Novel Useful Features and Algorithms for the Recognition of Emotions in Human Speech

Pierre-yves Oudeyer

Sony Computer Science Lab, Paris, France
py@csl.sony.fr

Abstract

Recent years have been marked by the development of robotic pets or partners such as small animals or humanoids. People interact with them using natural human social cues, in particular emotional expressions. It is crucial that robot can detect the emotional information contained in speech using only prosodic features, since this is often the only information that they can measure. We present here the first large scale experiment in which a large feature set space and a large machine learning algorithm space are searched concurrently. We describe new features which prove to be much more efficient than the traditional features used in the litterature.

1. Introduction

Recent years have been marked by the increasing development of personal robots, either used as new educational technologies or for pure entertainment/ Typically, these robots look like familiar pets such as dogs or cats (e.g. the Sony AIBO robot), or sometimes take the shape of young children such as the humanoids SDR-5 (Sony).

Among the capabilities that these personal robots need, one of the most basic is the ability recognize human emotions. Indeed, not only emotions are crucial to human reasoning, but they are central to social regulation. Emotional communication is at the same time primitive enough and efficient enough so that we use it a lot when we interact with pets, in particular when we tame them. This is also certainly what allows children to bootstrap language learning and should be inspiring to teach robots natural language.

In this paper, we present a set of experiments that formed the basis of a technology for automatically recognizing the emotions in speech based on prosodic features, and used now in certain entertainment robots such as the Sony AIBO or SDR-4. This work is the first (to our knowledge) large scale data mining experiment in which we compare most of the standard machine learning algorithms and explore the value of two hundred different features. As shown below, we found some new features of which efficiency seems to be significantly higher than the ones traditionally used in the literature. Besides, all the work presented here is based on the use of freely available softwares and thus can be reproduced with minor difficulties.

2. The acoustic correlates of emotions in human speech

It is possible to achieve our goal only if there are some reliable acoustic correlates of emotion/affect in the acoustic characteristics of the signal. A number of researchers have already investigated this question ([3]). Their results agree on the speech correlates that come from physiological constraints and correspond to broad classes of basic emotions, but disagree and are unclear when one looks at the differences between the acoustic correlates of for instance fear and surprise or boredom and sadness. Indeed, certain emotional states are often correlated with particular physiological states ([6] which in turn have quite mechanical and thus predictable effects on speech, especially on pitch, (fundamental frequency F0) timing and voice quality.

3. The recognition of emotions in human speech

3.1. Goal

As interesting interactions need to be 2-ways, it is necessary that robotic pets can also recognize the emotions of the humans who are interacting with them. Human generally do that by using all the context and modalities, ranging from linguistic content to facial expression and intonation. Unfortunately, using appropriately context is not easy for a machine in an uncontrolled environment: for instance robust speech recognition in such situations is out of reach for nowadays systems, and facial expression recognition needs both computational resources and video devices that robotic creatures most often do not have. For this reason we investigated how far we can go by using only the prosodic information of the voice. Furthermore, the speech we are interested in is the kind that occurs in everyday conversations, which means short unformal utterances, as opposed to the speech produced when one is asked to read emotionally a paragraph of for example a newspaper. Four broad classes of emotional content were studied: joy/pleasure, sorrow/sadness/grief, anger and calm/neutral.

3.2. Existing work

The first studies that were conducted (e.g. [8]) were not so much trying to get an efficient machine recognition device, but rather were searching for general qualitative acoustic correlates of emotion in speech (for example: happiness tends to make the mean pitch of utterances higher than in calm sentences). More recently, the increasing awareness that affective computing had an important industrial potential ([6]) pushed research towards the quest of performance in automatic recognition of emotions in speech. Unfortunately, to our knowledge, no large scale study using the modern tools developped in the machine learning community have been conducted. Indeed, most often, either only one or two learning schemes are tested (for e.g. in [7], [2]) or very few and simple features are used ([7], [2]), or only small databases are used - less than 100 examples per speaker (like foe e.g. in [2], [4], [7]) which makes that the power of some statistical learning schemes may have been overlooked.

Only ([4]) have tried to make some systematic data mining, using more than the traditional/standard set of features used by the rest of the literature: mean, max, min, max-min, variance of the pitch and intensity distributions, and of the lengths of phonemic or syllabic segments, or of pitch rising segments. Unfortunately, this work lacks many experiments: 1) only 3 kinds of learning schemes were used - support vector machines, gaussian mixtures and linear discriminants - which are far from being the best at dealing with data in which there are possibly many unrrelevant features, and in particular do not allow to derive automatically smaller set of features with optimal efficiency; 2) the feature set was explored by choosing one learning scheme and iteratively removing less useful features for classification: on one hand this is rather ad hoc since it is linked to a very particular learning schemes and selection procedure, on the other hand it does not allow to detect the fitness of groups of features. Finally, their work is based on speech generated by asking human subjects to read newspaper texts in an emotional manner , which does not correspond to our constraints. To our knowledge, only two research groups have tried to build automatic recognition machines of everyday speech are ([2], [7]). Yet, they could only use very small databases, very few simple features and 2 different learning algorithms. Finally, a general conclusion of this already existing corpus of research is that recognition rates above 60 percent even with only 4 basic emotions seems impossible if there are several speakers. The enormous speaker variability has been described in ([7]). As a conclusion, we chose to focus only on speaker dependant emotion recognition. This is not necessarily a bad point from an industrial point of view since it is targeted to robotic pets that may interact mainly only with their caretaker (and the fact that robots only manage to recognize their owner could even be a positive feature, because it is a source oc complicity between a robot and its caretaker).

Our methodology is an extension of the work of ([4]) in which we use more features (including new and crucial ones), more learning schemes, and more standard feature space exploration tools. A very large database of 2 speakers containing unformal short emotional utterances is used. All experiments were conducted using the freely available data mining software Weka [1] , which implements most of the standards data mining techniques.

3.3. The database

In order to have sufficiently large databases, we had to make some compromises (the recording conditions as described in ([7]) or ([2]) are rather poor and unpractical). So we used two japanese professional speakers (a man and a woman), who are both voice actor/actress and worked on many radio/TV commercials, Japanese dub of movies and animations. They were asked to imitate everyday speech by pronouncing short sentences or phrase like “Umm, I don’t know”, “Exactly!”, “See”, “Hello”, “I see”, “How are you?”, “What kind of food do you like?”, “Wonderful!”, “D’know”. Before each utterance, they had to imagine themselves in a situation where they could utter it, and which would correspond to one of the four emotional classes: joy/pleasure, sorrow/sadness/grief, anger, normal/neutral. If several emotions were compatible with the sentence meaning, then they were allowed to utter each of them. We ended with a database of 200 examples per speaker and per emotions, which makes 2000 samples in total. We know that having only two speakers restrains the generality of the results, but to our knowldege no one so far had the opportunity to have so many examples, even for one speaker, and so to use the power of modern statistical learning algorithms. Nevertheless, the making of more databases is planned.

3.4. Using data mining techniques

3.4.1. Features

The two main measures that can be done concerning the intonation are pitch and intensity, which we did, like in all the works reported above. For each signal, we also measured the intensity of its low-passed and high-passed version, the cutting frequency being chosen at 250 Hz (the particular value appears not to be crucial), the idea being to separate the signal into a pure prosodic component and a pure “spectral” component. Finally, for sake of exhaustivity, we made a spectral measure consisting in computing the norm of the absolute vector derivative of the first 10 MFCC components (mel-frequency ceptral components). All these measure were performed at each 0.01s time frame, using the Praat software, which is a signal processing toolkit freely available [2].

Each of these measures provides a serie of values that we had to transform to provide different points of view upon the data. So each serie of values was transformed into 4 series: the serie of its minimas, the serie of its maximas, the serie of the durations between local extrema of the 10Hz smoothed curve (which models rhytmic aspects of the signal), and the serie itself. Finally, to get features out of these series, we computed for each one: the mean, the maximum, the minimum, the difference between the maximum and the minimum, the variance, the median, the first quartile, the third quartile and the interquartile range, and the mean of the absolute value of the local derivative. In total we used 5*4*10 = 200 features.

3.4.2. Learning algorithms

There are many learning schemes that have been developed in the last 20 years (see [9], and they are often not equivalent: some are more efficient with certain types of class distributions than others, and some are better at dealing with many unrelevant features (which is the case here, as seen a posteriori) or with structured feature sets (in which this is the “syntactic” combination of the values of features which is crucial). As by definition we do not know the structure of our data and/or the (ir-)relevance of features, it would be a mistake to investigate our problem with only very few learning schemes. As a consequence, we chose to use a set of the most representative learning schemes, ranging from neural networks to rule induction or classification by regression. Also, we used one of the best meta-learning scheme, i.e. AdaBoostM1 ([9]), which allows generally the significant improvement in generalization performance for unstable learning schemes like decision trees (an unstable learning algorithm is one that can sometimes produce very different recognition machines when only a slight change in the learning database has been performed). We chose to use the Weka software, of which code and executable are freely available so that the experiment, though being large scale, can be easily reproduced. This software also provides means like automatic cross-validation, or the search of feature spaces with for e.g. genetic algorithms as we will see later. The list of all learning algorithms is given in table 4. More details about these algorithms can be found in [9].

[1] Weka web page: http://www.cs.waikato.ac.nz/ ml/

[2] Praat web page: http://www.praat.org

name	description
1-NN	1 nearest neighbours
5-NN	voted 2 nearest neighbours
10-NN	voted 10 nearest neighbours
Decision Tree/C4.5	C4.5 decision trees
Decision Rules/PART	PART decision rules
Kernel Density	Radial Basis Function Neural Net.
KStar	KStar
Linear Regression	classification via linear regression
LWR	classification via locally weighted regression
Voted Perceptrons	commitee of perceptrons
SVM 1	polynomial (deg. 1) Support Vector Machine
SVM 2	polynomial (deg. 2) Support Vector Machine
SVM 3	polynomial (deg. 3) Support Vector Machine
VFI	Voted features interval
M5Prime	clsiification via M5PRime regression method
Naive Bayes	Naive Bayes classification algorithm
AdaBoostM1/C4.5	Adaboosted version of C4.5
AdaboostM1/PART	Adaboosted version of PART

Table 1: Learning schemes

name	speaker 1	speaker 2
1-NN	82	87
5-NN	84	87
10-NN	83	87
Decision Trees/C4.5	84	93
Decision Rules/PART	84	94
Kernel Density	84	90
Kstar	81	85
Linear Regression	88	91
LWR	87	90
Voted Perceptrons	70	76
SVM degree 1	88	94
SVM degree 2	89	94
SVM degree 3	88	94
VFI	80	93
M5Prime	86	96
Naive Bayes	84	90
AdaBoost M1/C4.5	90	96
AdaBoost M1/PART	91	97

Table 2: Using all features

3.4.3. All features/All algorithms

In a first experiment, evaluation was conducted in which all algorithms were given all the (normalized) features, and were trained on 90 percent of the database and tested on the remaining 10 percent. This was repeated 10 times with each time a different 90/10 percent split (we performed a 10-fold cross-validation). Table 5 gives the average percentage of correct classification for the 10 folds.

We see from these results that very high success rate (between 92 and 97 percent, which is higher than any other reported result in the literature [3]. can be obtained thanks to the use of certain algorithms. Yet, the difference among algorithms is striking: whereas the best results are obtained with adaboosted decision trees and rules, some others perform 10 percent below (like nearest neighbours, RBF neural nets or Support Vector Machines, which are the ones typically used in other studies), or even 20 percent below (commitees of perceptrons). This illustrates our initial claim that one must be careful to try many different learning schemes when one wants to solve a problem about which we have very few prior or intuitive knowledge. It is not surprising that the best results are obtained with decision trees and rules since these kinds of algorithms are known to be very good at dealing with many unrelevant features, which seems to be the case here (if not, there would be less disparity between results).

[3]Of course, it is difficult to compare because databases are different, but at least the features and the algorithms used elsewhere are all strictly included in this study

feature	information gain
1: MEDIANINTENSITYLOW	1.44
2: MEANINTENSITYLOW	1.40
3: THIRDQUARTINTENSITYLOW	1.35
4: ONEQUARTINTENSITYLOW	1.34
5: MAXINTENSITYLOW	1.23
6: MININTENSITYLOW	1.14
7: THIRDQUARTMINIMASPITCH	0.72
8: THIRQUARTMAXIMASPITCH	0.72
9: THIRDQUARTPITCH	0.69
10: MAXMINIMASPITCH	0.67
11: MAXMAXIMASPITCH	0.67
12: MAXPITCH	0.67
13: MINMINIMASPITCH	0.59
14: MEDIANMINIMASPITCH	0.57
15: MEDIANMAXIMASPITCH	0.57
16: MINPITCH	0.52
17: MEDIANPITCH	0.52
18: MEANMINIMASPITCH	0.48
19: MEANMAXIMASPITCH	0.48
20:MEANPITCH	0.48

Table 3: Information Gain of 20 best features

3.5. Feature selection

After this first experiment, one naturally would like to see how the feature set could be reduced for three reasons: 1) small features set provide better generalization performance in general (see [9]); 2) obviously, it is computationally cheaper to compute less features; 3) it is interesting to see if the most useful features for the machine learning algorithms are the ones that are traditionally put forward in the psychoacoustic literature.

A first way of exploring the feature set is to look at the results of learning schemes like decision rules (PART), which are often used mainly as knowledge discovery devices:

```
If MEDIANINTENSITYLOW > 0.48 and
   MINIMASPITCH <= 0.07 and
   THIRDQUARTINTENSITY > 0.42 ==> CALM

ELSE If MEANINTENSITYLOW <= 0.58 and
   MEDIANINTENSITYLOW <= 0.29 ==> ANGRY

ELSE If THIRDQUARTINTENSITYLOW > 0.48 ==> SAD

ELSE ==> HAPPY
```

These four and surprisingly simple rules allow a percentage of correct classification in generalization of 94 percent for the speaker 2 database ! The striking fact is the repeated use of features related to the intensity of the low-pass signal.

In order to quantify the individual relevance of features or attributes, there is a measure often used in the data mining literature, which is the expected information gain, or mutual information between class and attribute. It corresponds to the difference between the entropies H(class) and H(class—attribute) (see [9], for details about how it is computed). Table 6 gives the 20 best attributes according to the information gain they provide.

This table confirms the great value of the features concerning the quartiles of the distribution of intensity values in the low-passed signals. It also show something rather surprising: among the 20 most individually informative features, only 3 (the 12, 16 and 20) are part of the standard set put forward in psychoacoustic studies ([5], [3], Williams 1972) or used in most of more application oriented research as in (Slaney et al. 1998, Breazal 2000).

Yet, one has to be aware that individual salience of a feature is only partially interesting: it is not rare that success comes from the combination of features. So in a first experiment, we tried to compare a feature set containing only the features 1 to 6 related to low-passed signal intensity (LPF), with a feature

learning scheme	(LPF) sp.1	(LPF) sp.2	(SF) sp.1	(SF) sp.2
1-NN	78	83	70	72
5-NN	84	82	72	75
10-NN	84	82	73	73
Decision Trees/C4.5	80	84	72	71
Decision Rules/PART	78	83	72	74
Kernel Density	82	85	71	74
Kstar	80	84	70	72
Linear Regression	63	68	72	78
LWR	75	71	75	80
Voted Perceptrons	51	70	60	58
SVM degree 1	63	68	73	78
SVM degree 2	71	70	77	50
SVM degree 3	76	85	78	82
VFI	78	76	64	70
M5Prime 83	85	76	80	
Naive Bayes	82	81	74	72
AdaBoost M1/C4.5	80	81	80	78
AdaBoost M1/PART	80	83	79	78

Table 4: Comparing “standard” features and “low-passed signal intensity” features

name	speaker 1	speaker 2
1-NN	87	92
5-NN	90	92
10-NN	87	91
Decision Trees/C4.5	85	92
Decision Rules/PART	86	93
Kernel Density	87	91
Kstar	86	90
Linear Regression	83	89
LWR	87	89
Voted Perceptrons	65	78
SVM degree 1	87	91
SVM degree 2	90	96
SVM degree 3	89	94
VFI	83	92
M5Prime	88	95
Naive Bayes	89	93
AdaBoost M1/C4.5	90	96
AdaBoost M1/PART	90	96

Table 5: Using the “optimal” feature set

set composed of the standard features (SF) used in (Breazal 2000, or Slaneyet al. 1998): mean, min, max, max-min, and variance of pitch and intensity of unfiltetered signal, plus mean length of syllabic segments (Results are similar if we add jitter and tremor as sometimes also used). Table N summarizes these experiments (each number corresponds again to the mean percentage of correct classification in generalization in 10-fold cross-validation).

This table shows that if one uses only the quartiles of the low-passed signal intensity, one still outperforms the combination of features used traditionally. Because here we have only few speakers this result has to be taken with caution, but it seems to indicate that previous work missed something crucial.

Finally, as we saw on this table, using only low-passed intensity features yields substantially lower results that when one used all features with decision rules. In order to attain our goal of finding a very efficient small set of features, we used an automatic search method: genetic algorithms. Populations of features (limited to 30) were generated and evolved using as fitness the 10-fold cross-validation with 2 algorithms: Naive Bayes and 5-Nearest Neighbours (we chose these mainly because they are fast to train). The outcome of this experiment was not obvious: within the selected feature set, not surprisingly, there were features related to the quartiles of low-passed signal intensity and features related to the quartiles of pitch, but also features with relatively low individual information gain: those related to the quartiles of the minimas of the unfiltered smoothed intensity curve. Also, we can note that again, the machine learning algorithms tend to always neglect features related to the variance or the range of distributions, whatever the measure. A final experiment using these 15 features along with all learning algorithms was conducted (max, min, median, 3rd quartile and 1st quartile of low-passed signal intensity, pitch and minimas of unfilterd signal intensity). Results are summarized in table 8.

We observe that we get very similar best results than initially, with more than 10 times less features. Moreover and interestingly, the variation between learning schemes is less important and algorithms which performed badly like nearest neighbours or Naive Bayes, behave now in a more satisfaying manner.

4. Conclusion

We showed that using on a large scale modern data mining techniques allowed to find non-obvious features which were missed in precedent studies. In particular, it is interesting to see that the features put forward in the psychoacoustic litterature are not the preferred ones of machine learning algorithms. As precedent studies seemed to show that multi-speaker emotion recognition was a very difficult task in principle, the present work suggest that speaker dependant recognition can reach very high scores, if adequate features and learning schemes are used. This work should serve as a basis for necessary additional experiments with more databases including speakers of very different languages. The use of only freely available softwares should allow other people who already posess these databases to help to pursue this research.

5. References

[1] Banse, R.; Sherer, K. R. 1996. Acoustic Profiles in Vocal Emotion Expression, Journal of Personality and Social Psychology, 70(3): 614-636.

[2] Breazal, C. 2000. Sociable Machines: Expressive Social Exchange Between Humans and Robots, PhD Thesis, MIT AI Lab.

[3] Burkhardt F.; Sendlmeier W. 2000. Verification of Acoustical Correlates of Emotional speech Using Formant-synthesis, in Proceedings of the ISCA Workshop Speech and Emotion.

[4] McGilloway S. et al. 2000. Approaching Automatic Recognition of Emotion from Voice: a Rough Benchmark, in Proceedings of the ISCA Workshop on Speech and Emotion.

[5] Murray E.; Arnott J.L. 1995. Implementation and Testing of a System for Producing emotion-by-rule in Synthetic Speech, Speech Communication, 16(4), pp. 369-390.

[6] Picard R. 1997. Affective Computing, MIT Press.

[7] Slaney M.; McRoberts G. 1998. Baby Ears: A Recognition System For Affective Vocalization, in Proceedings of ICASSP 1998.

[8] Williams U.; Stevens K.N. 1972. Emotions and Speech: some acoustical correlates, JASA 52, 1238-1250.

[9] Witten I.; Frank E. 2000. Data Mining, Morgan Kaufflan Publishers.