Timing levels in segment-based speech emotion recognition

Björn Schuller, Gerhard Rigoll

Angaben zur Veröffentlichung / Publication details:

Schuller, Björn, and Gerhard Rigoll. 2006. "Timing levels in segment-based speech emotion recognition." In *INTERSPEECH 2006 - ICSLP, Ninth International Conference on Spoken Language Processing, Pittsburgh, PA, USA, September 17-21, 2006*, 1818–21. ISCA Archive. https://doi.org/10.21437/Interspeech.2006-502.

Timing Levels in Segment-Based Speech Emotion Recognition

Björn Schuller and Gerhard Rigoll

Institute for Human-Machine Communication
Technische Universität München
{Schuller | Rigoll}@tum.de

10.21437/Interspeech.2006-502

Abstract

Additional sub-phrase level information is believed to improve accuracy in speech emotion recognition systems. Yet, automatic segmentation is a challenge on its own considering word- or syllable boundaries. Further more clarification is needed which timing level leads to optimal results. In this paper we therefore quantitatively discuss three approaches to segment-level features based on 276 statistical hi-level prosodic, articulatory and speech quality features. Apart from the choice of the optimal segmentation scheme also fusion of segments with respect to classification and combination of diverse timing levels is analyzed. Tests are carried out on the popular Berlin Database of Emotional Speech (EMO-DB). Significant improvement over existing works can be reported for combination of phrase-level features with relative time interval features.

1. Introduction

Emotion recognition is one of the challenges in order to approach human communication [1]. As analysis of affective intents in conversational speech is maturing, some general trends with respect to features, their selection and classification can be observed [2]. Most today's works rely on global phrase-wise statistics of derived low-level-descriptors (LLD) referring to intonation, intensity, duration and spectral development [3, 4]. Thereby a multivariate time-series, having a dynamic nature, is transformed into a single static feature vector for each emotional phrase or clip. Typical operations for such functional generation are linear moments as means, standard deviation, Kurtosis and Skewness, quartiles, quartile ranges, extremes and their ranges, or floor- and up-level-times. Such descriptive statistical analysis combined with static classification, e.g. by Support Vector Machines or Neural Nets, could be shown superior to direct dynamic processing by means of e.g. Hidden Markov Models (HMM) or Dynamic Bayesian Nets in the past [5]. However, recently sub-segment-based approaches are evolving questioning whether the phrase level is the right choice for functional generation [6, 7]. Popular are word-based statistics, which however require correct segmentation, or partitioning of an utterance without respect to the spoken content in relative intervals. Further more lower-level segmentation based on voicing probability or intensity levels seems a possibility. Apart from the right choice of segments also a fusion scheme for analysis of a complete utterance is crucial. The most simple choice would be majority voting over single segment decisions. Such votes may also be weighted by confidence levels or segment durations. Depending on the type of segment construction either a dynamic number of segments with respect to the phrase length or number of words, syllables or voiced segments contained is obtained, or a static number when splitting in relative time-intervals as thirds, or quarters. In this case a super vector can be constructed and classified in the conventional way. Especially within the fusion with other information sources as facial expression such timing levels may occur on a higher level, as facial expression analysis may deliver outputs on a multiple frames per second (fps) level. The aim of this paper therefore is the quantitative experimental discussion of sub-phrase-level timing in speech emotion recognition in order to boost performance compared to mere phrase-level features. Furthermore we aim at investigation whether emotion recognition is possible already within short speech clips, say macro-frames, compared to speaker- or phoneme recognition where an estimate is provided on a frame-basis.

The paper is structured as follows: Section 2 deals with the database used. In section 3 acoustic features are introduced in general. Section 4 is the core aspect of this paper: schemes for segmentation. Section 5 discusses classification and feature selection in the framework of segment-based speech emotion recognition. In the following section 6 experimental results are introduced and finally conclusions are drawn in section 7.

2. Database

In order to provide results on a public corpus we decided for the Berlin Emotional Speech database *(EMO-DB)* [8]. It consists of 816 phrases in total. The emotion set resembles the "Big 6", besides an exchange of surprise in favor of boredom. 10 German sentences of emotionally undefined content have been acted in these emotions by 10 professional actors, 5 of them female. Throughout perception tests by 20 subjects, 10 of them female, 488 phrases have been chosen that were classified as more than 60% natural and more than 80% clearly assignable. The database is recorded in 16 bit, 16 kHz under studio noise conditions. 84.3% mean accuracy is reported within a human perception test in [8]. As the spoken content is predefined in this database we do not use linguistic features in this work, as in our former works [9].

3. Acoustic Features

In former works [5] we compared static and dynamic feature sets for the prosodic analysis and demonstrated the higher performance of derived static features. Before single segments are constructed we will describe the general feature set that will be applied at each timing level. As an optimal set of such global features is broadly discussed [2,3,4], we consider an initially large set of 276 acoustic hi-level features which cannot all be described in detail here. However, the target is to

become utmost independent of the spoken content and ideally also of the speaker, but model the underlying emotion with respect to prosodic, articulatory and voice quality aspects. The feature basis is formed by the raw contours of zero crossing rate *(ZCR)*, pitch, first seven formants, energy, spectral development, and Harmonics-to-Noise-Ratio *(HNR)*. The following table shows the distribution of features among their general type. Thereby duration based features rely on common bi-state dynamic energy threshold segmentation and voicing probability.

Table 1: Distribution of the features

Type	**Pitch**	**Energy**	**Duration**	**Formant**
[#]	12	11	5	105
Type	**HNR**	**MFCC**	**FFT**	**ZCR**
[#]	3	120	17	3

In order to calculate the according Low-Level-Descriptors 20 ms frames of the speech signal are analyzed every 10 ms using a Hamming window function. Pitch is detected by the auto correlation function (ACF) with window compensation and dynamic programming (DP) for global error minimization. HNR also relies on the ACF. The values of energy resemble the logarithmic mean energy within a frame. Formants base on 18-point LPC spectrum and DP. We use their position and bandwidth, herein. For spectral development we use 15 MFCC coefficients and a FFT-spectrum out of which we calculate spectral flux, Centroid and 95%-roll-off-point after dB(A)-correction according to human perception. Low-pass SMA filtering smoothes the raw contours prior to the statistical analysis. First and second order regression coefficients are subsequently calculated. The higher level features are then derived by means of descriptive statistical analysis as linear moments, extremes, ranges, quartiles, or durations, and normalized. Overall the final per-segment feature vector consists of 276 features.

4. Segmentation Schemes

A variety of potential segmentation schemes exists in general. However, we focus on automatic segmentation without the necessity of word- or syllable-boundary detection, which is prone to errors and demands for considerable extra-effort as word alignment with an Automatic-Speech-Recognition (ASR) engine. We rather examine the impact on accuracy of fast and simple "blind" strategies that neglect spoken content but can easily be realized in real-time and partly allow for direct stream processing: Apart from the global time interval (GTI) we firstly investigate absolute time intervals (ATI). Thereby a phrase or speech stream is split into macro-frames at a set time interval, e.g. 500 msec. The idea is to test whether speech emotion recognition can be realized at a fixed frame rate compared to speaker recognition e.g. for fusion with a facial expression analysis stream. Interestingly, a dynamic number of macro-frames are obtained with respect to the overall length if a single phrase is analyzed. This demands for a different classification strategy as HMM or DBN, as a time-series on a higher level is obtained, or multi-instance learning with a suited fusion scheme as weighted majority vote.
As a second approach Relative Time Intervals (RTI) are obtained by splitting an utterance at fixed relative positions, e.g. halves or thirds. Afterwards, features are extracted for each part in the same manner as for GTI. A super-vector is constructed for classification by fusion of all segment features plus global features, which is named GRTI.
The variant ATIR combines absolute time intervals and the idea of relative positions. The advantage is the compensation of RTI's drawback that utterances of different lengths lead to sub-segments of different lengths. Likewise 500 msec segments are constructed at fixed relative positions as shown in figure 1. Thereby not all parts of the sample are contained in the analysis. GATIR compensates this by addition of global features. Figure 1 visualizes these variants and shows two utterances of different length to demonstrate the effect of each variant:

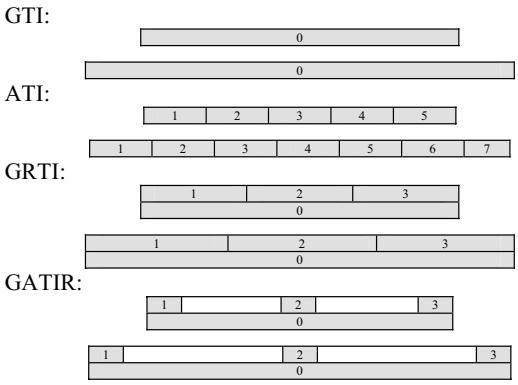

Figure 1: Notation segmentation schemes showing a short and a long utterance each. Numbers shown refer to segment-index. GTI: global time intervals, ATI: absolute time-intervals, RTI relative time intervals, ATIR: absolute time intervals at relative positions, GATIR: exemplary combination of segmental and global features

5. Classification and Feature Selection

In order to find the optimal classifier we demonstrate recognition performance for a multiplicity of classifiers. Among these are instance based nearest neighbor (1NN and kNN), a multi-layer Perceptron (MLP) as neural network representative, a decision tree (C4.5), Naïve Bayes (NB) and Bayesian Networks (BN), as well as ensemble construction with a single base classifier type by AdaBoosting, Bagging, and MultiBoosting and combination of diverse base classifiers by Stacking and Stacking with confidences (StackingC) with Multiple Linear Regression (MLR) as Meta-Level classifier. More details on classifiers and their setup is found in [9]. In order to classify a phrase that is split into several segments, a super-vector is constructed combining features of all segments. However, this is only possible if the number of segments is constant. ATI, as described in section 4, is therefore only considered herein to provide an impression of the possibility of emotion recognition out of short clips.

Apart from the choice of an optimal classifier also selection of the most relevant features is important as it saves computation time considering real-time processing and boosts performance as some classifiers are susceptible to high dimensionality. Therefore search for the right features seems mandatory. We chose SVM-SFFS within acoustic feature selection as it has proven a reasonable choice compared to NP-hard exhaustive

search and proved a good choice in former works [9]. The search is performed by forward and backward steps eliminating and adding features in a floating manner to an initially empty set. As relevance criterion the target classifier, namely SVM, is employed in a wrapper manner. Thereby a set is optimized as a whole rather than finding single attributes of high relevance. As a super vector is constructed, we can select features of different time segments in one pass. Thereby it can also be seen which time segment contributes mostly to the final decision. The optimal number of features is afterwards determined in accordance to the highest observed accuracy throughout selection.

6. Recognition Results

As a general mean of evaluation we employ 10-fold stratified cross validation (SCV). This validation method is used in many other works on the EMO-DB. However, the cross validation factor *j* has an important influence on the accuracies reported, as with a higher *j* more training data is available per run. Figure 2 in the next column therefore shows global features (GTI) with the full feature vector for diverse cross-validation factors using SVM as classifier for comparability reasons with other works serving as a baseline.
Table 2 clarifies the choice of the right classifier. For this test also only GTI was chosen.

Table 2: Comparison of classifiers and ensembles, full 276 dim. and optimal feature vector by SVM-SFFS, 10-fold SCV, EMO-DB, GTI

Accuracy [%]	All features	Optimal set
1NN	63.5	75.8
kNN	67.6	78.9
MLP	84.8	86.5
C4.5	61.1	61.5
AdaBoosting C4.5	72.3	74.6
Bagging C4.5	70.7	74.8
MultiBoosting C4.5	72.5	74.6
NB	73.6	74.0
BN	72.1	74.4
SVM	**84.8**	**87.5**
StackingC MLR kNN BN MBC4.5	78.1	83.2
StackingC MLR 1NN NB SVM C4.5	76.2	80.5
Voting 1NN NB SVM C4.5	76.0	79.9
StackingC MLR 1NN NB C4.5	75.4	79.9
Voting 1NN NB C4.5	73.2	78.5

As can be seen in the table SVM lead to the overall best result and are therefore chosen in the ongoing. The table also shows the accuracy for each classifier when the feature set is optimized by SVM-SFFS. Note that all classifiers show better performance with the reduced set, though Decision Trees are less prone to this effect, as they already prune features by information gain.

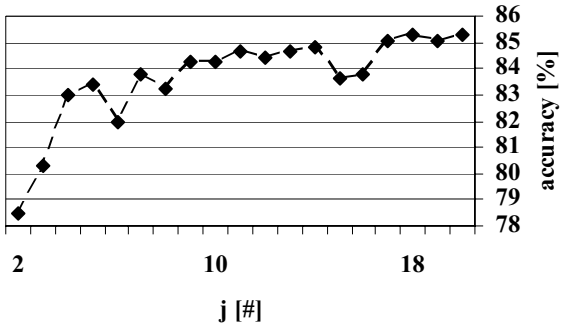

Figure 2: Influence of the cross-validation factor j, all features, j-fold SCV, EMO-DB

As a first variant apart from utterance-level features we consider absolute time intervals (ATI). The database is split into clips firstly, and then classified. Thereby the sample size of 488 is raised through segmentation. Table 3 shows obtained results and the number of clips for the two diverse splitting frame lengths chosen.

Table 3: Absolute Time Intervals, all 276 features, SVM, 10-fold SCV, EMO-DB, ATI

ATI	0.5 sec	1.0 sec
Clips [#]	2,569	1,430
Accuracy [%]	67.2	70.7

As can be seen in the table, emotion recognition is possible already on short parts of an utterance. Yet, recognition accuracy is expectantly lower even though more training material is available.
Next, we want to compare the other two introduced variants, which lead to a super feature vector: GRTI and GATIR. Table 4 shows for each type the chosen number *n* of segments, the accuracy obtained by once the full feature vector of *n* times 276 features and once the optimum number of features after SVM-SFFS reduction and the according accuracy. As reference conventional utterance-level only (GTI) performance is shown here, too.

Table 4: Comparison of segmentation schemes, full 276 dim. and optimal feature vector by SVM-SFFS, 10-fold SCV, EMO-DB

EMO-DB	Segments [#]	All Features [%]([#])	Optimized Set [%]([#])
GTI	1	**84.8** (276)	**87.5** (75)
GRTI	2+1	**86.9** (829)	**93.6** (295)
	3+1	**87.9** (1105)	**96.5** (304)
GATIR	3+1	**84.3** (1105)	**94.6** (293)

The table shows that the optimum result was obtained favoring 3+1 segments compared to 2+1 and taking GRTI rather than

GATIR. Feature reduction leads to an impressive improvement in any case. GTI is clearly outperformed, and thereby the results presented in [9, 10] are likewise outperformed.

It also seems interesting to highlight which segments show most relevance, and how good each segment would perform on its own. These numbers are shown in table 5 for the variants GATIR and GRTI having 3 segments plus the global features. GRTI segments always outperform GATIR ones already without feature reduction. This comes, as GATIR segments are always only 500 msec long and are therefore mostly shorter than GRTI ones. The decrease compared to the equivalently 500 msec long ATI features comes, as in this evaluation only the 488 samples of EMO-DB are available, compared to ATI where the database was split up firstly. In the case of the superior GRTI segments we also provide a per-segment feature-space optimization with according optimum number of features.

Table 5: Relevance of single segments

Segment Index	**1**	**2**	**3**
GATIR 3+1 All features [%]	63.7	66.7	57.9
GRTI 3+1 All features [%]	71.0	75.1	67.5
GRTI 3+1 Optimized set [%]([#])	75.9 (91)	81.9 (76)	80.7 (89)

As can be seen in the table the highest accuracy is observed for the middle part of an utterance in each case. Especially in the case of GRTI with optimized feature sets astonishingly high accuracy can be obtained. Likewise the middle third of an utterance is in terms of accuracy already close to regarding the whole utterance on this database.

7. Conclusions

Within this work we discussed three variants of sub-utterance-level features in speech emotion recognition. Absolute time intervals of 1 sec already led to 70.7% accuracy with respect to emotion recognition with a constant frame rate. However, processing a whole phrase leads to significantly higher accuracy, as one would expect. Astonishingly however, knowledge of the middle third already leads to 81.9% on EMO-DB compared to a maximum of 87.5% if only phrase-level features are used. By construction of a super-vector incorporating features on diverse time-levels overall accuracy could be raised as high as 96.5% which is the highest recognition rate reported on this database, yet. Thereby sub-dividing a phrase in relative parts proved superior to absolute lengths at relative positions. The choice of three segments further more proved superior to having only two segments, which led to 93.6% accuracy on EMO-DB. However, this high performance comes at a price: 305 features need to be extracted opposing 75 for phrase-level only. As future work the comparison to dynamic classification of absolute time intervals with HMM or DBN remains, which may be a new chance for dynamic modeling in speech emotion recognition. Alternatively multi-instance learning may be used to stick to static classifiers popular in speech emotion recognition. Thereby a weighting as the segment length or confidences for each segment could be used. Apart from that more intelligent segment construction shall be investigated as word-, or voicing segments [11]. Finally, we aim at demonstration of the effectiveness of the methods shown in this work on further databases as Danish Emotional Speech Database (DES) or Speech Under Simulated and Actual Stress (SUSAS).

8. Acknowledgements

The content of this paper highly benefits from the works of the student researcher Lorenz Mösenlechner.

9. References

[1] Shriberg, E., 2005. Spontaneous Speech: How People Really Talk And Why Engineers Should Care, *Proc. INTERSPEECH 2005*, ISCA, Lisbon, Portugal, pp. 1781-1784.

[2] Schuller, B.; Ablaßmeier, M.; Müller, R.; Reifinger, S.; Poitschke, T.; Rigoll, G., 2006. Speech Communication and Multimodal Interfaces, in *Advanced Man Machine Interaction*, K.-F. Kraiss (ed.), Springer, Berlin, Heidelberg, ISBN: 3-540-30618-8, pp. 141-190.

[3] Cowie, R.; Douglas-Cowie, E.; Tsapatsoulis, N.; Votsis, G.; Kollias, S.; Fellenz, W.; Taylor, J. G., Jan. 2001. Emotion recognition in human-computer interaction, *IEEE Signal Processing Magazine*, vol. 18, no. 1, pp. 32–80.

[4] Pantic, M; Rothkrantz, L..: Toward an Affect-Sensitive Multimodal Human-Computer Interaction, Sep. 2003. *Proccedings of the IEEE*, Vol. 91, pp. 1370-1390.

[5] Schuller, B.; Rigoll, G.; Lang, M., 2003. Hidden Markov Model-Based Speech Emotion Recognition. *Proc. ICASSP 2003*, Vol. II, Hong Kong, China, pp. 1-4.

[6] Rotaru, M.; Litman, D., 2005. Using Word-Level Pitch Features to Better Predict Student Emotions During Spoken Tutoring Dialogues, *Proc. INTERSPEECH 2005*, ISCA, Lisbon, Portugal.

[7] Batliner, A.; Steidl, S.; Hacker, C.; Nöth, E.; Niemann, H., 2005. Tales of Tuning – Prototyping for Automatic Classification of Emotional User States. *Proc. INTERSPEECH 2005*, ISCA, Lisbon, Portugal.

[8] Burkhardt, F.; Paeschke, A.; Rolfes, M.; Sendlmeier, W.; Weiss, B., 2005. A Database of German Emotional Speech, *Proceedings of the INTERSPEECH 2005,* ISCA, Lisbon, Portugal, 1517-1520.

[9] Schuller, B.; Müller, R.; Lang, M.; Rigoll, G., 2005. Speaker Independent Emotion Recognition by Early Fusion of Acoustic and Linguistic Features within Ensembles. *Proc. INTERSPEECH 2005*, ISCA, Lisbon, Portugal, 805-809.

[10] Vogt, T.; Andre, E., 2005. Comparing feature sets for acted and spontaneous speech in view of automatic emotion recognition, *Proc. ICME 2005*, Amsterdam, Holland.

[11] Shami, M. T.; Kamel, M. S., 2005. Segment-Based Approach to the Recognition of Emotions in Speech, *Proc. ICME 2005*, Amsterdam, Holland.