BRUTE-FORCING HIERARCHICAL FUNCTIONALS FOR PARALINGUISTICS: A WASTE OF FEATURE SPACE?

Björn Schuller[1]*,Matthias Wimmer*[2]*,Lorenz Mösenlechner*[3]*,Christian Kern*[1]*,Dejan Arsic*[1]*,Gerhard Rigoll*[1]

[1]Technische Universität München, Institute for Human-Machine Communication, D-80333, Germany
[2]Waseda University, Perceptual Computing Lab, Tokyo 169-8555, Japan
[3]Technische Universiät München, Intelligent Autonomous Systems Group, D-85748, Germany
schuller@IEEE.org

ABSTRACT

While the ”‘quasi-state-of-the-art”’ towards acoustic emotion recognition relies on multivariate time-series analysis of e.g. pitch, energy, or MFCC by statistical functionals as moments or extrema, only few respect statistical noise by outliers due to too long segments as turns. Such noise can be overcome by hierarchical functionals as means of extrema over smaller units as words or chunks. Segmentation of such units however usually relies on transcription. We therefore discuss hierarchical functionals based on automatic segmentation and their systematic generation as opposed to common expert-driven selection. To cope with rapidly growing feature spaces ¿5k, we discuss data-driven two-stage compression based on SVM-SFFS. Extensive test-runs are carried out on two known emotion and behavior corpora, and show superiority of the suggested approach.

***Index Terms*—** Emotion Recognition, Affect Recognition, Hierarchical Functionals, Feature Brute-Forcing, Feature Selection.

1. INTRODUCTION

The state-of-the-art approach towards acoustic emotion recognition is derivation of statistic functionals as mean, standard deviation, or extrema from a low-level-descriptor such as pitch, energy, or MFCC coefficients [1, 2, 3, 4]. This resembles a strong reduction of information, and allows for generalization in view of independence of spoken content. Mostly, such functionals are thereby derived over a whole turn of speech. However, such descriptive statistical analysis becomes prone to outliers with increasing unit length. To overcome this problem, more recent works often base on hierarchical functionals such as mean of extrema over consecutive smaller units, as words [3, 2], yet mostly relying on word boundaries by transcription. Also, these features are not generated in a systematic way, but rather by expert-knowledge. In this work we therefore aim at answering two questions: can we base hierarchical functionals on less intelligent pre-segmentation compared to word boundaries, more concretely absolute or relative time intervals, which can be carried out robustly? And, is brute-forcing of hierarchical functionals, which easily results in very high initial feature spaces ¿5k reasonable, or, put more straight forward: worth the effort? Such high dimensionality is however not intended for the actual classification. Moreover, we will use two-stage feature space compression to cope with high dimensionality while providing close-to-optimal set optimization at high de-correlation level, yet preserving most relevant features. To answer these questions we will introduce a set of base-level-functionals and features, explain segmentation and hierarchical functionals, and discuss extensive test-runs on two known data-sets in the ongoing.

2. FEATURES AND FUNCTIONALS

For every acoustic signal frame at 100 fps various features in the time as well as in the frequency domain are extracted. These will be denoted as low-level-descriptors (LLD). Selected LLD for this work aim at broad coverage of typical prosodic, spectral and cepstral, as well as voice quality features as found in [3, 4]. At the same time the basis for systematic feature generation shall be kept compact. Table 1 gives an overview of used LLD, herein.

Type	Abbreviation	LLD
Time Signal	T	Elongation, Centroid, ZCR
Energy	E	Log-Frame-Energy
Spectral	S	0-250 Hz, 0-650 Hz, Flux Roll-Off + δ, Centroid + δ
Pitch	P	F0
Formants	F	F1-7 Frequ. + δ, BW. + δ
Cepstral	C	MFCC 1-15 + δ + $\delta\delta$
Voice Quality	V	HNR

Table 1. Overview acoustic LLDs.

Next, functionals of LLD are calculated. Formally, a functional is a mapping of a function space to a number:

$$f:\ F \rightarrow \Re^1 \tag{1}$$

These functionals can then be used for static classification, e.g. by Support Vector Machines (SVM), as they describe the function of the change of features over time. Functionals in this work are typical statistical characteristics of LLD such as mean, median, minimum and maximum position and value, and standard deviation. We preferred simple over complex, following the findings presented in [4]. To better model changes over time of LLD, these functionals are also applied on speed (δ) and acceleration ($\delta\delta$) regression coefficients. In total, a feature set dimension of 622 is obtained by application of selected functionals to the introduced LLD. These will be denoted as base-level-functionals in the ongoing. While this number seems high already, it will be increased by a factor of up to ten in the ongoing. However, the aim is to provide a broad basis for feature selection rather than to actually classify emotions in a running engine by 5̃k features. Likewise, we choose the popular Sequential Floating Forward Search (SFFS) [5, 1] for feature space de-correlation at highest accuracy levels. SVM are used as wrapper, denoted as SVM-SFFS. Note that spaces will always be compressed to maximum accuracy by SVM in the ongoing. We use polynomial kernels and pair-wise multiclass discrimination.

3. SEGMENTATION OF AUDIO FILES

The performance of emotion classification highly depends on the information content of the functionals extracted by the feature extractor. As one functional can give only a very raw description over time, the idea is to divide the whole audio signal at analysis into several parts and extract base-level functionals for every part. After extraction all resulting functionals for every part of the audio file are joined together to form one combined feature vector - that is super vector. To keep segmentation schemes as simply computable as possible with respect to real-time processing, effectiveness and robustness, no dynamic segment boundaries (prone to error themselves) like words are used. Note however that e.g. voiced segments or pause/non-pause are almost equally derived, yet not considered, herein. In addition to the whole audio signal which can be seen as one segment and will be denoted as global time interval (GTI), two further segmentation schemes were investigated:

- Absolute Time Intervals (ATI)
- Relative Time Intervals (RTI)

These segmentation schemes are discussed in this section.

3.1. Absolute Time Intervals

In this segmentation scheme the audio stream is split up into frames with fixed size, e.g. 500 msec. In general this allows for incremental emotion recognition or easy fusion with other modalities as video, that operate on fixed frame rates (cf. [6]). Depending on the length of audio samples at analysis, accordingly a variable number of segments is generated. On its own, this demands special requirements to the classification algorithm as time warping by e.g. Hidden-Markov-Models or Dynamic Bayesian Networks, or Multi-Instance Learning. Figure 1 visualizes this scheme for two symbolic audio files with different length.

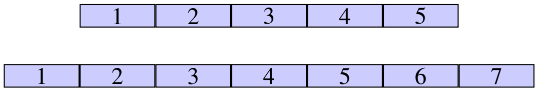

Fig. 1. Schematic view of Absolute Time Intervals (ATI) segmentation.

3.2. Relative Time Intervals

In this segmentation approach, the audio signal is divided into a fixed number of segments, as halves, thirds, etc. Therefore, the number of extracted functionals is constant for every analyzed audio file and all classification algorithms for a static number of input vectors can be used as more common in the field of acoustic emotion recognition. As ATI, this segmentation scheme leads to a better modeling of the function described by the functionals. Again, functionals for every segment and for the whole audio file can be composed into a super-vector. Figure 2 accordingly visualizes the segments, the functionals are calculated for. The number of segments is therein set to three and segmentation for two audio files with different lengths is shown.

Note that also mixed forms as Absolute Time Intervals at Relative Positions are an option, as discussed in [6]. However these will not be followed, herein.

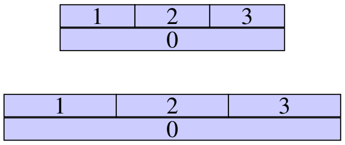

Fig. 2. Schematic view of Relative Time Intervals (RTI) segmentation.

3.3. Hierarchical Functionals

Apart from adding sub-turn entity, that is segmental features, to a super-vector as described before and shown in [6], we now calculate functionals of functionals. Looking at the effect of e.g. segmentation according to the RTI scheme, the change of functionals is represented only by the different values of functionals for each segment. The description of the function of a functional by its statistic properties as described further increases the feature information, and allows for static modeling in the case of ATI: the problem of the ATI scheme was that the resulting feature vectors had variable lengths related to the lengths of the whole audio sample. Calculation of functionals of the function of segment's functionals provides an elegant solution here: by dumping the values of functionals for each segment and keeping only the functionals calculated for their change one static vector is obtained.

In detail, first a segmentation as described by the ATI scheme or the RTI scheme is performed. Then a set of default functionals are calculated for every functional of the segments: in order to keep the feature space limited considering systematic hierarchical functional brute-forcing and the original space size of 0.6k, we decided for a limited set of only 7 hierarchical functionals. Again, we preferred simple over complex, as first two statistical moments over higher ones, following the findings in [4]: mean, median, standard deviation, position and value of extrema are used.

4. DATABASES

4.1. Berlin Emotional Speech

The Berlin Emotional Speech Database [7] (known as EMO-DB) is an audio only database of German emotional utterances spoken by 10 professional actors (5 female). The recordings took place in the anechoic chamber of the Technische Universität Berlin. For each of 7 emotions as seen in Table 2 10 sentences of emotionally neutral content were spoken be each speaker. Overall 896 datasets were recorded. The database was independently annotated by 20 people with respect to naturalness and assignability. For our experiments only the datasets with ¿60% of the annotators agreeing upon naturalness and ¿80% upon assignability to an emotion where chosen in accordance to other works. This final class distribution is shown in Table 2.

	anger	boredom	disgust	fear	happiness	neutral	sadness	TOTAL
[#]	127	79	38	55	58	78	53	**488**

Table 2. Distribution of emotions, database EMO-DB.

4.2. Airplane Behavior Corpus

The Airplane Behavior Corpus (ABC), introduced in [8], is a database crafted for the special target application of public transport surveillance, consisting of elicited behavior. There is a broad discussion in the community with respect to acted vs. spontaneous data, which we will not address herein. However, it is believed, that mood induction procedures favor realism in behavior. Therefore a script was used, which lead subjects through a guided storyline. 8 subjects in gender-balance from 25 a to 48 a (mean 32 a) took part in the recording. The language throughout recording is also German, and a total of 11.5h video was recorded and annotated independently after pre-segmentation by three experienced male labelers within a closed set as seen in Table 3. This table also shows the final distribution of samples with total inter-labeler-agreement. The average length of the 396 clips in total is 8.4s.

	aggressive	cheerful	intoxicated	nervous	neutral	tired	TOTAL
[#]	87	100	31	70	68	40	**396**

Table 3. Distribution of behaviors, database ABC.

5. RESULTS

In the research of behavior or emotion recognition data is usually sparse. As most popular evaluation strategy j-fold stratified cross validation (SCV) can therefore be named: SCV allows for testing and disjunctive training on the whole corpus available. We therefore use 10-fold SCV in the ongoing and present mean accuracies throughout cross-folds.

In our first experiment with hierarchical functionals, features were extracted in the common way for the whole sequence (GTI), as well as for its thirds according to the RTI segmentation scheme. Then, the default hierarchical functionals named in sect. 2 were calculated for every base-level-functionals of the three parts. With the 622 functionals for the whole sequence, the $3 \cdot 622$ functionals of each part and the $7 \cdot 622$ hierarchical functionals, the resulting combined feature vector has a dimensionality of 6842. Considering high computational demand of the very effective SVM-SFFS, dimensionality had to be pre-reduced. Therefore, the SFFS algorithm was first applied to the original global 622 functionals on the first layer. Then, functionals for the three parts and hierarchical functionals for the 200 best features were taken out and recombined to form a new data vector, still containing the 200 functionals of the whole sequence, $3 \cdot 200$ functionals for every part and $7 \cdot 200$ functionals for the hierarchical functionals. Thus, it had an overall dimensionality of 2200. Then SFFS was repeatedly applied on the hierarchical layer.

Next, the performance of the hierarchical functionals according to the ATI segmentation scheme is considered. Again, the base-level functionals introduced before were extracted for every audio file. Further, every audio sequence was divided into sub-segments all having a length of 0.5 sec. All 622 functionals were extracted for every sub-segment and functionals of these functionals were calculated. Therefore, the resulting data vector is consisting of the 622 functionals of the whole sequence and the $7 \cdot 622$ hierarchical functionals. Thus, the overall dimensionality is 4976.

Table 4 shows the classification results with 10-fold SCV for the EMO-DB.

Scheme	Functionals	All Features	Optimized Set
GTI	1	**79.7%** (622)	**90.4%** (154)
RTI	3RTI+GTI+HF	**78.5%** (6842)	**92.6%** (480)
ATI	GTI+HF	**80.1%** (4976)	**92.4%** (795)

Table 4. Results for hierarchical functionals (HF) with RTI and ATI segmentation on EMO-DB, SVM, 10-fold SCV. Feature numbers [#] are provided in ().

Interestingly, the RTI and the ATI segmentation scheme with hierarchical functionals are performing nearly equally although the RTI feature vector contains functionals for every of the three frames in addition to the functionals of the whole sequence and the hierarchic functionals. By contrast, the ATI feature vector contains only the functionals of the whole feature vector and the hierarchical functionals. This may indicate that most of the important information for each sub-sequence can be described by hierarchical functionals.

Table 5 shows classification accuracies for different numbers of features. As described, the according number indicates the amount of best performing features given by the SFFS algorithm. At the beginning, a relatively strong increase can be seen while the accuracy keeps nearly constant for higher feature dimensions. But the ATI scheme seems to contain more information in the first 200 features than the RTI scheme. The accuracy difference is about 2%. For higher feature dimensions, the results are nearly equal. The stability of classification rates for higher feature dimension indicates a saturation of feature information.

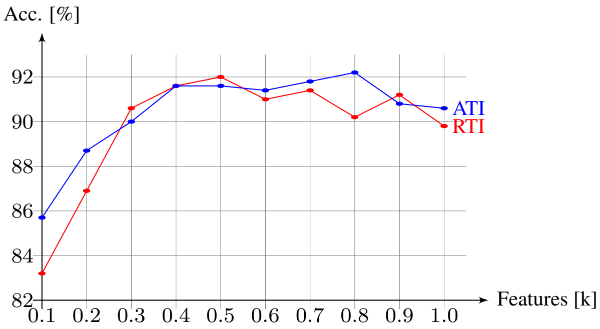

Table 5. Accuracy over space dimension by SVM-SFFS for RTI and ATI and hierarchical functionals. EMO-DB, SVM, 10-fold SCV.

Our next results base on the ABC database in accordance to the settings for EMO-DB. Note that the average sequence length is rather long for the ABC database. Thus, finer temporal modeling by hierarchical functionals of segments according to the ATI segmentation may be beneficial compared to the RTI scheme: with the RTI segmentation, a fixed number of segments with a dynamic length was created. On databases with short sequences an almost equal number of segments will be created with the ATI and the RTI segmentation scheme with a low number of segments. Hence, the results for ATI and RTI are nearly similar for short sequences. But for longer sequences, one segment of the RTI scheme contains noticeably more information and the common statistical description by functionals becomes inexact. In contrast, the modeling of the sequence of functionals with a significantly higher number of segments should lead to a considerable increase of classification accuracy due to better smoothing of outliers. For the ABC database, the ATI segmentation scheme with a segment length of 0.5 sec leads to an average number of 17 segments per audio sample.

Scheme	Functionals	All Features	Optimized Set
GTI	1	**53.9%**	**71.1%**
		(622)	(190)
RTI	3RTI+GTI+HF	**55.2%**	**75.4%**
		(6842)	(290)
ATI	GTI+HF	**73.4%**	**80.0%**
		(4976)	(277)

Table 6. Results for hierarchical functionals (HF) with RTI and ATI segmentation on the ABC database. Feature numbers [#] are provided in ().

As can be seen in Table 6, the ATI segmentation scheme is clearly outperforming the other segmentation schemes. Interestingly, even for the unoptimized set of features it shows an impressively higher classification accuracy. The increase in comparison with GTI is about 10% absolute. By having a closer look at the features selected by the SFFS algorithm, it can be seen that 81% of the functionals used in the optimal set are hierarchical functionals. Figure 3 shows the contribution of the different types of features to the optimized feature set. Note that the high percentage of for example MFCC features likely results from the high number of derived functionals of this type. More interestingly, Figures 7 and 7 reveal the percentage of the three different kinds of functionals. For the RTI scheme these are global, the subsequences', and hierarchical functionals. It can be seen that hierarchical functionals are mostly superior to the other kinds of functionals in terms of presence after selection. This can also be seen in Figure 8, based on the optimized set for the ATI segmentation scheme. Due to the lack of subsequence functionals, the dominance of hierarchical functionals is even stronger than for the RTI segmentation scheme, here. Interestingly, most feature types show similar behavior with respect to distribution among global and hierarchical features. The segmental features present in RTI seem to be equally "'absorbed"' in the ATI case.

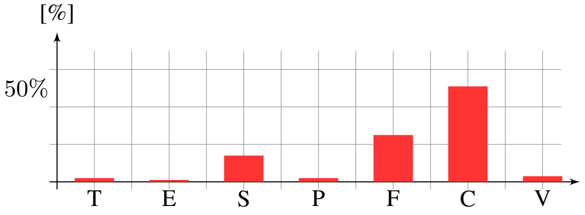

Fig. 3. Contribution of feature types to the optimized feature set. For abbreviations see Table 1.

Rel. Frequ. [%]	T	E	S	P	F	C	V
global	14	50	5	20	14	17	8
segmental	0	0	24	60	22	15	46
hierarchical	86	50	71	20	64	68	46

Table 7. Comparison of the contribution of global, segmental, and hierarchical functionals to the optimized set by SVM-SFFS by feature group. ABC database, RTI-segmentation scheme.

Rel. Frequ. [%]	T	E	S	P	F	C	V
global	14	66	18	0	18	20	0
hierarchical	86	33	82	100	82	80	100

Table 8. Comparison of the contribution of global, segmental, and hierarchical functionals to the optimized set by SVM-SFFS by feature group. ABC database, ATI-segmentation scheme.

6. CONCLUSION AND OUTLOOK

Within this paper we presented brute-forcing of hierarchical functionals for acoustic emotion recognition based on absolute and relative time intervals for simple but fast and robust pre-segmentation. Thereby absolute time intervals pre-dominate in the case of longer segments as turns. Two-stage compression of the generated large spaces by SVM-SFFS was shown to highly boost accuracies. Likewise even such straight-forward segmentation helps to improve over turn-segmentation. As space optimization has to be carried out only once prior to recognition in an application setting, the second initial question also clearly has to be answered positively: the (low) extra effort of hierarchical functional brute-forcing seems worth the effort.

Future research will deal with other databases such as AIBO [3, 4] and hierarchical functionals based on ASR-based word-boundary detection in comparison to the strategies discussed, herein.

7. REFERENCES

[1] D. Ververidis and C. Kotropoulos, "Emotional Speech Classification Using Gaussian Mixture Models and the Sequential Floating Forward Selection Algorithm," in *Proc. Multimedia and Expo*, Amsterdam, 2005, pp. 1500–1503.

[2] T. Vogt and E. Andre, "Comparing Feature Sets for Acted and Spontaneous Speech in View of Automatic Emotion Recognition," in *Proc. Multimedia and Expo*, Amsterdam, 2005, pp. 474–477.

[3] A. Batliner, S. Steidl, B. Schuller, D. Seppi, K. Laskowski, T. Vogt, L. Devillers, L. Vidrascu, N. Amir, L. Kessous, and V. Aharonson, "Combining Efforts for Improving Automatic Classification of Emotional User States," in *Proceedings of IS-LTC 2006*, Ljubliana, 2006, pp. 240–245.

[4] B. Schuller, A. Batliner, D. Seppi, S. Steidl, T. Vogt, J. Wagner, L. Devillers, L. Vidrascu, N. Amir, L. Kessous, and V. Aharonson, "The Relevance of Feature Type for the Automatic Classification of Emotional User States: Low Level Descriptors and Functionals," in *Proc. INTERSPEECH 2007*, Antwerp, Belgium, 2007, pp. 2253–2256.

[5] P. Pudil, J. Novovicova, and J. Kittler, "Floating search methods in feature selection," *Pattern Recognition Letters*, vol. 15, pp. 1119–1125, 1994.

[6] B. Schuller and G. Rigoll, "Timing Levels in Segment-Based Speech Emotion Recognition," in *Proc. INTERSPEECH 2006*, Pitsburgh, PA, 2006, pp. 1818–1821.

[7] F. Burkhardt, A. Paeschke, M. Rolfes, W. Sendlmeier, and B. Weiss, "A database of german emotional speech," in *Proceedings of the 9th Eurospeech - Interspeech 2005*, Lisbon, Portugal, 2005, ISCA, pp. 1517–1520.

[8] B. Schuller, M. Wimmer, D. Arsic, D. Rigoll, and B. Radig, "Audiovisual Behavior Modeling by Combined Feature Spaces," in *Proc. ICASSP 2007*, Honolulu, Hawaii, 2007, vol. II, pp. 733–736.