Patterns, Prototypes, Performance: Classifying Emotional User States

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

[1] Fondazione Bruno Kessler – irst, Trento, Italy
[2] Lehrstuhl für Mustererkennung, Universität Erlangen-Nürnberg, Germany
[3] Institute for Human-Machine Communication, Technische Universität München, Germany
[4] Multimedia Concepts and their Applications, University of Augsburg, Germany
[5] LIMSI-CNRS, Spoken Language Processing Group, Orsay Cedex, France
[6] Dep. of Communication Disorders, Sackler Faculty of Medicine, Tel Aviv University, Israel
[7] Tel Aviv academic college of engineering, Tel Aviv, Israel

Abstract

In this paper, we report on classification results for emotional user states (4 classes, German database of children interacting with a pet robot). Starting with 5 emotion labels per word, we obtained chunks with different degrees of prototypicality. Six sites computed acoustic and linguistic features independently from each other. A total of 4232 features were pooled together and grouped into 10 low level descriptor types. For each of these groups separately and for all taken together, classification results using Support Vector Machines are reported for 150 features each with the highest individual Information Gain Ratio, for a scale of prototypicality. With both acoustic and linguistic features, we obtained a relative improvement of up to 27.6%, going from low to higher prototypicality.

Index Terms: emotion, prototypes, feature types, automatic classification

1. Introduction

In word recognition, it is either/or: either the word to be recognized 'is there', or not. This holds true even if the signal-to-noise ratio is unfavourable, or the speaker is slurring. In emotion recognition, it is different: neither do we know exactly, what an emotion is, nor do we *know* — in a literal meaning (ground truth) — in the actual case which emotion is expressed by the speaker up to which extent. Thus we have to resort to specific strategies: we can use acted speech — with or without subsequent perceptual evaluation — or we can use 'spontaneous' speech with some external (context-) annotation, or human labelers. Acted data are relatively easy to get and can be designed in such a way that there are enough cases per class; produced by a good actor, they can be conceived as pure and (sort of) prototypical. However, their relevance for modelling realistic data is doubtful. Realistic data, on the other hand, is normally unbalanced and sparse. Moreover, there is ample evidence that realistic emotions are not either/or but can be mixed or more or less pronounced, i.e. prototypical. Normally, more than one labeler is employed that either can annotate explicitly whether the emotion is pronounced/mixed or not, or we can resort to majority voting for assigning an emotion label. For instance, for the database dealt with in this paper, we employed five labelers and normally use as majority voting an agreement of at least three out of these five labelers [1, 2, 3]. This is a reasonable but of course not the only possible strategy: we could have used another threshold, for instance, at least four out of five. Less or not prototypical cases constitute a sort of waste-paper-basket category; in order to get a somehow balanced sample, they are often, together with the majority of the default class (i.e., neutral), discarded. No matter which strategy has been chosen, it should be described fully, together with possible consequences: if we do not select but deal with the whole database, classes might be 'noisy' and classification performance rather low. This is, however, a realistic setting. If we select most prototypical cases, classification performance might be considerably higher but we will not be dealing with a fully realistic scenario.

The 'prototypical' study on automatic classification of emotion first selects the data and the class assignment — which is then kept constant throughout the study — and tries to optimize performance by varying, for example, features and feature selection, or (types of) classifiers. The focus of the present study is the opposite: we vary the degree of prototypicality and keep everything else constant. For that, we use a very large feature vector with subsequent feature selection, and state-of-the-art classification procedures. The idea behind is, of course, that using only more prototypical cases yields higher classification performance. We do not know of any other study so far where degrees of prototypicality and classification performance were systematically investigated.

Another question we want to address is whether the use of prototypical cases — and let's assume that acted data are prototypical as well — can be beneficial for classification: if we take such clear cases and classify all cases, even the less clear ones, will performance be higher, the same, or less? In this paper, we cannot discuss in what way acted data are different from spontaneous, prototypical data — they surely are. Seen from a pattern recognition / data mining point of view, we simply can focus on classification performance. The third topic we want to deal with is whether the relevance of different acoustic and linguistic feature types changes if we go over from less prototypical to more prototypical constellations.

In Sec. 2 we introduce the database used and describe the labeling process resulting in each pattern assigned to a unique

The initiative to co-operate was taken within the European Network of Excellence HUMAINE under the name CEICES (Combining Efforts for Improving automatic Classification of Emotional user States). 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 responsibility lies with the authors.

emotional class. In the same section we define prototypes. In Sec. 3 we give a short overview of the feature types employed. Experimental setup and prototypes' performance are described in Sec. 4. In Sec. 5 we summarize the most important results.

2. Chunks: Labels and Prototypes

The database consists of German recordings of children communicating with Sony's AIBO pet robot; it is described in more detail in [1, 2, 3]. The basic idea was to collect 'natural' emotional speech within a Wizard-of-Oz scenario: a human operator causes the AIBO to perform a fixed, predetermined sequence of actions, provoking emotional reactions in the child. The whole database comprises about 9.2 hours of speech without pauses collected at two different schools from 51 children (age $10-13$, 21 male, 30 female). We segmented the data into 'turns' of variable length, using as criterion a pause of $\geq$ 1.5 sec. By that we got many one-word turns, but even turns with up to > 50 words. Five labelers (advanced students of linguistics) listened to the turns in sequential order and annotated independently from each other each word as *neutral* or as belonging to the one of ten other classes. If three or more labelers agreed, the label was attributed to the word (majority voting). All in all, 48401 words were labeled.

As emotional units match better with semantically meaningful chunks than turns containing up to > 50 words, we clustered words into chunks. Eventually, these chunks which represent the patterns in the classification process, were labeled by mapping word labels onto chunks by the procedure described below. The chunking was based on syntactic and prosodic information. First, we performed a coarse, manual, syntactic labeling 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. Eventually, we applied a prosodic, additional criterion: if the pause between words is ≥ 500 msec, we assume a chunk boundary; the length of the pauses between words was obtained from the manually corrected word segmentation. This is a reasonable, heuristic criterion for segmenting syntactically ill-formed sequences which often can be observed in scenarios such as 'giving commands to a pet (robot)'. Before labeling the chunks, as some of the labels were very sparse, we resorted to down-sampling *neutral* and *emphatic* classes, while *touchy*, *reprimanding*, and *angry* were mapped under the same cover class *angry*. Patterns assigned to the classes *surprised*, *helpless*, *irritated*, *bored*, *joyful* and *rest* were discarded due to sparse data. Weighted kappa for the four (cover) classes *motherese*, *neutral*, *emphatic*, and *angry* is 0.59. The mapping of word- onto chunk-based labels followed basically the strategy described in [2]: for each chunk, we pooled the labels given by our 5 labelers (for a chunk of n words, we obtain $n \times 5$ labels). For the chunks to be mapped onto **N***eutral*[1], 60% of the (word) labels had to be *neutral*. If 40% or more are non-neutral, then the chunk is **A***ngry*, **M***otherese*, or **E***mphatic*: the chunk is mapped onto **M***otherese* if at least 50% of the non-neutral labels are *motherese*, otherwise, if *emphatic* is more (as) frequent than (as) *anger*, the turn is mapped onto **E***mphatic*. The remaining chunks, which are neither **E***mphatic* nor **M***otherese*, are defined as **A***ngry*. By that we employ some 'markedness' criterion: **M***otherese* is more marked than **E***mphatic* and **A***ngry*, and all are more marked than **N***eutral* (see the example below in this section). This procedure yielded 4543 chunks (914 **A***ngry*, 586 **M***otherese*, 1045 **E***mphatic*, and 1998 **N***eutral* — cf. 'w/o' (without) in Fig. 1) with 2.9 words per chunk on average.

The $l = n \times 5$ word-based emotion labels assigned to each chunk (made of n words) can serve to select the most *prototypical* patterns. More in general, we can construct ensembles of prototypical chunks that have a certain expectation of being representative of the emotion they are assigned to. This expectation might be simply estimated by looking at the labelers' agreement on the final chunk label L: $\#L/l \times 100$ in percentage. By using opportune thresholds thr we can construct prototypical datasets with growing degrees of representativity. How prototypes are obtained, can be better explained with the following example:

chunk	**A**	**M**	**E**	**N**	# words	label	agreem.
0052	7	0	7	11	5	**E**	28%
0056	1	0	1	13	3	**N**	87%
0037	2	8	0	10	4	**M**	40%

Chunk 0052, annotated as **E***mphatic* following the 'markedness' procedure described above, has $7/(5 \times 5) = 28\%$ of the labels supporting this decision, while chunk 0056 has $13/(3 \times 5) = 87\%$ agreement on the **N***eutral* choice. Therefore, a threshold thr=60% would discard chunk 0052 (and 0037 as well) from the prototype list while it would retain chunk 0056. The higher the threshold thr (which reflects the degree of representativity of a certain group of prototypes), the lower the retaining number of patterns. In the upper part of Fig. 1 we draw the distributions of the prototypes over the class labels and over the threshold thr. Note, that for thresholds above 90% the number of patterns of the classes **E***mphatic* and **M***otherese* is almost negligible, especially in relation to the neutral patterns. Below in Fig. 1, we sketch the number of words per chunk as thr increases. In general, the number of words per chunk thereby decreases, co-varying with higher inter-labeler correspondence for shorter chunks. Differently for what happens with **N***eutral* patterns — thr is weakly correlated with # of words per **N***eutral* chunk, which is always above 3.5 — **E***mphatic* is falling more

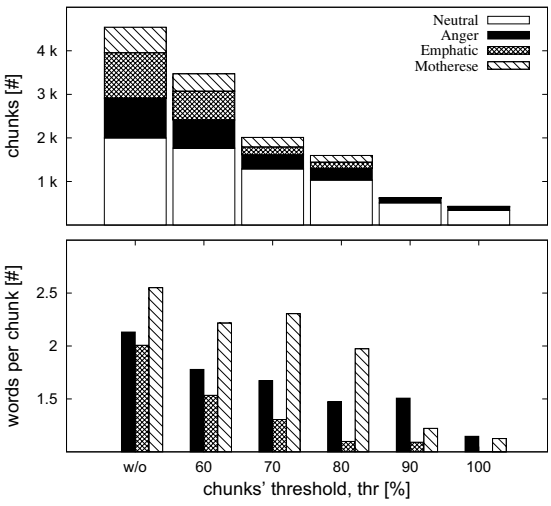

Figure 1: *Above: distributions of the chunks over the emotional classes and over some representative prototypes' thresholds thr. Below, words per chunk, for all but* **N**eutral *class, and for different thr. The initial distribution is labeled with 'w/o' (without). Thresholds are in percentage.*

[1]Labels given to chunks have initial letter boldfaced: this letter is used in the figures to identify the respective emotional class.

steeply, maybe because of frequent emphatic stand-alone words such as 'stop', while **M***otherese* is slowly falling until thr=80, probably because child-directed speech often consists of more-word chunks (e.g. 'yeah, that's fine').

3. Features

The 4232 features used in this study can be clustered according to the *type* of Low Level Descriptors (LLDs). We concentrate on a characterisation in phonetic and linguistic terms — *what* has been extracted rather than *how* it has been extracted.

Acoustic features comprise: **Duration** ($\# = 391$) LLDs model temporal aspects of the speech signal. Positions of prominent energy or F0 values on the time axis are attributed to this type as well. **Energy** (265) features model intensity, based on the amplitude in different intervals. **Pitch** (333): the acoustic equivalent to the perceptual unit pitch is measured in Hz and often made perceptually more adequate by, e.g., logarithmic transformation. Intervals, characterising points, or contours are being modelled. **Spectrum** (656) or formant (spectral maxima) LLDs model spoken content, esp. lower ones. Higher formants 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 mean spectrum over a chunk averages out formant information, giving general spectral trends. **Cepstrum** (1699): MFCC features 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. **Voice quality** (153) features comprise HNR, jitter, shimmer, and other measures of microprosody. They are based in part on pitch and intensity but reflect voice quality such as breathiness or harshness. **Wavelets** (216) give a short-term multi-resolution analysis of frequencies over time.

Linguistic features include: **Bag Of Words, BOW** (476) which are well known from document retrieval tasks, showing good results for emotion recognition as well. Each term within a vocabulary is represented by an individual feature modelling the term's (logarithmical and normalized) frequency within the current phrase. Terms are clustered with Iterated Lovins Stemming. **Part Of Speech, POS** (31) classes (frequencies in the chunk) represent a coarse taxonomy of six lexical and morphological main word classes based on the spoken word chain.

thr.	**w/o**	**60**	**70**	**80**	**90**	**100**
#	4543	3472	2012	1597	630	430
acoustic features						
RR	61.7	69.6	77.3	80.3	91.6	89.5
CL	60.7	66.0	73.3	70.9	60.8	58.2
F	61.2	67.8	75.3	75.3	73.1	70.5
linguistic features						
RR	62.6	68.8	75.7	78.5	86.8	90.5
CL	62.2	67.9	75.6	73.5	80.1	86.2
F	62.4	68.3	75.6	75.9	83.3	88.3
acoustic & linguistic features						
RR	63.6	72.8	80.5	81.7	92.2	90.9
CL	62.5	71.4	79.1	73.6	68.1	73.9
F	63.0	72.1	79.8	77.5	78.3	81.5

Table 1: *Classification performance of different prototypical groups of chunks. Figures [%] are obtained by 3-fold SVM speaker-independent stratified cross-validation. Each experiment was conducted on 150 selected (IGR) features only.*

Higher **Semantics** (12) features (frequencies in the chunk) are based on a coarse taxonomy into six classes, (partly scenario-specific) most relevant words, word classes, and emotional valence (negative vs. positive), based on the spoken word chain.

Different types of normalizations are applied to the LLDs' base contours listed above before functionals are applied: features are often both – raw and normalized within each chunk. Linguistic features are not extracted fully automatically as the word transliteration was manually checked. A more detailed overview of the extracted features per site is given in [3].

4. Experiments and Results

Different grades of prototypicality, cf. Tab. 1 (increment: 10%) and Fig. 2 (increment: 5%), were classified with Support Vector Machines: linear kernel, one-against-one multi-class discrimination, sequential minimal optimization [4]; this constellation provided best accuracy in [3]. To eliminate the bias towards the most frequent classes, cf. **N***eutral* in Fig. 1, we up-sampled the training sets by n-times complete instance repetition per class except **N***eutral* so that we finally approximated uniform class distributions. Note that the multiplying factors n depend on the thresholds thr chosen for the tuning of level of prototypization. Classification results were obtained by partitioning the data set into three balanced splits meeting the following requirements (in the 'w/o' configuration): no splitting within-subject chunks, quite similar distribution of chunk labels, and balance between the two schools and genders. The three splits were exploited in a 3-fold cross validation framework. Performance figures are reported as F-measure, defined here as the uniformly weighted harmonic mean of the class averaged recall (CL) and overall recognition rate (RR). This is a slightly different definition from the standard one [5]. However, this approach seems to be more adequate in a multi-classification problem.

In the first set of experiments we focused on evaluating the effects of prototyping on classification, using different subsets of prototypes obtained with the procedure described above. Results are reported in Tab. 1 using the 150 best acoustic or linguistic features, or the 150 best acoustic and linguis-

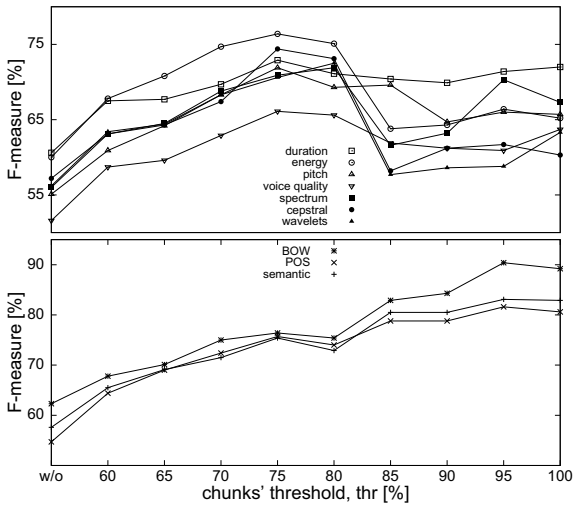

Figure 2: *Classification results using single feature types (LLD), each one previously reduced to 150 by IGR. The lines represent LLD performance trends over growing prototyping thresholds. Acoustic features are grouped above, linguistic features below.*

tic features. For feature selection, we used open-loop non-decorrelating compression by highest Information Gain Ratio (IGR), cf. [3]. Results for individual LLDs (Fig. 2) were obtained by applying IGR within each group separately: this feature selection approach is fair across LLDs because the original number of features is very unequal (see Sec. 3: 1699 cepstral vs. 153 voice quality). Groups with less than 150 features were not reduced: for instance, F-measures obtained using POS features only are computed by exploiting 31 features.

Note that the results in Tab. 1 (and Fig. 2) are not directly comparable across increments — the test sets are prototypized as well, and therefore contain different patterns. Nonetheless, they can be taken as exemplars for experiments with different grades of prototypes. First of all, we see that prototypical patterns clearly allow to boost classification performance, also in spite of the reduced and more unbalanced dataset.[2] The 'rubicon' for such an improvement seems to be between a *thr* of 70 and 80, i.e. 75 (Fig. 2): for 'acoust. & ling. features' F-measure is slightly better than 79.8% (*thr*=70, Tab. 1), namely 80.4% (not shown); this amounts to a relative improvement on 'w/o' of 27.6%.[3] Fig. 1 reveals that this is at the transition of a real 4-class problem to a 'mutilated' 4-class problem which is rather a 2-class problem because of sparse data for **M***otherese* and **E***mphatic*. This holds for all acoustic features but **duration** which is, in a way, a sort of linguistic feature as well: content words are on average longer than function words, and by looking at the linguistic features, we see that all three types, i.e. POS as well, are not really impeded by the $thr \geq 75$.

In Fig. 2, classification results over different prototypical sets are reported for the single LLDs. The trends basically reflects what we already saw with both acoustic and linguistic features together: the former set is more sensible to the curse of dimensionality which becomes more compelling as *thr* increases, apart from **duration**, cf. above. The differences between the feature types (e.g., best are **energy** and **duration**) has been discussed in more detail in [3].

To check the effectiveness of prototypes in real-life conditions, we trained our classifier on prototypes and tested on all the original (*thr*=w/o) patterns. That way, the test sets always consist of the same 4543 chunks. Classification performance in these conditions are reported in Tab. 2. To some extent, these results confirm similar findings in other areas (e.g. [6]): data cleaning (or data pruning) is effective for removing meaningless and mislabeled patterns. However, "suspicious patterns may not be garbage patterns" as noisy data too is needed to make the classifier learn those difficult patterns (both ambiguous and atypical). The minimum of the generalization error in Tab. 2 is for 'acoust. & ling. features' at *thr*=60; higher *thr*'s obviously result in over-cleaning the training data. This relatively low threshold probably means that although ambiguous, many patterns are still important and do characterize spontaneous emotional speech. It is noticeable though that F-measures do not heavily decrease until *thr*=80, where the curse of dimensionality problem gets overwhelming.

thr.	**w/o**	**60**	**70**	**80**	**90**	**100**
#	4543	3472	2012	1597	630	430
acoustic features						
RR	61.7	63.1	61.6	61.4	56.6	55.2
CL	60.7	58.5	56.7	51.9	43.6	40.8
F	61.2	60.7	59.0	56.3	49.3	46.9
linguistic features						
RR	62.6	62.9	63.2	62.1	60.0	58.2
CL	62.2	60.0	59.4	54.8	49.3	45.4
F	62.4	61.4	61.2	58.2	54.1	51.0
acoustic & linguistic features						
RR	63.6	66.4	65.9	63.2	58.5	56.8
CL	62.5	63.6	62.1	54.9	46.3	42.9
F	63.0	65.0	63.9	58.8	51.7	48.9

Table 2: *Classification performance of SVM* trained *on different prototypical data sets: here,* test *patterns are* not *prototypes. Figures [%] are obtained by 3-fold SVM cross-validation. Each experiments was conducted on 150 selected (IGR) features only. Differently from Tab.1, # is the number of* training *patterns only.*

5. Conclusions

Our data and our results can be seen as being typical for realistic databases: a tidy, balanced set of classes is not given, and can even less be maintained when going over to more prototypical constellations. However, we could demonstrate that the degree of prototypicality chosen clearly amounts to a marked difference in classification performance, e.g. for *thr*=w/o vs. 75 to 17.4 percent points (27.6% relative improvement). This difference is higher than the one normally obtained by optimizing feature sets or classifiers, cf. for our data [2, 3]. Second, Tab. 2 reveals that prototypes cannot fully model variability in the data and, used for training, yield minor improvements. Even if our prototypes cannot simply be put on the same level as acted data, this result makes it less probable that using acted data for training is the solution for the sparse data problem.

6. References

[1] A. Batliner, S. Steidl, C. Hacker, E. Nöth, and H. Niemann, "Tales of Tuning – Prototyping for Automatic Classification of Emotional User States," in *Proc. Interspeech*, Lisbon, 2005, pp. 489–492.

[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 *Proc. IS-LTC 2006*, Ljubliana, 2006, pp. 240–245.

[3] 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*, Antwerp, 2007, pp. 2253–2256.

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

[5] J. Makhoul, F. Kubala, R. Schwartz, and R. Weischedel, "Performance measures for information extraction," in *Proc. DARPA Broadcast News Workshop*, Herndon, VA, USA, 1999.

[6] N. Matic, I. Guyon, L. Bottou, J. Denker, and V. Vapnik, "Computer aided cleaning of large databases for character recognition," in *Proc. Int. Conf. on Pattern Recognition*, 1992, pp. 330–333.

[2] As expected, there is a clear negative correlation between *thr* and number of words per chunk, cf. Fig. 1: **M***otherese*: -0.95, **N***eutral*: -0.52, **E***mphatic*: -0.84, and **A***ngry*: -0.87: the higher *thr*, the lower the number of words; this holds esp. for the non-neutral classes. For them, there is an even — albeit weak — negative correlation between number of words and labelers' agreement, e.g., for w/o: **M***otherese*: -0.26, **N***eutral*: 0.03, **E***mphatic*: -0.49, and **A***ngry*: -0.39.

[3] If we look at 'acoust. & ling. features' in Tab. 1, we see that RR and CL are balanced for *thr*=w/o,60, 70. (This is true for 'linguistic features' as well but not for 'acoustic features'.) This means in turn that recall for all four classes is rather balanced, for instance, for a *thr* of 70, **M***otherese*: 79.0%, **N***eutral*: 82.3%, **E***mphatic*: 80.6%, and **A***ngry*: 74.3%. If RR and CL are not balanced, the default class **N***eutral* is classified better, and esp. **M***otherese* worse.