The Relevance of Feature Type for the Automatic Classification of Emotional User States: Low Level Descriptors and Functionals

Björn Schuller[1], Anton Batliner[2], Dino Seppi[3], Stefan Steidl[2], Thurid Vogt[4], Johannes Wagner[4], Laurence Devillers[5], Laurence Vidrascu[5], Noam Amir[6], Loic Kessous[6], Vered Aharonson[7]

[1]**TUM:** Institute for Human-Machine Communication, Technische Universität München, Germany
[2]**FAU:** Lehrstuhl für Mustererkennung, Friedrich-Alexander-Universität Erlangen, Germany
[3]**FBK:** Fondazione Bruno Kessler - irst, Trento, Italy
[4]**UA:** Multimedia Concepts and their Applications, University of Augsburg, Germany
[5]**LIMSI:** Spoken Language Processing Group, LIMSI-CNRS, Orsay Cedex, France
[6]**TAU:** Dep. of Communication Disorders, Sackler Faculty of Medicine, Tel Aviv University, Israel
[7]**AFEKA:** Tel Aviv academic college of engineering, Tel Aviv, Israel
batliner@informatik.uni-erlangen.de

10.21437/Interspeech.2007-612

Abstract

In this paper, we report on classification results for emotional user states (4 classes, German database of children interacting with a pet robot). Six sites computed acoustic and linguistic features independently from each other, following in part different strategies. A total of 4244 features were pooled together and grouped into 12 low level descriptor types and 6 functional types. For each of these groups, classification results using Support Vector Machines and Random Forests are reported for the full set of features, and for 150 features each with the highest individual Information Gain Ratio. The performance for the different groups varies mostly between $\approx 50\%$ and $\approx 60\%$.

Index Terms: emotional user states, automatic classification, feature types, functionals

1. Introduction

The study of 'Speech and Emotion' during the recent years can be characterized by three trends: (1) the trend towards more natural, real-life data, (2) the trend towards taking into account not only some 'prototypical' emotions but emotional, affective states in a broader sense, and (3) the trend towards a thorough exploitation of the feature space, resulting in hundreds or even thousands of features used for classification. The database described in section 2 has been recorded and processed in this vein. First results reported in [2] showed that pooling together features extracted at different sites indeed improved classification performance; however, a systematic investigation examining the contributions of different types of features has not yet been carried out.

The 'holy grail' of automatic classification is to find 'the' optimal set of features, consisting of the most important independent features. The difficulty of this task, due to factors such as the huge number of possible features that can be extracted from speech signals, and due to the computationally demanding methods for classifying highly dimensional features spaces, would have required feature space de-correlation and reduction, e.g. through transformations like Principal Component Analysis (PCA). However, in this paper we did not follow this approach because we would not have found the answer to the aforementioned question: which *types* of features contribute to which extent to classification performance, and therefore to modelling the phenomenon we are interested in. Neither did we opt for comparing selection and classification results obtained at each site separately; instead we tried to unify as many factors as possible, such as feature selection and classification itself to enable a more reliable comparison between feature types. We approached these goals by pooling together feature vectors computed at different sites. The various sites are rooted in different traditions, focussing on acoustics only or on a combination of acoustics and linguistics; some sites followed a 'brute-force' method of exploiting the feature space, other sites computed features in a knowledge-based way. Of course, this has not been done in a pure form; thus some hybrid strategies were used as well. By pooling together all these features we at least come closer to modelling diversity. In this paper, our intention was to concentrate on dealing separately with feature types (Low Level Descriptors LLDs and functionals), and have a closer look at their respective impact on classification performance.

2. Material and Annotation

The database used is a German corpus with recordings of children communicating with Sony's AIBO pet robot; it is described in more detail in [2] and other papers quoted therein. The children were led to believe that the AIBO is responding to his or her commands, but the robot is actually being controlled by a human operator who causes 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 from 51 children (age 10 - 13, 21 male, 30 female; about 9.2 hours of speech without pauses). The recordings were segmented automatically into 'turns' using a pause threshold of 1500 msec. 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. If three or more labelers agreed, the label was attributed to the word (majority voting MV); in parentheses, the number of cases with MV is given: *joyful* (101), *surprised* (0), *emphatic* (2528), *helpless* (3), *touchy*, i.e., irritated (225), *angry* (84), *motherese* (1260), *bored* (11), *reprimanding* (310), *rest*,

i.e. non-neutral, but not belonging to the other categories (3), *neutral* (39169). 4707 words had no MV; all in all, there were 48401 words. Some of the labels are very sparse. Therefore, *neutral* and *emphatic* were down-sampled, and *touchy* and *reprimanding*, together with *angry*, were mapped onto **A***ngry* as representing different but closely related kinds of negative attitude. This more balanced 4-class problem consists of 1557 words for **A***ngry* (**A**), 1224 words for **M***otherese* (**M**), 1645 words for **E***mphatic* (**E**), and 1645 for **N***eutral* (**N**). As semantically meaningful chunks can probably be better mapped onto emotional units than turns containing up to > 50 words, we clustered words into chunks. Eventually, these chunks of words were labelled by mapping word labels onto chunks by MV. We performed a coarse syntactic labelling manually with the following chunk triggering boundaries: at main clauses, free phrases, and between adjacent /Aibo/ instances because repetitions of vocatives make emotional colouring more likely. Spontaneous speech, especially in such scenarios as 'giving commands to a pet (robot)', are quite often not well-formed syntactically: no clear structural indication is found, let alone interpunction. We therefore used a prosodic criterion in addition: if the pause between words is $\geq$ 500 msec, we assume a chunk boundary. The length of the pauses between words was obtained from the manually corrected word segmentation. The mapping of word-based onto chunk-based labels followed the basic strategy described in [2]. This procedure yielded 4543 chunks (914 **A***ngry*, 586 **M***otherese*, 1045 **E***mphatic*, and 1998 **N***eutral*) with an average length of 2.9 words per chunk. Inter-labeller correspondence is dealt with in [7].

3. Features: Extraction and Grouping

The following arrangement into 'knowledge-based' vs. 'brute-force' has to be taken with a grain of salt; it rather describes the starting point and the basic approach. FAU for instance uses a knowledge-based approach for the computation of word-based features and then a 'blind', 'brute-force' approach for the subsequent computation of chunk-based features. The six part-of-speech (POS) classes used by some of the sites were AUX (auxiliaries), PAJ (particles, articles, and interjections), VERB (verbs), APN (adjectives and participles, not inflected), API (adjectives and participles, inflected), and NOUN (nouns, proper nouns), annotated for the spoken word chain. Some sites used Praat [3], the other own procedures for feature extraction.

3.1. 'knowledge-based' computation, sequential: chunks, based on word statistics using correct segmentation

FAU: 92 *acoustic features:* word-based computation of pauses, energy, duration, and F0; for energy: maximum (max), minimum (min), mean, absolute value, normalized value, and regression curve coefficients with mean square error; for duration: absolute and normalized; for F0: min, max, mean, and regression curve coefficients with mean square error, position on the time axis for F0 onset, F0 offset, and F0 max; for jitter and shimmer: mean and variance; normalization for energy and duration based on speaker-independent mean phone values; for all these word-based features, min, max, and mean chunk values computed based on all words in the chunk. 24 *linguistic features* (# of classes per chunk and normalized as for # of words in chunk): POS features; higher semantic features: vocative, positive valence, negative valence, commands and directions, interjections, and rest.
FBK: 26 *acoustic features:* similar to FAU but no F0 onset and offset values, no jitter/shimmer; normalization of duration and energy done on the training set without backing off to phones but using information on the number of syllables in addition; 6 *linguistic features:* POS features.

3.2. 'knowledge-based' computation for chunks

LIMSI: 90 *acoustic features:* min, max, median, mean, quartiles, range, standard deviation for F0; the regression curve coefficients in the voiced segments, its slope and its mean square error; calculations of energy and of the first 3 formants and their bandwidth; duration features (speaking rate, ratio of the voiced and unvoiced parts); voice quality (jitter, shimmer, Noise-to-Harmonics Ratio (NHR), Harmonics-to-Noise Ratio (HNR), etc.). 13 *linguistic features:* POS, nonverbals and disfluencies.
TAU: 222 *acoustic features:* Five families of features: pitch based, duration based, intensity based, spectral, and voice quality based; different levels of functionals applied to the raw contours: from basic statistics to curve fitting methods to methods based on perceptual criteria. Several duration features computed on the lengths of voiced segments and pauses, and spectral features based on Mel Frequency Cepstral Coefficients (MFCC) and Long Term Average Spectrum (LTAS).

3.3. 'brute force' computations for chunks

UA: 1586 *acoustic features:* pitch, energy, 12 MFCCs, 10 cepstral coefficients based on wavelet transformation, HNR and short-term spectra, as well as different views on the time series such as considering only local max or min, or distances, magnitudes and steepness between adjacent extrema. From each of these series of values, mean, max, min, range, median, first quartile, third quartile, interquartile range, and variance. Chunk length added to the vector as a durational feature. The proportion of voiced to unvoiced frames, several normalised and positional features of pitch and energy.
TUM: 1718 *acoustic features:* a systematic generation by acoustic LLD extraction, filtering, derivation, and application of functionals on the chunk level. As LLDs pitch, HNR, jitter, shimmer, energy, MFCCs 1-16, formants 1-7 with amplitude, position, and bandwidth, and a selection of spectral features; derived LLDs comprising derivatives and crossed LLDs; functionals covering the first four moments, extremes, quartiles, ranges, zero-crossings, roll-off, and higher level analysis. 489 *linguistic features:* frequencies of bag of words using the manual transliteration of the spoken word chain, POS, non-verbals, and disfluencies.

3.4. Grouping into Low Level Descriptor Types and Functionals

In the following grouping, we shortly describe the breakdown into types of LLDs on the one hand, and types of functionals on the other hand. As for LLDs, we concentrate on a characterisation in phonetic and linguistic terms (*what* has been extracted); as for functionals, we concentrate on the way *how* these features have been extracted:
voice quality: jitter/shimmer and other measures of microprosody, NHR, HNR and autocorrelation. They are based in part on pitch and intensity but reflect voice quality such as breathiness or harshness.
F0: This is the acoustic equivalent to the perceptual unit pitch; it is measured in Hz and often made perceptually more adequate by logarithmic transformation etc. Intervals, characterising points, or contours are being modelled.

spectral and formants: Formants (i.e. spectral maxima) are known to model spoken content, especially lower ones. Higher ones however also represent speaker characteristics. Each one is fully represented by position, amplitude and bandwidth. As further spectral features band-energies, roll-off, centroid or flux are used. Long term average spectrum over a chunk averages out formant information, giving general spectral trends.
cepstrum: MFCC features — as homomorphic transform with equidistant band-pass-filters on the Mel-scale — tend to strongly depend on the spoken content. Yet, they have been proven beneficial in practically any speech processing task. They emphasise changes or periodicity in the spectrum, while being relatively robust against noise.
wavelets: Wavelets give a short-term multi-resolution analysis of time, energy and frequencies in a speech signal. Compared to similar parametric representations such as MFCCs, they are superior in the modeling of temporal aspects.
energy: These features model intensity, based on the amplitude in different intervals, with implicit or explicit normalisation. They can model intervals or characterising points.
duration: These features model temporal aspects; normally the basic unit is milliseconds for the 'raw' values. Different types of normalization are applied. Positions of prominent energy or F0 values on the time axis are attributed to this type as well.
non-verbals, disfluencies: Such as laughter or breathing, and filled pauses or hesitations.
part of speech (POS): A coarse taxonomy of main word classes based on the spoken word chain.
higher semantics: A coarse taxonomy of (partly scenario-specific) most relevant words, word classes, and emotional valence (negative vs. positive), based on the spoken word chain.
bag of words: They are well known from document retrieval tasks [4], and have shown good results for emotion recognition as well [2]. Each term within a vocabulary is represented by an individual feature that represents the term's (logarithmical and normalized) frequency within the current phrase. Terms are thereby clustered with Iterated Lovins Stemming [5].

In the following breakdown of the functionals, we mostly provide figures for the sub-sets as well:
percentiles: quartiles 1/2/3 (245/259/245), quartile ranges lower/upper/total (74/74/212) and other percentiles (87).
specific functions (distributional, spectral, regressional): a blend of several more 'unusal' functionals: several complex statistical functionals (95), micro variation (2), number of segments/intervals/reversal points (6/2/2), ratio (8), error (3), linear/quadratic regression coefficients (10/15), and DCT coefficients 1-5 (2 each).
extremes: min/max by value (283/338), min/max position (107/110), range (285), and min/max of slope (2/5), as well as on-/off-position (1/1).
higher statistical moments: standard deviance (164), variance (137), skewness (79), kurtosis (79), length (12), and zero-crossing-rate (76).
means: first moment by arithmetic mean (353) and centroid (74).
sequential and combinatorial: functionals of any type under the premise that a minimum of two functionals has been applied in either a sequential way (e.g. mean of max) or combinational way (e.g. ratio of mean of two different LLD).

4. Classification of feature types

The data was partitioned into three balanced splits meeting the following requirements (in order of priority): no splitting of within-subject chunks, similar distribution of labels, balance between the two schools, and balance between genders. For the training set, we upsampled all classes but **N***eutral*: 3x **M***otherese*, 2x **E***mphatic*, and 2x **A***ngry*. We computed a 3-fold cross-validation with support-vector-machines SVM (linear Kernel, one-against-one multiclass discrimination, Sequential Minimal Optimization SMO) and Random Forests RF from [8]. Results are given in Table 1 where we report the F value which is used in the interest of having a unique performance measure; here, F is defined as the uniformly weighted harmonic mean of RR and CL: $2 \cdot CL \cdot RR/(CL + RR)$. RR is the overall recognition rate (number of correctly classified cases divided by total number of cases or weighted average); CL is the 'class-wise' computed recognition rate, i.e. the mean along the diagonal of the confusion matrix in percent, or unweighted average. The F measures for SVMs and RFs represent a trade-off between CL and RR. Results are reported for the full set of features in each sub-group, and for the best 150 features per group using Information Gain Ratio (IGR) selection. Note that we did not optimize the whole sets but choose features with individual high IGR. This is not the best feature selection but it is fair across the sub-groups because the number of features is very unequal, cf. 1699 cepstral vs. 153 voice quality features. For the reduced set, groups with less than 150 features were not taken into account. Due to the problem of repeated measurements [2], we refrain from interpreting differences in terms of significance. Note that some 22 features computed at different sites could not be attributed to the LLDs; in addition some 40 features used for the types could not be attributed to the functionals.

5. Discussion and concluding remarks

Now we shortly address the most important results:

feature selection, classifiers, and full vs. reduced set: Our setting displays very high dimensionality (thousands of attributes) and a comparatively small number of patterns (4543). Moreover, many features are extracted by different sites adopting similar strategies and algorithms. These are the well-known problems *curse of dimensionality* and *feature correlation* which both can trouble a reliable classification process. Therefore we present two groups of results, one with the complete feature sets, and another with IGR reduced sets. IGR allows performance gain for SVMs for highly dimensional groups (such as cepstral and spectral), although it is probably not the best approach for dealing with correlated features as it maximizes relevance rather than minimizes redundancy. Therefore, we should refrain from interpreting SVM results without feature reduction. On the other hand, RFs are almost insensible to feature reduction as they basically work on many small, random sub-samples of the features' domain. The two classifiers yield comparable trends for the individual groups.

acoustic features: we can tell apart three groups, most relevant being duration and energy, of medium relevance all other types except voice quality which is least relevant. Grosso modo, this is in accordance with the literature. All acoustic features together obtain better results than single groups, revealing that, in principle, no group should be left out.

linguistic features: most relevant is the full set of words (bag-of-words) but higher semantics and even the very coarse POS modelling are competitive. It might turn out that POS are most robust if it comes to real ASR processing. If ASR is robust, linguistic features will obviously be a good choice. The linguistic features are better exploited by SVMs than by RFs. Differently from acoustics, all linguistic features do not perform

better than the outstanding bag-of-words set. This probably depends on the fact that other groups (such as POS) are simply looser 'quantizations' of the vocabulary.

functionals: Classification of functionals must be carried out with caution: differences might depend for example on the underlying speech contours. Therefore we decided to keep linguistic features aside. Means, higher statistical moments, and esp. sequential+combinatorial functionals are most relevant; means might not be that prone to outliers as extremes/percentiles are; higher statistical moments might be good at approximating curve shapes; a sequential processing might better model the contribution of both smaller (words) and higher (chunks) units, and specific prosodic structuring.

In this paper, we confined classification to feature types, and for feature reduction, to individual IGR. This made the comparison across types reasonably fair. Further, this approach allowed to consider features that display interesting behaviour in conjunction with other, but not alone. The next steps to be taken are to find out most important features — both for individual types and for the whole set of features — and to have a look at specific combinations of types: for instance, we could imagine an added value if we combine the temporal aspects encoded in wavelets with the spectral aspects encoded in MFCCs. The caveat has to be made that there is no strict balance as for functionals and LLDs; for instance, the more powerful sequential functionals were only computed for F0, energy, and duration but not for other types of LLDs. Moreover, acoustic features can 'hide' linguistic information; a simple example is overall length which is of course highly correlated with number of words. Thus we should not take our results as final proof but as indication and guidelines for future research. Then we will find out whether any LLD profits from sequential modelling; it might as well be the case that this holds only for the 'traditional' prosodic feature types, modelling types of rhythmicality, but not for the other LLDs.

The overall performance reported is within the expected range; an adequate feature selection and reduction for the whole set will most likely result in higher performance. Better results could as well easily be produced by processing only the more selective sub-set of prototypes with a higher MV, cf. [1]. As we were interested in the relative importance of feature types, we extracted values based on manual word and chunk segmentation and on the spoken word chain. One of the next steps to be taken is thus classification based on automatic segmentation and word recognition output. If the figures from [6] can be reproduced, this will not result in a marked deterioration of performance.

Table 1: *Classification results: number of features #, F values for full (#) and for reduced (150, by IGR) set.*

feature set		**full**		**reduced**	
type	#	F_{SVM}	F_{RF}	F_{SVM}	F_{RF}
Low Level Descriptors					
voice quality	153	51.5	51.1	51.6	50.8
F0	333	56.1	56.6	55.1	55.1
spectral/formants	656	54.4	57.1	56.0	56.6
cepstral	1699	52.7	55.7	57.1	56.3
wavelets	216	56.0	56.5	56.3	56.7
energy	265	58.5	59.3	60.0	60.0
duration	391	55.1	60.1	60.6	59.8
all acoustic	3713	57.7	62.5	61.2	60.9
disfluencies	4	26.8	25.2	–	–
non-verbals	8	24.8	24.2	–	–
part of speech	31	54.7	54.1	–	–
higher semantics	12	57.6	57.7	–	–
bag of words	476	62.6	60.2	62.3	58.6
all linguistic	531	62.6	60.2	61.7	59.0
all	4244	61.0	64.0	63.1	61.7
functionals (without linguistic features)					
percentiles	1196	53.8	55.5	56.6	54.1
specific	153	54.5	57.0	54.3	56.6
extremes	1132	53.4	57.1	57.0	57.1
higher stat. mom.s	547	57.6	58.6	58.9	59.0
means	427	59.8	59.8	61.3	60.4
sequential+comb.	218	61.2	61.2	60.5	61.6
all functional	3673	57.4	62.3	61.2	60.8

6. Acknowledgements

The initiative to co-operate was taken within the European Network of Excellence HUMAINE under the name CEICES (**C**ombining **E**fforts for **I**mproving automatic **C**lassification of **E**motional user **S**tates). This work was partly funded by the EU in the projects PF-STAR under grant IST-2001-37599 and HUMAINE under grant IST-2002-50742. The responsibilty lies with the authors.

7. References

[1] A. Batliner, S. Steidl, C. Hacker, E. N¨oth, and H. Niemann. Tales of Tuning – Prototyping for Automatic Classification of Emotional User States. In *Proc. 9th Eurospeech - Interspeech 2005*, pages 489–492, Lisbon, 2005.

[2] 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*, pages 240–245, Ljubliana, 2006.

[3] P. Boersma. Praat, a system for doing phonetics by computer. *Glot International*, 5:341–345, 2001.

[4] T. Joachims. Text categorization with support vector machines: learning with many relevant features. In C. Nédellec and C. Rouveirol, editors, *Proceedings of ECML-98, 10th European Conference on Machine Learning*, number 1398, pages 137–142, Chemnitz, DE, 1998. Springer Verlag, Heidelberg, DE.

[5] J. B. Lovins. Development of a stemming algorithm. *Mechanical Translation and Computational Linguistics*, 11:22–31, 1968.

[6] B. Schuller, D. Seppi, A. Batliner, A. Meier, and S. Steidl. Towards more Reality in the Recognition of Emotional Speech. In *Proc. of ICASSP 2007*, pages 941–944, Honolulu, 2007.

[7] S. Steidl, M. Levit, A. Batliner, E. N¨oth, and H. Niemann. "Of All Things the Measure is Man": Automatic Classification of Emotions and Inter-Labeler Consistency. In *Proc. of ICASSP 2005*, pages 317–320, Philadelphia, 2005.

[8] I. H. Witten and E. Frank. *Data mining: Practical machine learning tools and techniques, 2nd Edition*. Morgan Kaufmann, San Francisco, 2005.