Exploring the benefits of discretization of acoustic features for speech emotion recognition

Thurid Vogt, Elisabeth André

Multimedia Concepts and their Applications, University of Augsburg,Germany
{vogt,andre}@informatik.uni-augsburg.de

Abstract

We present a contribution to the Open Performance subchallenge of the INTERSPEECH 2009 Emotion Challenge. We evaluate the feature extraction and classifier of EmoVoice, our framework for real-time emotion recognition from voice on the challenge database and achieve competitive results. Furthermore, we explore the benefits of discretizing numeric acoustic features and find it beneficial in a multi-class task.
Index Terms: speech emotion recognition, discretization of features

1. Introduction

Emotion recognition from speech has made considerable advances in the last years. The number of research studies of emotional speech databases has grown, and also first applications and prototypes have been developed [1, 2]. There are large EU projects (e. g. Callas[1] and Semaine[2]) that push real-time emotion recognition. The real-time recognition of emotion in speech is also our goal, for which we have developed EmoVoice, our framework for real-time emotion recognition from voice [3], that has already been integrated in a number of prototypes and showcases. However, real-time processing sometimes requires to accept lower recognition accuracies compared to offline research systems. In this contribution to the Open Performance subchallenge of the INTERSPEECH 2009 Emotion Challenge we evaluate our methodology to assess if it is competitive. Since our main focus lies on the acoustic features, we also explore here whether a discretization of numeric acoustic features can make the classification problem easier. Though being promising, discretization has not been investigated extensively so far. For instance, Casale and colleagues [4] achieve an improvement by feature discretization on two small databases with acted emotions. Here, we study the effects of discretization on a large database with spontaneous emotions such as the Challenge database.

The rest of this paper is organized as follows: first, we briefly characterize the challenge database. Next we present our methodology to speech emotion recognition, which includes feature extraction, feature selection and classification. Afterwards, we present and discuss our results on the database with two pre-processing strategies.

2. Database

All experiments in this paper are carried out on the database provided by the organizers of the INTERSPEECH 2009 Emotion Challenge, the German FAU AIBO database, a large database with children's emotional speech. The challenge is split into two tasks: the 2-class problem with the classes negative (NEG) and non-negative (IDL), and the 5-class problem with the classes anger (A), emphatic (E), neutral (N), positive (P) and rest (R). More information on this database can be found in the introductory paper of the challenge [5].

3. Features

Our approach to feature extraction is for the most part a generative one, which means that starting from basic acoustic observations, or low level descriptors (LLD), that are calculated equidistantly over the classification unit (here chunks), we systematically apply statistical functions to derive features on chunk level. The LLDs we use are pitch, energy, MFCCs, the short-term frequency spectrum, and the harmonics-to-noise ratio. The feature set consists of 7 feature types, which are pitch, energy, duration, spectral, cepstral, voiced segments (derived from pitch) and voice quality features. The distribution of features on types is shown in Table 1. Note that the features are categorized here according to the basic feature type they were derived from. This is not necessarily a phonetically valid categorisation, — for example Batliner et al. [6] denote features related to the position of pitch on the time axis as durational features — but rather a practical categorisation. For each sequence of LLD values within a chunk we calculate the following 9 statistical functions: mean, minimum, maximum, range, standard deviation, median, first quartile, third quartile and interquartile range.

Table 1: *Number of features per types.*

Feature type	Number of features
pitch	208
energy	110
duration	4
spectrum	45
cepstrum	1053
voiced segments	19
voice quality	12
Σ	1451

As a generative approach usually yields a high number of features, in our case almost 1500, we apply a feature selection algorithm to obtain a smaller set with the most relevant features for the particular data set. Furthermore, we explore whether discretization of continuous features is beneficial because it may help the classifier to better estimate parameters.

The feature extraction is implemented within the EmoVoice framework for real-time speech emotion recognition. Our main

[1] http://www.callas-newmedia.eu
[2] http://www.semaine-project.eu

goal is to use fast and fully automatic algorithms for feature extraction and classification even if that may degrade recognition accuracy to some extent. For example, we do not use word information as this would require a time-consuming speech recognizer to precede the emotion recognition. Our method for feature extraction is similar to the one used to achieve the baseline results for the challenge [5]. However, the feature set we give to the selection process is bigger and contains additional feature types, spectral and voicing based features.

3.1. Pitch

Pitch was estimated for frames of 80 ms length at a rate of 10 ms with the algorithm described by Boersma [7], whose high accuracy is generally accepted and which is part of the popular phonetics software Praat [8]. Pitch is very prone to inter-speaker differences, thus not only raw pitch was considered, but also logarithmised pitch, and a normalization was obtained by subtracting the median from the logarithmised pitch values [9]. Additionally, the unlogarithmised pitch mean, median, first and third quartile values were normalized by minimum and maximum pitch of the respective segment. Presumably, however, features modeling the dynamic behavior of pitch, so for example related to the temporal distance between local extrema or relative magnitude of extrema, are more important than the specific (raw or normalized) values of pitch. Therefore, the above described series of raw, logarithmised and median-subtracted log pitch values were further transformed into the series of the local maxima, the local minima as well as the difference, distance and slope between adjacent local extrema. Furthermore, first and second derivation were obtained. Then, the 9 statistical functions were applied to these pitch series.

Further pitch features are the position of the overall pitch maximum, which approximates the main chunk accent, and the position of the overall pitch minimum. As indicators for pitch contours, the number of minima, maxima, falling and rising values were obtained. All these values were normalized by the number of pitch values in the segment, ending up with 208 pitch related features in total.

3.2. Energy

Energy was obtained using the ESMERALDA environment for speech recognition [10] where it is calculated as mean adjusted and logarithmised signal energy for frames of 16 ms length at a rate of 10 ms.

As for pitch, the series of only the local maxima and only the local minima were created from the energy curve, as well as difference, distance and slope between adjacent local extrema. First and second order derivation together with the series of their local maxima and local minima were further added to the number of 12 resulting energy related series, to which again the statistical functions were applied.

Also, the position of the global maximum and the number of local maxima, both normalized by the number of frames in the segment, were added to the feature vector, which finally contained 110 energy related features.

3.3. Duration

The duration features fall out of the generative approach as they are not based on low-level descriptors. They include the chunk length, measured in seconds, and the zero-crossing rate to roughly decode speaking rate. Furthermore, pause is obtained as the proportion of non-speech calculated by a voice activity detection algorithm from the signal energy [10] and also approximated by the ratio of unvoiced pitch frames to the total number of pitch frames in the chunk.

3.4. Spectrum

In order to obtain spectral features, FFT was applied to frames of the acoustic signal of 16 ms length at a rate of 10 ms, thus yielding a series of short-term frequency spectra. Since especially information on the slope of the spectrum was regarded as important, for each short-term spectrum, the distance between the 10th and the 90th percentile, the slope between weakest and strongest frequency, as well as two linear regression coefficients were calculated by ordinary least-square estimation. Furthermore, the center of gravity of the spectrum according to the following formula

$$CG = \frac{\sum f_i \cdot E_i}{\sum E_i} \tag{1}$$

was obtained, where f_i is the ith frequency of the spectrum and E_i the energy of f_i in the spectrum. It parametrizes the spectral balance between high and low frequencies for a signal segment [11]. Each of these 5 values yielded a new 1-dimensional time series, to which the statistical functions listed earlier were applied.

3.5. Cepstrum

MFCC calculation was again obtained from ESMERALDA. Twelve coefficients were derived at the same frame length and rate as signal energy. In ESMERALDA, additionally an adaptation to the microphone channel is performed.

Adding the first and second derivatives of each coefficient, this yielded 36 MFCC time series for each signal segment. From each series, the local maxima and minima series were also derived. Furthermore, a condensed representation — the average over all 12 coefficients, for each basic, first and second derivation — was contrasted to the single features. This produced in total 1053 MFCC related features.

3.6. Voiced segments

The length and distribution of voiced and unvoiced segments, as calculated by the pitch algorithm, in a speech signal is related to voice characteristics. Therefore, the lengths of both the voiced and the unvoiced segments in the chunks were counted. To the resulting two series of values, the statistical functions were applied, yielding e. g. the mean length of voiced segments in the unit. Furthermore, the number of voiced segments normalized by the number of pitch frames in the chunk was calculated.

3.7. Voice quality

In order to model voice quality, jitter and shimmer of the glottal pulses of the whole segment, as well as the number of glottal pulses normalized by the segment length in seconds are added to the feature set. Furthermore, statistics were derived from the Harmonics-to-Noise ratio calculated for frames of 160 ms length at a rate of 10 ms.

3.8. Discretization

Discretization of numeric features can help a classifier to better estimate parameters because each discrete value can accumulate many examples to learn from. This is especially true for

small databases when there are not many examples for the classifier to learn from, but may also be beneficial for relatively large databases such as the database investigated here. In order to discretize our features, we use the filter for supervised discretization by Fayyad and Irani's MDL method [12] from the Weka data mining software package [13]. It finds suitable intervals of values and thus maps a continuous domain into discrete nominal features. The number of intervals for our features ranges from 2 to 8 resp. 9 for 2 and 5 classes.

3.9. Feature selection

The extraction of acoustic features yields in total 1451 features. In order to discard irrelevant features (which may degrade classification performance) and to reduce the search space (which accelerates classification), we carry out a feature selection before classification. We decided on correlation-based feature subset selection (CFS) [14] from Weka, a selection technique that searches for subsets with features that are highly correlated with the class but not with each other and that we already found suitable in previous work [15].

For the experiments with discretized features, this was done before the feature selection. However, discretization has only a minor effect on the selected subset. This holds for the distribution on feature types, which is shown in Table 2 for the 2-class and 5-class problem, each with and without discretization, but as well for the actual selected features. Obviously, 2 classes can be modeled with considerably less features than 5 classes.

Table 2: *Distribution of features on types after feature selection, with and without prior discretization (D).*

Feature type	Number of features			
	2-class cont.	2-class dis.	5-class cont.	5-class dis.
pitch	15	16	27	26
energy	17	17	22	23
duration	0	0	0	0
spectrum	15	15	21	21
cepstrum	24	24	75	74
voiced segments	4	4	4	5
voice quality	3	3	5	5
Σ	78	79	154	154

4. Classification

Since the number of test set evaluations was limited for the challenge, results were only obtained from one classifier, the Naïve Bayes classifier from Weka. The same algorithm is also used in our EmoVoice framework. It is a simple, but fast algorithm that matches our requirements of real-time performance in various European projects we are conducting, even if this may slightly degrade the overall recognition performance. Still, as we will show in the next section, it is able to achieve comparable or even better results than the baseline results in [5] that were obtained from a more sophisticated SVM classifier, thus it can be considered a competitive classifier. Since the distribution of classes is unbalanced, especially for the 5-class problem, we also apply the Synthetic Minority Oversampling Technique (SMOTE) from Weka as suggested by [5], to obtain a more balanced class distribution for the training set. In particular, we double the number of instances in the classes A, P and R (5-class problem) and add a further 90 % of instances to class NEG (2-class problem).

5. Evaluation results

Table 3 shows our results for the Open Performance Challenge. Results are obtained for the 2-class and the 5-class problem and show the effects of pre-processing by discretization and training set balancing. As the class distribution is unbalanced also in the test set, result discussions are restricted to the unweighted average recall as the primary measure to optimize.

Table 3: *Classification results for the Open Performance Challenge. Pre-processing strategies: discretization (D) and balancing of training set by SMOTE (B). Recall and Precision are given both as unweighted average (UA) and weighted average (WA).*

	Preprocessing		**Recall** [%]		**Precision** [%]	
			UA	WA	UA	WA
2-class	-	-	66.4	66.6	64.0	71.3
	D	-	66.3	64.4	63.6	71.6
	D	B	66.3	64.3	63.7	71.6
5-class	-	-	35.9	23.8	28.7	57.8
	D	-	39.0	43.7	30.9	58.7
	D	B	39.4	41.1	30.6	59.2

Obviously, discretization and training set balancing have no effect in the 2-class problem, while they improve accuracy considerably in the 5-class problem. In the latter, discretization increases the unweighted averaged recall by 3.1 % (without SMOTE) and 3.3 % (with SMOTE), which is a significant improvement (confidence margin is ± 1.0). This is probably due to the training set being large enough to let the classifier learn continuous parameters for 2 classes so that discretization is not necessary. However, for 5 classes, discretization has a positive effect which suggests that the training set, though huge compared to other databases, is still not large enough for continuous features.

In comparison to the baseline results given in [5] (see Table 4 for a summary) we find that our results without discretization and SMOTE exceed the baseline results without pre-processing by 3.7 % for the 2-class problem and 7 % for the 5-class problem. But, since we cannot achieve a further improvement by discretization and training set balancing for the 2-class problem, we also cannot outperform the best baseline result for this task. However, in the 5-class problem, we actually can exceed the best baseline result by application of discretization and SMOTE significantly by 1.2 %.

Table 4: *Summary of the baseline results of [5]*

Task		**UA Recall** [%]
2-class	no pre-proc.	62.7
	best result w. pre-proc.	67.7
5-class	no pre-proc.	28.9
	best result w. pre-proc.	38.2

6. Conclusion

We evaluated our feature set and classifier, which are specifically tailored for real-time processing, on the INTERSPEECH 2009 Emotion Challenge database. In particular, we explored the use of discretized numeric features for this large database and we found that discretization is not only beneficial for small databases, but also for larger databases if there are many classes. We assume this result holds also for other databases than the one investigated here and for other feature sets, which we will explore in our future work. Our result may also indicate that the Aibo database, though it is comparably huge, is still not large enough to estimate — with continuous features — 5 classes sufficiently, at least some of the classes with few samples, and it further proves the need for large databases for speech emotion recognition. As currently, there are only few large databases available for speech emotion recognition, discretization is presumably useful in many cases.

Furthermore, with our techniques, we could exceed the baseline result of the challenge for the 5-class problem and achieved comparable results for the 2-class problem though we used a faster but possibly less powerful classification algorithm.

7. Acknowledgements

This work was partially financed by the EU in the CALLAS Integrated Project.

8. References

[1] S. W. Gilroy, M. Cavazza, R. Chaignon, S.-M. Mäkelä, M. Niranen, E. André, T. Vogt, J. Urbain, M. Billinghurst, H. Seichter, and M. Benayoun, "E-tree: emotionally driven augmented reality art," in *Proceedings of the 16th ACM International Conference on Multimedia*. Vancouver, BC, Canada: ACM, 2008, pp. 945–948.

[2] C. Jones and J. Sutherland, "Acoustic emotion recognition for affective computer gaming," in *Affect and Emotion in Human-Computer Interaction*, ser. LNCS, C. Peter and R. Beale, Eds. Heidelberg, Germany: Springer, 2008, vol. 4868.

[3] T. Vogt, E. André, and N. Bee, "EmoVoice — A framework for online recognition of emotions from voice," in *Proceedings of Workshop on Perception and Interactive Technologies for Speech-Based Systems*. Kloster Irsee, Germany: Springer, June 2008.

[4] S. Casale, A. Russo, G. Scebba, and S. Serrano, "Speech emotion classification using machine learning algorithms," in *2008 IEEE International Conference on Semantic Computing*, Santa Clara, CA, USA, August 2008, pp. 158–165.

[5] B. Schuller, S. Steidl, and A. Batliner, "The Interspeech 2009 Emotion Challenge," in *Interspeech*, ISCA, Ed., Brighton, UK, 2009.

[6] A. Batliner, J. Buckow, R. Huber, V. Warnke, E. Nöth, and H. Niemann, "Boiling down prosody for the classification of boundaries and accents in german and english," in *Proceedings of Eurospeech 2001*, Aalborg, Denmark, September 2001, pp. 2781–2784.

[7] P. Boersma, "Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound," in *Proceedings of the Institute of Phonetic Sciences*, University of Amsterdam, 1993, pp. 97–110.

[8] P. Boersma and D. Weenink, "Praat: doing phonetics by computer (version 5.1.04) [computer program]," http://www.praat.org/, April 16th 2009, last retrieved.

[9] A. Kießling, "Extraktion und Klassifikation prosodischer Merkmale in der automatischen Sprachverarbeitung," Ph.D. dissertation, Technical Faculty, University Erlangen-Nuremberg, 1996.

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

[11] R. J. J. H. van Son and L. C. W. Pols, "An acoustic description of consonant reduction," *Speech Communication*, vol. 28, no. 2, pp. 125–140, June 1999.

[12] U. M. Fayyad and K. B. Irani, "Multi-interval discretization of continuousvalued attributes for classification learning," in *Thirteenth International Joint Conference on Articial Intelligence*, Chambéry, France, 1993, pp. 1022–1027.

[13] I. H. Witten and E. Frank, *Data Mining: Practical machine learning tools with Java implementations*, 2nd ed. San Francisco: Morgan Kaufmann, 2005.

[14] M. A. Hall, "Correlation-based feature subset selection for machine learning," Master's thesis, University of Waikato, New Zealand, 1998.

[15] T. Vogt and E. André, "Comparing feature sets for acted and spontaneous speech in view of automatic emotion recognition," in *Proceedings of International Conference on Multimedia & Expo*, Amsterdam, The Netherlands, July 2005.