references generally produces higher saliences. However, "Neid" (meaning "grudge") seems to be a systematic mis-recognition of "nein" ("no"), if this word is said negatively, as it appears with higher emotional salience. The recognizer sometimes contracts "stehen bleiben" into "stehenbleiben", diminishing frequency and salience of the component words. These two effects (other examples were observed, too) seem to balance each other, so that emotion classification using hypotheses generally works as well as emotion classification using references.

Hypotheses			References		
sal	word	emo	sal	word	emo
	...			...	
0.41	fein	IDL	0.41	fein	IDL
0.41	schon	IDL	0.41	aber	IDL
0.42	stehen	NEG	0.45	bleib	NEG
0.44	nein	NEG		...	
0.44	sollst	NEG	0.49	stehen	
0.45	stehenbleiben	NEG	0.53	halt	NEG
	...		0.56	Aibolein	NEG
0.46	bleiben	NEG	0.56	bleiben	NEG
	...			...	
0.56	Aibolein	NEG	0.63	nein	NEG
0.58	Neid	NEG	0.63	pfui	NEG
	...		0.65	tanzen	NEG
0.71	stoppen	NEG	0.99	stoppen	NEG

Table 2: *Selection (deletions marked by "..." for brevity) of most salient uni-grams, with a minimum count of 10, for models trained on CMU/ T-Labs (see below) hypotheses and references. Higher values for "sal" denote higher salience.*

3.2. Bag of Words Classifier

Our second approach to linguistic analysis is based on the bag of words (BOW) concept [8]: the idea behind this approach is the representation of text in a numeric feature space. Each feature thereby represents the occurrence of a specific word in a sentence. To classify these BOW features, we used Support Vector Machines (SVM) and a discriminatively learned simple Bayesian Network, namely Discriminative Multinomial Naive Bayes (DMNB) [9]. Both are capable of dealing with high dimensional feature spaces. DMNB is a statistical classifier, learning its parameters from the data distribution. SVM is a distance classifier based on the expansion of the feature vector using kernel functions and spanning of a hyperplane in a high dimensional space to optimally separate classes.

3.3. Speech Transcription

Transcriptions were generated using two different systems, in order to be able to analyze effects pertaining to specifics of a particular ASR setup, like normalization, noises, etc. However, no systematic effects could be identified so far, see Table 3.

A speaker-independent ASR system CMU/ T-LABS was first trained on about 14 h of close-talking, clean 16 kHz 'background' speech, recorded from adults reading German newspaper texts, using JRTk/ Ibis [10]. The acoustic model uses 2 000 context-dependent acoustic models. These were trained using Maximum Likelihood (ML) and employ 32 Gaussians

WA (in %)	FAU Hypos	CMU/ T-Labs Hypos
dev	76.3 (2.9|5.6)	82.9 (5.0|3.1)
test	77.4 (2.4|5.7)	81.0 (6.5|1.7)

Table 3: *Word accuracy of adapted CMU/ T-Labs and unadapted FAU recognizers (deletions\|insertions in brackets).*

with diagonal covariance matrices each in a 42-dimensional MFCC-based feature space after Linear Discriminant Analysis (LDA), also using Vocal Tract Length Normalization (VTLN) and speaker-based Cepstral Mean Subtraction/ Cepstral Variance Normalization. A baseline language model was also trained using tri-grams on German Broadcast News type text data and transcripts, using a 60 k vocabulary.

For the INTERSPEECH 2009 Emotion Challenge, we used a vocabulary of about 5 000 words, and 4 500 word types including 300 domain-specific words, appearing at least two times. Out-of-Vocabulary (OOV) rate is around 2 % on the test data (including fragments).

These models were adapted to the childrens' speech using Maximum A-Posteriori (MAP) adaptation. For development on the training data, we used 10-fold Cross-Validation (CV) on the speakers, in order to match conditions on unseen test data as much as possible. The Language Model (LM) was also adapted to the target domain using a matching context independent interpolation of 3-gram background and in-domain LMs for development. Averaged perplexity on the training data is 55.

During tests, the baseline acoustic model was adapted to the test speaker incrementally using feature-space unsupervised constrained Maximum Likelihood Linear Regression (MLLR), and VTLN. For tests on the unseen evaluation test data, we loaded an acoustic and language model trained on the full training set. Speaker adaptation was performed using automatically determined speaker clusters, which was found to give virtually the same error rate as when using known speakers.

A second set of experiments has been conducted using the speech-to-text engine FAU SYSTEM developed at FAU Universität Erlangen-Nürnberg [11]. The first 12 standard MFCC features are being used, while the first MFCC coefficient is replaced by the sum of the energies of the 22 Mel filterbanks, together with first derivatives.

This system is based on semi-continuous hidden Markov models (SC-HMM) modelling polyphones. A polyphone is modelled by its own HMM if it can be observed at least 50 times in the training set. All HMM states share the same set of Gaussian densities; the size of the codebook is 500. However, full covariance matrices are used in contrast to most other systems based on continuous HMMs. We use Baum-Welch re-estimation for training and Viterbi decoding. As language model we use back-off bi-grams.

It is interesting to note that the children in the training partition of the database (OHM) have a higher vocabulary of 703 words and 253 fragments than the MONT students in the test set, which has a vocabulary size of 383 words and 158 fragments. The vocabulary of this ASR system consists of all words (but no word fragments) of both the training and the test set – all in all 813 words. Hence, 158 vocabulary words (types) of the test set are out of OOV, which amounts to a total of 2.1 % OOV tokens.

4. Experiments

To investigate robustness of the approaches, we report results on development and independent test sets. The development

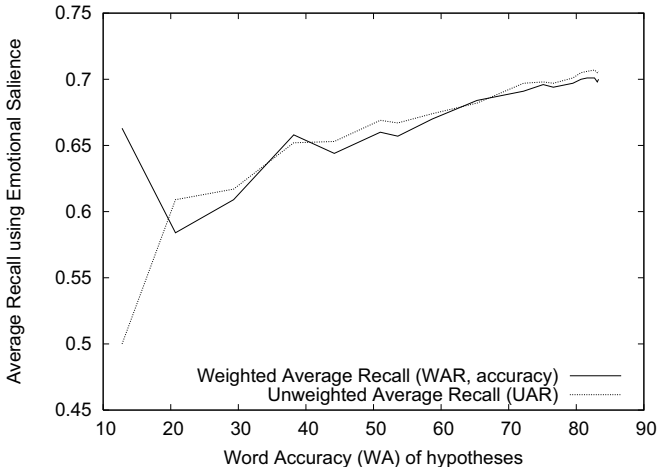

Figure 1: *Recall versus word accuracy using 10-fold CV on the training data and an emotional salience classifier trained on CMU/ T-Labs hypotheses.*

data was used to train the speech recognition, even though parameters were optimized using cross-validation. For brevity of presentation, some results will only be presented on the CMU/ T-Labs hypotheses, which have a lower WER.

4.1. Hypothesis quality dependency on development data

Figure 1 shows weighted and unweighted recall for different word accuracies of training and testing hypotheses, achieved by varying the decoding beam, using emotional salience. Because of OOV words and spontaneous speech, 100% accuracy cannot be achieved. The baseline performance on references is UAR=71.2 % and WAR=70.5 %, while the priors are UAR=50.0 % and accuracy or WAR=66.3 %. As the model shown in Table 2 is more 'compact' for the training on hypotheses (it contains 422 words, instead of 797), it is also more consistent, so that performance degrades only minimally (to UAR=70.6 %, WAR=70.0 % for the best hypotheses), when compared to training and testing on the references. For word accuracies below about 60 %, the performance can even become worse than chance.

4.2. Dependency on hypothesis quality on independent data

Figure 2 shows results on independent test data. Using the references results in UAR=67.0 % and WAR=64.6 %, so there is a mismatch in absolute performance, but not with respect to the training method. Although the word accuracy of the recognizer is not affected (see Tables 3 and 4), the emotion recognition degrades slightly, but gradually, as the word accuracy falls below ≈65 %. As there is no big discrepancy between "matched" and "best" training, it does not pay to transcribe training or adaptation data with more care than testing data.

Figure 3 shows the performance using a BOW SVM classifier instead of one using emotional salience. This exhibits generally better performance, but UAR degrades sharply at ≈77 %.

Using the best hypotheses, emotional salience seems to be more affected by the training/ testing mis-match. On the other hand, BOW SVM classifiers seem to be more affected by the quality of the hypotheses used for training. For all systems, training on good word hypotheses is superior to training a model on references, presumably due to the match between training and testing conditions. This is especially true, if word confidences or other side information from the speech recognizer can be incorporated, which is not available for transcripts. In addi-

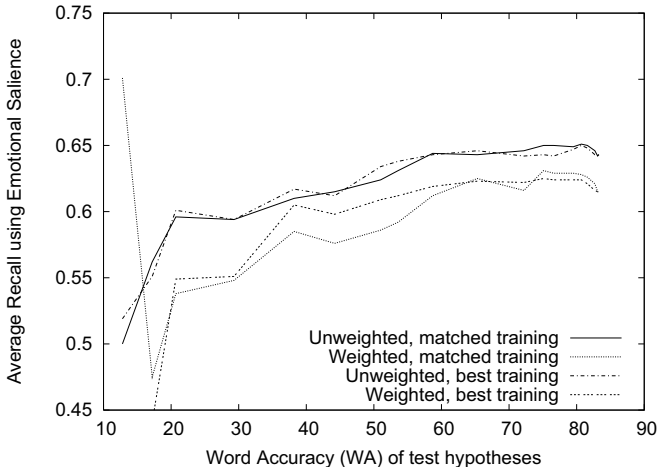

Figure 2: *Recall versus word accuracy using independent test data, for an emotional salience model as in Figure 1. "Matched training" refers to a system trained on CMU/ T-Labs hypotheses with roughly the same word accuracy as in the test case, while "best training" refers to a system trained on the best available hypotheses (with 82.9 % word accuracy).*

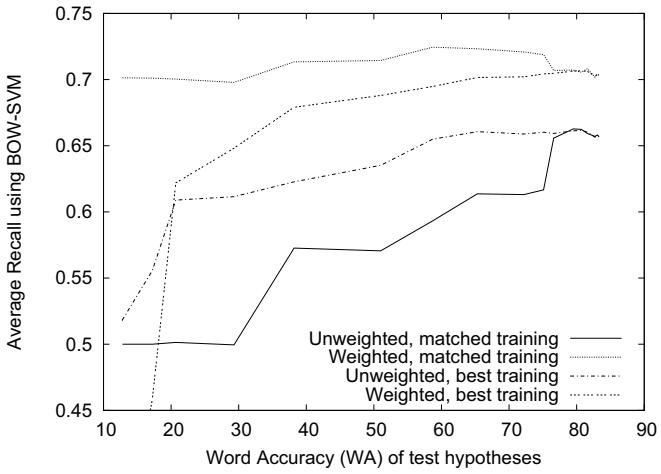

Figure 3: *Recall versus word accuracy using BOW SVM classifier, comparable to Figure 2, on CMU/ T-Labs hypotheses. The weighted average recall is above chance level for the best training, but the "matched training" case degrades with decreasing word accuracy of the training transcriptions.*

tion, systematic mistakes such as the frequent mis-recognition of "nein" as "Neid" by the German recognizer will be trained into the model automatically.

5. Summary and Outlook

This paper compares approaches to automatically discriminate between utterances labeled as "angry" and "non-angry", using only the words of children, who are engaged in a dialog with an Aibo robot dog. Previous experiments have shown that worthwhile gains can be achieved by combining linguistic and acoustic features for emotion recognition, which is also what humans do. In this paper, we investigated the accuracy level needed in order to be able to use speech recognition hypotheses during training and testing of classifiers based on linguistic features, so that these can be used without word-level manual transcriptions.

Our results show that word accuracy can be well below 100 % for both training and testing with a model, with a gradual deterioration to even sub-chance levels occuring for low word accuracies. An emotional salience model seems to be less sensitive to training with imperfect transcriptions than a BOW model, which also exhibits sharper drop-off points, as the "matched" and "best" curves in Figures 1 and 3 deviate less when using emotional salience. Also, as some speech recognition mistakes tend to be systematic for emotional speech, references are not crucial in order to obtain best overall performance.

Recall (in %)	on FAU Hypotheses			on CMU/ T-Labs Hypotheses
	BOW Model		Emotional	
	SVM	DMNB	Salience	
UAR train	65.9	66.1	70.3	74.2
WAR train	71.4	70.6	69.5	72.8
UAR test	62.4	64.9	64.2	64.3
WAR test	72.8	67.1	62.2	61.5

Table 4: *Comparison of FAU and CMU/ T-Labs hypotheses for training emotional salience, and using various classifiers.*

Future experiments will therefore address the integration of linguistic and acoustic parameters, using word-level segmentations. Also, word confidences generated by ASR could successfully be included in emotional salience [4], so that their inclusion in BOW based approaches should be investigated, too. Separate acoustic models could also be trained for separate, known emotions, whose use could provide even more side information for an integrated classifier.

6. References

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

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

[3] B. Schuller, A. Batliner, S. Steidl, and D. Seppi, "Emotion recognition from speech: Putting ASR in the loop," in *Proc. ICASSP*, Taipei; Taiwan, May 2009.

[4] F. Metze, T. Polzehl, and M. Wagner, "Fusion of acoustic and linguistic speech features for emotion detection," in *Proc. International Conference on Semantic Computing (ICSC)*. Berkeley, CA; USA: IEEE, Sep. 2009.

[5] B. Schuller, S. Steidl, and A. Batliner, "The INTERSPEECH 2009 Emotion Challenge," in *Proc. INTERSPEECH*. Brighton; UK: ISCA, 2009, pp. 312–315.

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

[7] A. Batliner, R. Kompe, A. Kießling, M. Mast, H. Niemann, and E. Nöth, "M = Syntax + Prosody: A syntactic–prosodic labelling scheme for large spontaneous speech databases," *Speech Communication*, vol. 25, no. 4, pp. 193–222, September 1998.

[8] T. Joachims, "Text categorization with support vector machines: learning with many relevant features," in *Proceedings of ECML-98, 10th European Conference on Machine Learning*, C. Nédellec and C. Rouveirol, Eds. Springer, Heidelberg, 1998, pp. 137–142.

[9] J. Su, H. Zhang, C. X. Ling, and S. Matwin, "Discriminative Parameter Learning for Bayesian Networks," in *Proc. ICML*, Helsinki, 2008, pp. 1016–1023.

[10] H. Soltau, F. Metze, C. Fügen, and A. Waibel, "A One-pass Decoder based on Polymorphic Linguistic Context Assignment," in *Proc. Automatic Speech Recognition and Understanding (ASRU)*. Madonna di Campiglio, Italy: IEEE, Dec. 2001.

[11] G. Stemmer, *Modeling Variability in Speech Recognition*. Berlin: Logos, 2005.