COMPARING FEATURE SETS FOR ACTED AND SPONTANEOUS SPEECH IN VIEW OF AUTOMATIC EMOTION RECOGNITION

Thurid Vogt[*,†]*, Elisabeth André*[*]

[*]Augsburg University, Germany
Multimedia concepts and applications
[†]Bielefeld University, Germany
Applied Computer Science
{andre,vogt}@informatik.uni-augsburg.de

ABSTRACT

We present a data-mining experiment on feature selection for automatic emotion recognition. Starting from more than 1000 features derived from pitch, energy and MFCC time series, the most relevant features in respect to the data are selected from this set by removing correlated features. The features selected for acted and realistic emotions are analysed and show significant differences. All features are computed automatically and we also contrast automatically with manually units of analysis. A higher degree of automation did not prove to be a disadvantage in terms of recognition accuracy.

1. INTRODUCTION

Many features for emotion recognition from speech have been explored. However, there is still no agreement on a fixed set of features. We present a data-mining experiment where we computed a large set of acoustic features providing different views on pitch, energy and MFCC time series of the data. Then, we automatically selected from these the best subsets for given data sets. This approach is quite common in speech emotion recognition (e. g. see [1],[2],[3]), but unlike previous work we start with more than 1000 features as opposed to just a few hundred.

In view of a future online emotion recognition system we want to investigate the following questions: Does a large number of features provided to the selection algorithm enable the selection of a better feature set? What degree of automation is feasible, i. e. which analysis units and features can be calculated automatically in an online system and still yield good results? Acted and realistic emotions have been compared before (e. g. [2],[3]) but not in regard to feature sets. Therefore, the question arises of how do optimal feature sets for both types of data differ?

The steps of the feature extraction from the speech signals are described in the next section. Then we present the databases on which we performed the experiments and give our evaluation results.

This work was partially funded by a grant from the DFG in the graduate program 256 and by the EU Network of Excellence Humaine.

2. FEATURE EXTRACTION

Prosodic features that are commonly used in the literature for speech emotion recognition are based on pitch, energy, MFCCs (Mel Frequency Cepstral Coefficients), pauses, duration and speaking rate, formants and voice quality features (e. g. [1], [2],[3]). Features are derived from these measurements over a given time segment. In our approach, we compute a multitude of features and then select the most relevant ones for the given application. While this concept is followed also by others, this work is intended to be more exhaustive: instead of reducing from 100–200 features, we start with almost 1300 features.

The process of feature extraction can be divided into 3 steps: choosing a segment length, calculating features over that segment and then reducing the feature set to the most relevant ones. These steps are now explained in more detail.

2.1. Segment length

Single pitch or energy values are not meaningful for emotions, but rather their behavior over time. Therefore, normally statistics such as mean, minimum or maximum from time series of these measures are computed. Thus, the time series of values have to be segmented into chunks from which to compute the statistics. These time segments have to be chosen very carefully as they have to fulfil two conflicting conditions: 1) emotion changes can occur very quickly, but the segment length sets the temporal resolution of recognizable changes, 2) reliable statistical features can often only be computed over longer segments. To find the best trade-off we experimented with several kinds of segments.

One possibility is to use a fix segment length, e. g. 500 ms. Other units can be linguistically motivated such as words, words with context, segments delimited by pauses or whole utterances. While whole utterances usually exhibit very distinctive contours for emotional states, they are

not practical for online recognition of emotions in natural speech, because no changes of the user's emotional state within the utterance could be recognized. Furthermore, for most linguistic units a speech recognizer is sufficient, but for utterance segmentation additional language processing software is needed. However, we can use whole utterances as units in acted speech, which is normally recorded in this segmentation, and consider the achieved recognition accuracy as an upper bound of what is possible. Words can be problematic as they are sometimes only a few milliseconds long so that not even a single pitch value can be estimated confidently. Thus, we tested words in context and segments divided by longer pauses for spontaneous emotions and, as a reference, words, words with context, utterances and 500 ms segments for acted emotions. Words in context consist of a word with its leading and subsequent word.

2.2. Feature calculation

We chose as the basis of our features pitch, energy and MFCC time series.

Pitch is obtained using the algorithm described in [4] with values ranging from 75 – 600 Hz calculated every 10 ms over a frame of 80 ms. Energy and 12 MFCCs are obtained from the ESMERALDA environment for speech recognition [5]. Values are computed every 10 ms for a frame length of 16 ms. We also use first and second derivatives of energy and MFCCs.

From these basic series we derive the following further series: from pitch, the series of the minima and the maxima, temporal distances, magnitudes and slopes between minima and maxima and also between maxima and minima; from energy again the series of the minima and the maxima, temporal distances, magnitudes and slopes between minima and maxima and between maxima and minima; from energy derivatives, the series of the minima and the maxima; from MFCCs, a series with the mean of all 12 MFCCs and the same for the first and second derivatives, as well as for all these series, the series of the minima and the maxima;

For each of these series, mean, maximum, minimum, range between minimum and maximum, variance, median, first quartile, third quartile and interquartile range of a segment are determined (like [1]). These values build up the feature vector.

A few additional features are joined to the feature vector: In order to diminish gender differences, pitch mean, median, first and third quartile are normalised by minimum and maximum pitch of the respective segment following the formula $mean_{norm} = \frac{mean-min}{max-min}$ (accordingly for median and quartiles).

Further features are the position of the overall pitch maximum, which approximates the main accent in linguistically motivated segments, the number of pitch and energy minima and maxima per segment as indicators for pitch and energy contours, and the ratio of the number of voiced frames to the total number of frames in a segment as coarse measure for pauses.

Speaking rate is not explicitly represented in the feature vector but the temporal distance between the energy minima and maxima is an approximation for that.

Although some of the features have only approximative character, their advantage is that they can be computed very fast, which is important in respect to the intended use of online feature extraction. Altogether, the features accumulate to a total of 1280.

2.3. Feature selection

The feature vector as described in the last section contains a lot of features, many of them probably redundant or not relevant. But the purpose of computing so many features is to let the data decide on the most significant features.

The data mining software Weka [6] was used for finding the best feature subset. We decided on correlation-based feature selection (CFS, [7]) as feature evaluator and Best-First search to find the best subset of features.

CFS is especially good with classifier Naïve Bayes (see section 4.1), because Naïve Bayes performs badly when features are highly correlated. CFS exactly removes those correlated attributes.

In general, selection reduces the originally 1280 features to about 90-160. This is significant and speeds up classification a lot, particularly, since feature selection needs to be done only once for every application.

3. DATABASES

3.1. Actors database

This database was recorded at the Technical University, Berlin [8]. 10 professional speakers (5 male, 5 female) were asked to pretend 6 different emotions (anger, joy, sadness, fear, disgust and boredom) as well as a neutral emotional state in 10 utterances each. The content of the utterances was emotionally neutral. The utterances from the collected material that were perceived as unnatural by test persons were discarded, ending up with 493 utterances total (female: 286/male: 207).

The recordings are characterised by a very high quality because they were originally intended to be used for emotional speech synthesis. This database is a comparably easy task for emotional speech recognition, but quite far from realistic settings.

3.2. Wizard-of-Oz database

Data from Wizard-of-Oz (WOZ) studies comes very close to real life data as people behave naturally and do not fol-

low a script. To get closer to natural emotions our features were also evaluated on the SmartKom corpus. This WOZ database was recorded at the University of Munich within the SmartKom project [9]. Subjects interacted with a multimodal dialogue system, not knowing that their emotional state was observed. While these emotions can be considered quite realistic, unfortunately, the biggest part of the speech is emotionally neutral. Another problem results from the fact that emotions were labelled considering both audio and visual information. Sometimes these labeled emotions are hardly identifiable from the speech signal alone. As a consequence, this corpus represents a greater challenge for emotion recognition than the corpus with the acted emotions.

The following emotions, referred to in SmartKom as "user states", were labeled: strong joy, weak joy, surprise, helplessness, weak anger and strong anger, as well as emotionally neutral segments. Emotions are distributed very unequal, with almost 90 % of the speech being neutral, though this should also be the case in most real applications.

4. EVALUATION

4.1. Classification

The Weka data mining software is again used as a toolbox for classification. All experiments described here use Naïve Bayes as learning scheme. Other schemes were also tested, but no big differences were observed and Naïve Bayes has the advantage of being fast, even when dealing with high-dimensional data. Furthermore, it still performs satisfactorily when the majority of instances belongs to only one class, which is the case for the SmartKom Corpus. This is decisive for our purposes as we want to test the feature extraction and keep the classifier constant.

4.2. Results

4.2.1. Acted emotions:

Acted emotions were evaluated in 4 different arrangements: all 7 emotions (anger vs. joy vs. sadness vs. fear vs. disgust vs. boredom vs. neutral), evaluation (anger/sadness/fear/disgust/boredom vs. neutral vs. joy), activation (anger/joy/fear /disgust vs. neutral vs. boredom/sadness) and emotional/non-emotional (anger/joy/sadness/fear/disgust/boredom vs. neutral). Class-wise recognition accuracy obtained from 10-fold cross-validation over all utterances is given.

Table 1 shows recognition results for the 4 arrangements comparing the full feature set with the reduced feature set and using the whole utterance as segment. Reducing the feature set brings an average improvement of 6.4 %. Besides, classification with the reduced feature set is faster.

Table 2 shows results for different segment lengths with reduced feature sets for all 7 emotions. A considerable decrease in recognition accuracy can be observed when segment length gets shorter. Though all results are well above chance level, looking at the results in respect to usefulness in an application, only words-in-context come off well.

	7 emotions	Evaluation	Activation	Emo./Non-Emo.
Full set	69.1 %	67.1 %	85.4 %	81.9 %
Reduced set	77.4 %	72.5 %	88.6 %	85.3 %

Table 1. Comparing the full feature set with the reduced feature set.

Segment length	Recognition accuracy
Whole utterance	77.4 %
Word in context	53.2 %
500 ms	44.5 %
Word	34.1 %

Table 2. Comparing segment lengths (reduced feature sets)

4.2.2. WOZ emotions:

We evaluated our approach with the same scheme as [10] who also build an emotion recognition system for the SmartKom Corpus. They used a different extract, but results should be comparable, as the expression of emotions is consistent throughout the corpus and the amount of data in the extracts is similar.

Our results (see Table 3) are similar to theirs but our feature set was computed completely automatically whereas their features are partially annotated by hand (prosodic peculiarities) and they additionally use part-of-speech flags. Obviously, a higher degree of automation is no disadvantage. We suppose that the larger feature set compensates for this.

As for the two units of analysis, the longer unit (segments delimited by pauses) again comes off better, but the difference is not as striking. The reason for that is that in spontaneous speech, phrase and word contours are less distinct. When comparing results for the reduced and the full feature sets differences are also not as big. In some cases, the reduced feature set performs equal or worse than the full feature set. Still, feature selection has the advantage of faster classification.

4.3. Selected features

As in [1] the selected features are not those one would normally suspect. Generally we can say, the more classes, the more features are necessary. In acted speech, pitch-related features play a dominant role. For spontaneous emotions, the focus lies more on MFCCs, rather low coefficients are

Different granularities of user states							Full set	Reduced set	Full set	Reduced set
							Pauses as borders		Word with context	
joyful strong	joyful weak	surprised	neutral	helpless	angry weak	angry strong	26	25.6	28.4	28
joyful		surprised	neutral	helpless	angry		37.5	38.7	31.2	35.7
joyful			neutral	helpless	angry		39	40.6	39.5	36.1
joyful			neutral	problem			48.3	51.6	44.2	42.4
no problem				helpless	angry		50.3	51.9	45.9	45.4
no problem				problem			68.3	73.3	59.3	59.4
not angry					angry		59.9	61.1	59.1	50.5

Table 3. Recognition results in % for natural emotions using segments delimited by pauses and words with context as units.

selected and mainly the first derivatives. The extrema series of pitch and energy are more important than the basic series.

Pauses are a very important feature for acted emotions because there is a high proportion of pauses in the sad emotion. This aspect, however, can not be generalised to realistic emotions as pauses are always discarded there.

5. CONCLUSIONS

Our results demonstrate the very different demands of acted and realistic data. Consistent with other work, we found acted emotions to be more easily recognized than realistic emotions and the impact of feature selection to be higher for acted speech. The novel contribution of this paper is that we looked closer on the differences in the selected feature sets for acted and spontaneous emotions. Good feature sets for acted and spontaneous emotions showed to overlap little as pitch related features were predominantly used for acted speech and MFCC (esp. low coefficients) related features for spontaneous emotions. These differences suggest that, when intending to recognize natural emotions, it may often not make sense to use acted data even only for a first test of methods.

Finally, we did not find a high degree of automation in feature extraction and unit segmentation to be a disadvantage and we assume the large number of features we provided to the selection process to be responsible for this.

6. REFERENCES

[1] P.-Y. Oudeyer, "The production and recognition of emotions in speech: features and algorithms," *Int. Journal of Human-Computer Studies*, vol. 59, no. 1–2, pp. 157–183, 2003.

[2] A. Batliner, K. Fischer, R. Huber, J. Spilker, and E. Nöth, "How to find trouble in communication," *Speech Communication*, vol. 40, pp. 117–143, 2003.

[3] D. Küstner, R. Tato, T. Kemp, and B. Meffert, "Towards real life applications in emotion recognition," in *ADS Workshop 04*, Kloster Irsee, Germany, 2004, pp. 25–35.

[4] P. Boersma, "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*, U. of Amsterdam, 1993, pp. 97–110.

[5] G. A. Fink, "Developing HMM-based recognizers with ESMERALDA," in *Lecture notes in Artificial Intelligence*, V. Matoušek et al., Eds., vol. 1962, pp. 229–234. Springer, Berlin, Heidelberg, 1999.

[6] I. H. Witten and E. Frank, *Data Mining: Practical machine learning tools with Java implementations*, Morgan Kaufmann, San Francisco, 2000.

[7] M. A. Hall, "Correlation-based feature subset selection for machine learning," M.S. thesis, U. of Waikato, New Zealand, 1998.

[8] F. Burkhardt, *Simulation emotionaler Sprechweise mit Sprachsynthesesystemen*, Ph.D. thesis, TU Berlin, Germany, 2001.

[9] S. Steininger, F. Schiel, O. Dioubina, and S. Raubold, "Development of user-state conventions for the multimodal corpus in SmartKom," in *Proc. Workshop 'Multimodal Resources and Multimodal Systems Evaluation' 2002*, Las Palmas, 2002, pp. 33–37.

[10] A. Batliner, V. Zeißler, C. Frank, J. Adelhardt, R. P. Shi, and E. Nöth, "We are not amused - but how do you know? User states in a multi-modal dialogue system.," in *Proc. EUROSPEECH 2003*, Geneva, 2003, pp. 733–736.