The Production and Recognition of Emotions in Speech: Features and Algorithms

Oudeyer Pierre-Yves

Sony CSL Paris,

6, rue Amyot,

75005 Paris, France

Abstract

Recent years have been marked by the development of robotic pets or partners such as small animals or humanoids. The interactions with them are very different from those with traditional computers : instead of having human beings using robotic conventions, robots should learn to communicate in a humanised fashion. In particular, they need to be able to express and recognize emotions. This can be done in part using speech, which has the advantage to be computationally cheap and practical to implement in real world robots. Nevertheless, research in this area is still very young. We present here algorithms that allow a young robot to express its emotions like babies do. They are very simple and efficiently provide life-like speech thanks to the use of concatenative speech synthesis. We describe a technique which allows to control continuously both the age of a synthetic voice and the quantity of emotions that are expressed. This is useful since personal robots may grow up and have many degrees of emotions. Also, we present the first large-scale data mining

experiment about the automatic recognition basic emotions in unformal everyday short utterances. We focus on the speaker dependant problem. We compare a large set of machine learning algorithms, ranging from neural networks, Support Vector Machines or decision trees, together with 200 features, using a large database of several thousands examples. We show that the difference of performance among learning schemes can be substantial, and that some features which were previously unexplored are of crucial importance. An optimal feature set is derived through the use of a genetic algorithm. Finally, we explain how this study can be applied to real world situations in which possibly very few examples are available. Furthermore, we describe a game to play with a personal robot which allows to teach it examples of emotional utterances in a natural and rather unconstrained manner.

Key words:

emotions, speech, robots, production, recognition

---

1 Introduction

Recent years have been marked by the increasing development of personal robots, either used as new educational technologies (Druin and Hendler, 2000) or for pure entertainment (Fujita and Kitano, 1998; Kusahara, 2000). Typically, these robots look like familiar pets such as dogs or cats (e.g. the Sony AIBO robot), or sometimes they take the shape of young children such as the humanoids SDR3-X (Sony). The interactions with these machines are radically different from the way we interact with traditional computers. So far humans have been learning to use very unnatural conventions and devices such as key-

Email address: `py@csl.sony.fr` (Oudeyer Pierre-Yves).

URL: `www.csl.sony.fr/ py` (Oudeyer Pierre-Yves).

boards or dialog windows, and need to know how computers work to be able to use them. Opposite to that, personal robots should try themselves to learn the natural conventions (such as natural language or social rules like politeness) with the appropriate modalities (such as speech or touch) that humans have been using for thousands of years.

Among the capabilities these personal robots need, the most basic is the ability to grasp human emotions (Picard 1997), and in particular they should be able to recognize human emotions as well as express their own emotions. Indeed, not only emotions are crucial to human reasoning, but they are central to social regulation (Halliday, 1975) and in particular to control dialog flows. Emotional communication is both 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 (Halliday, 1975) and should be inspiring to teach robots natural language.

Apart from the words that we use, we express our emotions in two main ways : the modulation of facial expression (Ekman, 1982) and the modulation of the intonation of the voice (Banse and Sherer, 1996). Whereas research about automated recognition of emotions in facial expressions is now very rich (Samal and Yengar, 1992), research dealing with the speech modality, both for automated production and recognition by machines, has only been active for very few years (Bosch, 2000). In this paper, we present the results of our research which consisted in providing to robots means to express emotions vocally and in enabling them to recognize some basic emotional information in its caretaker's voice. Both aspects are original : as far as production is concerned, and unlike most of existing work, we are dealing with cartoon-like meaningless speech, which has different needs and constraints than for example trying

to produce naturally sounding adult-like normal emotional speech. For example we would like the emotions to be recognized by people of different cultural or linguistic background. Our work has similarities with the one of (Breazal, 2001), but we use concatenative speech synthesis and our algorithm is more simple and completely specified. As far as the recognition of emotions is concerned, we present here 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 which seem to be more efficient 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. A web site [1] containing some accompanying material such as sounds and graphs is also available.

Next section presents general information about the acoustic correlates of emotion in speech, which form the basis of our work. Section 3 presents our algorithm for the production of emotion as well as its validation with human subjects. Section 4 presents the results of our data mining experiment concerning learning algorithms and useful features in the recognition of emotions in the human voice.

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 num-

[1] www.csl.sony.fr/py

ber of researchers have already investigated this question (Fairbanks 1940, Burkhard and Sendlmeier 2000, Banse and Sherer 1996). 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 (Picard 1997) which in turn have quite mechanical and thus predictable effects on speech, especially on pitch, (fundamental frequency F0) timing and voice quality. For instance, when one is in a state of anger, fear or joy, the sympathetic nervous system is aroused, the heart rate and blood pressure increase, the mouth becomes dry and there are occasional muscle tremors. Speech is then loud, fast and enunciated with strong high frequency energy. When one is bored or sad, the parasympathetic nervous system is aroused, the heart rate and blood pressure decrease and salivation increases, producing speech that is slow, low-pitched and with little high frequency energy (Breazal, 2001).

Furthermore, the fact that these physiological effects are rather universal means that there are common tendencies in the acoustical correlates of basic emotions across different cultures. This has been precisely investigated in studies like (Abelin and Allwood 2000) or (Tickle 2000) who made experiments in which American people had to recognize the emotion of either another American or a Japanese person only using the acoustic information (the utterances were meaningless, so there were no semantic information). Reversely, Japanese listeners were asked to decide which emotions other Japanese or American people were trying to convey. Two results came out of it : 1) there was only little difference between the performance in detecting the emotions conveyed

by someone speaking the same language or the other language, and this is true for Japanese as well as for American subjects ; 2) subjects were far from being perfect recognizer in the absolute : the best recognition score was 60 percent (This result could be partly explained by the fact that subjects were asked to pronounce nonsense utterances, which is quite unnatural, but is confirmed by studies asking people to utter semantically neutral but meaningful sentences (Burkhart and Sendlmeier 2000)). The first result indicates that our goal to build a machine that can express affect, both with meaningless speech and in a way recognizable by people from different cultures with the accuracy of a human speaker, is attainable in theory. The second result shows that we should not expect perfect recognition, and compare the machine's performance in relation to human performance. The fact that humans are not so good is mainly explained by the fact that several emotional states have very similar physiological correlates and thus acoustic correlates. In actual situations, we solve the ambiguities by using the context and/or other modalities. Indeed, some experiments have shown that the multi-modal nature of the expression of affect can lead to a McGurk effect for emotions (see Massaro 2000): a face showing emotion A and speaking with emotion B is perceived as expressing either only one of the two emotions or sometimes even a third one. Also, different contexts may lead people to interpret the same intonation as expressing different emotions for each context (see Cauldwell 2000). These findings indicate that we shall not try to have our machine generate utterances that make fine distinctions ; only the most basic affect categories should be investigated.

A number of experiments using computer based techniques of sound manipulation have been conducted to explore which particular aspects of speech reflect emotions with most saliency. (Murray and Arnott, 1993; Banse and

Scherer, 1996; Burkhardt and Sendlmeier, 2000; Williams and Stevens, 1972) basically all agree that the most crucial aspects are those related to prosody : the pitch (or f0) contour, the intensity contour and the timing of utterances. Some more recent studies have shown that voice quality (Gobl and Chasaide, 2000) and certain co-articulatory phenomena (Kienast and Sendlmeier, 2000) are also reasonably correlated with certain emotions.

3 The generation of cartoon emotional speech

3.1 Goal

The goal of this research is quite different from most of existing work in synthetic emotional speech. Whereas traditionally (see Cahn 1990, Iriondo et al. 2000, Edgington 1997, Iida et al. 2000) the aim is to produce adult-like naturally occuring emotional speech, here the target is to provide a young creature with the ability to express its emotions in an exaggerated/cartoon manner, while using nonsense words (this is necessary for us because we use this in experiments with robots to which we try to teach language : this pre-linguistic ability to use only intonation to express basic emotions serves to bootstrap learning; yet, we will not give more details about this point since it falls far beyond the scope of this paper). The speech should sound lively, not repetitive, and similar to infants' babbling. Finally, we wanted people from very different linguistic and cultural background are able to recognize easily the creature's emotions.

Additionally, we wanted to have algorithms as simple as possible and to control as few parameters as possible : in brief, what is the simplest manner to

transmit emotions with prosodic variations ? Also, the speech had to be both of high quality and computationally cheap to generate (robotic creatures have usually only very scarce resources). For these reasons, we chose to use a concatenative speech synthesizer (Dutoit and al., 1993), the MBROLA software freely available on the web [2], which is an enhancement of more traditional PSOLA techniques (it produces less distortions when pitch is manipulated). The price of quality is that very few control over the signal is possible, but this is compatible with our need of simplicity.

Because of all these constraints, we have chosen to investigate only five emotional states so far, corresponding to calm and one for each of the four regions defined by the two dimensions of arousal and valence: anger, sadness, happiness, comfort.

3.2 Existing work

As said above, existing work has concentrated on adult-like naturally sounding emotional speech, and most of projects have tackeled only one language. Many of them (see Cahn, 1990; Murray and Arnott, 1995; Burkhardt and Sendlmeier, 2000) have used formant synthesis as a basis, mainly because it allows detailed and rich control of the speech signal : one can control voice quality, pitch, intensity, spectral energy distributions, harmonics-to-noise ratio or articulatory precision which allows to model many co-articulation effects occurring in emotional speech. The drawbacks of formant synthesis are that quality of the produced speech remains not satisfying (voices are often still quite unnatural). Furthermore, the algorithms developped in this case are

[2] MBROLA web page : http://tcts.fpms.ac.be/synthesis/mbrola.html

complicated and necessitate the control of many parameters, which renders their fine tuning quite impractical (see Cahn, 1990 for a discussion). Unlike these works, (Breazeal, 2000) has described a system which is very similar to ours: based on (Cahn, 1990), she made a system for her robot Kismet that allows it to produce meaningless emotional speech. Like the work of Cahn, it relies heavily on the use of a commercial speech synthesizer of which many parameters are often high level (for example, specification of the pitch baseline of a sentence) and implemented in an undocumented manner. As a consequence, this is hardly reproducible if one wants to use another speech synthesis system. On the contrary, the algorithm we will describe here is completely specified, and can be used directly with any PSOLA-based system (besides, the one we used here can be freely downloaded, see above). Another drawback of Breazeal's work is that the synthesizer she used was formant based, which does not correspond to our constraints.

Because of their very superior quality, concatenative speech synthesizers (Dutoit et al., 2000) have gained popularity in the recent years, and some have tried to use them to produce emotional speech. This is a challenge significantly more difficult than with formant synthesis since only the pitch contour, the intensity contour and the duration of phonemes can be controlled (and yet, there are narrow contraints over this control). To our knowledge, two approaches have been presented in the literature. The first one, as described in (Iida et al., 2000), uses one speech database for each emotion as the basis of the prerecorded segments to be concatenated in the synthesis. This gives satisfying results but is quite impractical if one wants to change the voice or add new emotions or even control the degree of emotions. The second approach consists in (see for example Edgington, 1997) making databases of human produced

emotional speech and computing the pitch and intensity contours and apply them to sentences to be generated. This brings some problems of alignments, partially solved using syntactic similarities between sentences. Finally, Edgington (1997) showed that this method gave quite unsatisfying results (speech sounds unnatural and emotions are not very well recognized by human listeners). Finally, these two methods are unapplicable to our work since there would be great difficulties to make speech databases of exaggerated/cartoon baby voices.

The approach we take here is from an algorithmic point of view completely generative (it does not rely on the recording of human speech that would serve as input), and uses concatenative speech synthesis as a basis. We will show that it allows to express emotions as efficiently as with formant synthesis, but with simpler controls and the liveliness of concatenative speech synthesis.

3.3 A simple and complete algorithm

Our algorithm consists in generating a meaningless sentence and specifying the pitch contour and the duration of phonemes (the rhythm of the sentence). For the sake of simplicity, we specify only one target per phoneme for the pitch, which reveals enough. We could have fine control over the intensity contour, but as we will show, this is not necessary, since manipulating the pitch can create the auditory illusion of intensity variations. We only control the overall volume of sentences. Our program generates a file which is fed into the MBROLA speech synthesizer. This file looks like :

```
l 448 10 150 80 158 ;; means : phoneme ‘‘l’’ duration 448 ms,
                    ;; at 10 percent of 448 ms
                    ;; try to reach 150 Hz, at 80 percent
                    ;; try to reach 158 Hz
9~ 557 80 208
b 131 80 179
@ 77 20 200 80 229
b 405 80 169
o 537 80 219
v 574 80 183.0
a 142 80 208.0
n 131 80 221.0
i 15 80 271.0
H 117 80 278.0
E 323 5 200
```

The first step of the algorithm is to generate a sentence composed of random words, each word being composed of random syllables (of type CV or CCV). Initially, the duration of all phonemes is constant and the pitch of each phoneme is constant equal to a pre-determined value (noise is added, which is crucial if one wants the speech to sound natural; we tried many different kinds of noise, and this does not make significant differences; for the perceptual experiment reported below, gaussian noise was used). Then the pitch and duration informations of this sentence are altered so as to yield a particular affect. Deformations consist of deciding that a number of syllables become stressed, and in applying a certain stress contour on these syllables as well as

some duration modifications. Also, all syllables are applied a certain default pitch contour and duration deformation. For each phoneme, we give only one pitch target fixed at 80 percent of the duration of the phoneme. Let us now state more precisely the different steps of the algorithm (words in capital letters denote parameters of the algorithm that need to be set for each emotion) :

```
1 Choose the number of words of the sentence
       (random number between 2 and MAXWORDS);
2 Create the words :
3  For each word, choose the number of syllables
4               (random number between 2 and MAXSYLL), and
5               decides with probability PROBACCENT whether
                the word is accented or not ;
6 If the word is accented then choose randomly one
7            of its syllables and mark it as accented ;
8 Create the syllables :
9  For each syllable
10   choose whether this is a CV or a CCV syllable
11                (CV syllable have probability 0.8) ;
12   instantiate the C's and V by picking randomly a
13                consonnant or vowel in the phoneme database ;
14   set the duration of each phoneme to MEANDUR + random(DURVAR) ;
15   let e = MEANPITCH + random(PITCHVAR)
16     set the pitch of consonnants to e - PITCHVAR
17     set the pitch of vowels to e + PITCHVAR
```

```
18   if the syllable is accented then
19    add DURVAR to the duration of its phonemes ;
20    if DEFAULTCONTOUR = rising
21       set the pitch of consonants to MAXPITCH - PITCHVAR
22       set the pitch of the vowel to MAXPITCH + PITCHVAR
23    if DEFAULTCONTOUR = falling
24       set the pitch of consonants to MAXPITCH + PITCHVAR
25       set the pitch of the vowel to MAXPITCH - PITCHVAR
26    if DEFAULTCONTOUR = stable
27       set the pitch of phonemes to MAXPITCH
28
29 Change the contour of the last word :
30 if not LASTWORDACCENTED
31    let e = PITCHVAR/2
32        if CONTOURLASTWORD = FALLING
33        for each syllable in word
34          add -(i+1)*e pitch of phonemes to their value
                         (i = index of phoneme in syllable)
35          e = e + e
36        if CONTOURLASTWORD = RISING
37        for each syllable in word
38          add +(i+1)*e pitch of phonemes to their value
39                       (i = index of phoneme in syllable)
40          e = e + e
41  else
42    if  CONTOURLASTWORD = FALLING
43    for each syllable in word
```

```
44            add DURVAR to the duration of its phonemes ;
45            set the pitch of consonants to MAXPITCH + PITCHVAR
46            set the pitch of the vowel to MAXPITCH - PITCHVAR
47      if CONTOURLASTWORD = RISING
48      for each syllable in word
49            add DURVAR to the duration of its phonemes ;
50            set the pitch of consonants to MAXPITCH - PITCHVAR
51            set the pitch of the vowel to MAXPITCH + PITCHVAR
52
53 Set the loudness volume of the complete sentence to VOLUME.
```

A few remarks can be made concerning this algorithm. First, it is useful to have words instead of just dealing with random sequences of syllables because it avoids to put accents on adjacent syllables too often. Also it allows to express more easily the operations done on the last word. Typically, the maximum number of words in a sentence (MAXWORDS) does not depend on the particular affect, but is rather a parameter than can be freely varied. A key aspect of this algorithm are the stochastic parts : on the one hand, it allows to produce, for a given set of parameters, a different utterance each time (mainly thanks to the random number of words, the random constituents of phonemes of syllables or the probabilistic attribution of accents) ; on the other hand, details like adding noise to the duration and pitch of phonemes (see line 14 and 15 where random(n) means "random number between 0 and n") are fundamental to the naturalness of the vocalizations (if it remains fixed, then one perceives clearly that this is a machine talking). Finally, let us remark that here accents are implemented only by changing the pitch and not the loudness.

Nevertheless, it gives satisfying results since in human speech an increase in loudness is correlated to an increase in pitch. Of course here we had to exaggerate the pitch modulation, but this is fine since as we explained earlier, our goal is not to reproduce faithfully the way humans express emotions, but to produce a lively and natural caricature of the way they express emotions (cartoon-like). Finally, a last step is added to the algorithm in order to get a voice typical of a young creature : the sound file sampling rate is overriden by setting it to 30000 or 35000 Hz as compared to the 16000 Hz produced by MBROLA (this is equivalent to playing the file quicker). Of course, so that the speech rate remains normal, it is initially made slower in the program sent to MBROLA. Only the voice quality and pitch are modified. This last step is necessary since no child voice database exists for MBROLA. So a female adult voice was choosen.

Now that we have described in details the algorithm, let us give (see table 1) examples of the parameters' values obtained for 5 affects : calm, anger, sadness, happiness, comfort. The way these parameters were obtained was by first looking at studies describing the acoustic correlates of each emotion (e.g. Murray and Arnott 1993, Sendlmeier and Burkhartd 2000), then deducing some coherent initial value for the parameters and modifying them by hand, and trial and error until it gave a satisfaying result.

3.4 Validation with human subjects

In order to evaluate the algorithm described in section 3.3, an experiment was conducted in which human subjects were asked to describe the emotion

	Calm	Anger	Sadness
LASTWORDACCENTED	NIL	NIL	NIL
MEANPITCH	280	450	270
PITCHVAR	10	100	30
MAXPITCH	370	100	250
MEANDUR	200	150	300
DURVAR	100	20	100
PROBACCENT	0.4	0.4	0
DEFAULTCONTOUR	RISING	FALLING	FALLING
CONTOURLASTWORD	RISING	FALLING	FALLING
VOLUME	1	2	1

	Comfort	Happiness	
LASTWORDACCENTED	TRUE	TRUE	
MEANPITCH	300	400	
PITCHVAR	50	100	
MAXPITCH	350	600	
MEANDUR	300	170	
DURVAR	150	50	
PROBACCENT	0.2	0.3	
DEFAULTCONTOUR	RISING	RISING	
CONTOURLASTWORD	RISING	RISING	
VOLUME	2	0	

Table 1

Parameter values for different emotions

they felt when hearing a vocalization produced by the system [3]. More precisely, each subject first listened to 10 examples of vocalizations, with emotion randomly chosen for each example, so that they got used to the voice of the system. Then they were presented a sequence of 30 vocalizations (unsupervised serie) , each time corresponding to an emotion randomly choosen, and were asked to make a choice between "Calm", "Anger", "Sadness", "Comfort" and "Happiness". They could hear each example only once. In a second experiment

[3] Some sample sounds are available on the associated web page www.csl.sony.fr/py

with different subjects, they were initially given 4 supervised examples of each emotion, which means they were presented vocalization together with a label of the intended emotion. Again they were presented 30 vocalizations that they had to describe with one of the word cited above. 8 naive adult subjects were in each experiment : 3 French subjects, 1 English subject, 1 German subject, 1 Brazilian subject, and 2 Japanese subjects (none of them was familiar with the research or had special knowledge about the acoustic correlates of emotion in speech). Table 2 shows the results for the unsupervised serie experiment. The number in the (rowEm,columnEm) means the percentage of times a vocalization intended to represent rowEm emotion was perceived as columnEm emotion. For instance in the Table 2,we see that 76 percent of vocalizations intended to represent sadness were effectively perceived as sadness.

The results of the unsupervised serie experiment have to be compared with experiments done with human speech instead of machine speech. They show that for similar setups, like in (Tickle 2000) in which humans were asked to produce nonsence emotional speech, at best humans have 60 percent success, and most often less. Here we see that the mean result is 57 percent, which compares well to human performance. The errors are of two types: the most frequent is related with a confusion with the neutral/calm emotion. This is the less annoying error since it does not involve a confusion between aroused/not aroused and negative/positive. There are also (but much less) confusion between anger and happiness, but not between comfort and sadness, which means that confusions about valence appear only for aroused speech. Finally, there are nearly no confusions between aroused and not aroused speech.

A second unsupervised experiment was performed, similar to the one reported here except that the calm affect was removed. A mean success of 75 percent

	Calm	Anger	Sadness	Comfort	Happiness
Calm	**36**	1	1	30	30
Anger	0	**65**	0	0	35
Sadness	20	0	**76**	4	0
Comfort	**45**	0	16	39	0
Happiness	5	30	0	5	**60**

Table 2

Confusion matrix for the unsupervised series

was obtained, which is a great increase and is much better than human performance. This can be explained in part by the fact that here the acoustical correlates of emotions are exagerated. The results presented here are similar to those reported in (Breazal 2001), which proves that using a concatenative synthesizer with a lot less parameters still allows to convey emotions (and in general provides more life-like sounds).

Examination of the supervised serie shows that the presentation of only very few vocalizations with their intended emotion (4 exactly for each emotion), results increase very much : now 77 percent success is achieved. We see that confusions involving the neutral emotion and confusions between anger and happiness have nearly disappeared. Similarly, an experiment in which the calm affect was removed was conducted, which gave a mean success of 89 percent. This supervision is something that can be implemented quite easily with digital pets : many of them use for combinations of color LED lights to express their "emotions", and the present experiment shows that it would be enough to visually see the robot a few times while it is uttering emotional sentences to be able later to recognize its intended emotion just by listening to it.

	Calm	Anger	Sadness	Comfort	Happiness
Calm	**76**	3	4	14	3
Anger	0	**92**	0	0	8
Sadness	8	0	**76**	16	0
Comfort	15	0	5	**77**	3
Happiness	4	20	0	8	**68**

Table 3

Confusion matrix for the supervised series

3.5 Varying continuously the age of the voice and the degree of emotion

Typically, robotic pets are initially "babies" and shall grow up and develop along with time and interactions with humans (or other pets). It seems natural that their voice evolves accordingly, and in a continuous manner. To our knowledge, this problem has not already been addressed in the literature. Generally, one has several voice databases (for the segments used by the speech concatenizer) to choose from, and corresponding to different ages. Yet, on one hand each database is made with a different person (having a voice of the desired age), which means that it is clear to the human ear that the voice is also from a different person, which of course is not acceptable in our case. On the other hand, only a limited number of databases are available (because having a lot is impractical and requires a lot of memory), which means that age can not vary in a smooth manner.

We found a solution to this problem, which is rather simple but sufficient (see the associated web page for samples). When one has a vocalization signal, in order to change only the age of the voice, it is enough to override the sample rate and then use the PSOLA algorithm to modify the length of the new sound so that it remains the same as in the original sound. In our case, the 'default' age is a signal sampled at 32000 Hz (which we use in the validation experiment

in next section) : if we want to make the signal sound 'older', then we can override the sample rate to for example 28000 Hz, and then use PSOLA to shorten the signal back to its original time length.

Furthermore, it would be obviously useful that robotic pets may be able to vary the degrees of emotion that they express: for instance they could make a difference between happy and very happy. Again, we did not find this question addressed anywhere in the literature. We propose to add to each set of parameters of 'normal' degree of emotion (those described in previous section), an associated set of similar parameters corresponding to the highest degree of emotion for a give emotion (for e.g. very very happy). For example, to the parameter MEANPITCH (= 400) of (normal) happiness, we add a parameter MEANPITCH2 (= 500); then we define a variable *delta* taking values in [-1;1] which determines the degree of one emotion: 0 is for normal, 1 for maximum and -1 for minimum. When a vocalization is to be generated, *delta* has to be set and the actual mean pitch of the utterance becomes : MEANPITCH + *delta**MEANPITCH2. We are in fact making some kind of local linear models of emotion degrees. This experimentally gives satisfying results, and requires to specify only one additional set of parameters, while allowing an infinite range of nuances.

4 Validation of age and emotion degree control

In order to validate the techniques presented in the precedent part, we made some tests with the eight human subjects used above. As far as age control is concerned, each subject was presented pairs of utterances with a random emotion and asked which one looks older than the other. The re-sampling fre-

quencies were taken in the range [25000;35000] Hz. Each subject was presented 50 pairs of utterances (the difference of re-sampling frequency was always superior to 1000 Hz and random). The mean rate of correct age ranking was 92.4 percent, which is satisfying for our goal.

For the evaluation of the control over the degree of emotion, the same test was repeated except that the age was fixed (32000 Hz), and pairs consisted in two utterances of the same emotion (but each time random), with a different (random) degree. Human subjects had to evaluate which utterance expressed with a higher degree the emotion. Again 50 pairs were presented to each subject. The mean rate of correct ranking was 85.1 percent, which is again satisfaying for our goal.

5 The recognition of emotions in human speech

5.1 Goal

It is necessary that robotic pets can also recognize the emotions expressed by the humans who are interacting with them. Human beings generally do that by using all the context and modalities, ranging from lexic to facial expression and intonation. Unfortunately, using appropriately the context is not an easy thing 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 prosodic information voice. Furthermore, the speech we are interested in is the kind that occurs in everyday conversations,

which means short informal 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.

5.2 Existing work

As opposed to the automatic recognition of emotions with facial expression (Samal and Iyengar, 1992), research using the speech modality is still very young (Bosch, 2000). The first studies that were conducted (e.g. Murray and Arnott 1993, Williams and Stevens, 1972) 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 has an important industrial potential (Picard, 1997) pushed research towards the quest for performance in automatic recognition of emotions in speech (Bosh, 2000). Unfortunately, to our knowledge, no large scale study using the modern tools developped in the data mining and machine learning community has been conducted. Indeed, most often, either only one or two learning schemes are tested (for e.g. in Polzin and Waibel 2000, Slaney and McRoberts 1998, Breazal 2001) or very few and simple features are used (Polzin and Waibel 2000, Slaney and McRoberts 1998, Breazal 2001, Whiteside 1997), or only small databases are used - less than 100 examples per speaker (as in Breazal 2001, McGilloway et al. 2000, Slaney et al. 1998) which means that the power of some statistical learning schemes may have been overlooked.

Only (McGilloway and al. 2000) 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. This work has some drawbacks: 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 irrelevant 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 scheme 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 (Breazal 2001, Slaney et al. 1998). Yet, they only used very small databases, very few features and two 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 (Slaney et al. 1998). As a conclusion, we chose to focus only on speaker dependent 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 only with their caretakers (and the fact that robots only manage to recognize their owner could even be a positive feature, because it is a source of complicity between

a robot and its caretaker).

Our methodology is an extension of the work of (McGilloway and al. 2000) in which we use more features (including new and crucial ones), more learning schemes, and more powerful feature space exploration tools. A very large database of six speakers containing informal short emotional utterances is used. All experiments were conducted using the freely available data mining software Weka [4] , which implements most of the standards data mining techniques.

5.3 The database

In order to have sufficiently large databases, we had to make some compromises (the recording conditions as described in (Slaney et al, 1998) or (Breazal 2001) were too impractical for us to make several thousands samples). So we used six Japanese professional speakers (men and women), who are both voice actor/actress and worked on many radio/TV commercials, Japanese dubbing of movies and animations. They were asked to imitate everyday speech by pronouncing short sentences or phrase like "Great !", "Exactly!", "See", "Hello", "I see", "How are you?", "What kind of food do you like?", "Wonderful!", "What is your name ?" (these are of course the english translation of the japanese utterances). They had to imagine that they would utter these sentences to a pet robot. Before each utterance, they had to imagine themselves in a situation where they could pronounce 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,

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

then they were allowed to utter each of them. Each example in the database was evaluated by human subjects who had to decide if they were appropriate or not (whether the utterance's intonation compatible with the emotion or not). We ended with a database of 200 examples per speaker and per emotion, which makes 4800 samples in total. We know that having only six 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. Another potential drawback of the database is that there might be a self-entrainment of the speakers: as they are asked to perform a particular task with their voice intonation, they might produce less variable speech than in natural situations.

5.4 Using data mining techniques

5.4.1 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). 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 measures were performed at each 0.01s time frame, using the Praat software, which is a signal processing toolkit freely available [5]. In particular, the pitch was computed using the algorithm described in (Boersma, 1993), which is

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

known to be very accurate.

Each of these measures provides a time series of values that we had to transform to produce different points of view upon the data. So each serie of values was transformed into 4 series : the series of its minima, the series of its maxima, the series of the durations between local extrema of the 10Hz smoothed curve (which models rhytmic aspects of the signal), and the series 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.

5.4.2 Learning algorithms

There are many learning schemes that have been developed in the last 20 years (see Witten and Frank, 2000), 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 (Witten and Frank, 2000), which allows generally significant

name	description
1-NN	1 nearest neighbour
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
SVM 4	Gaussian kernel 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 4

Learning schemes

improvement on 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 (for e.g. with 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 (Witten and Frank, 2000).

name	mean correct generalization rate across 6 speakers
1-NN	84.5
5-NN	85.2
10-NN	84.4
Decision Trees/C4.5	94.1
Decision Rules/PART	94
Kernel Density	85.2
Kstar	81
Linear Regression	89.7
LWR	88.3
Voted Perceptrons	75.9
SVM degree 1	92.1
SVM degree 2	91.2
SVM degree 3	90.9
SVM 4	91.5
VFI	88.2
M5Prime	90.4
Naive Bayes	89.8
AdaBoost M1/C4.5	95.7
AdaBoost M1/PART	94.8

Table 5

Using all features

5.4.3 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 rates are obtained (95.7 percent). These figures are higher than any other reported in the literature but

one has to be careful since the number of classes and the types of classes are most of the time unique to each paper. For example, (Slaney and McRoberts, 1998) report rates around 80 percent for the speaker dependant recognition, with only three classes (but they were different than ours: approval, prohibition, attention). The best way to compare is in fact to look at the results of individual learning schemes and features in our paper, since all the learning schemes and features used in the papers that we quote are included here. 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).

5.5 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 Witten and Frank, 2000); 2) obviously, it is computationally cheaper to compute fewer 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
   THIRDQUARTMINIMASPITCH <= 0.07 and
   THIRDQUARTINTENSITY > 0.42 ==> CALM

ELSE If MEANINTENSITYLOW <= 0.58 and
   MEDIANINTENSITYLOW <= 0.29 and
   THIRDQUARTMAXIMASPITCH > 0.1 ==> ANGRY

ELSE If THIRDQUARTINTENSITYLOW > 0.48 ==> SAD

ELSE ==> HAPPY
```

These four and surprisingly simple rules allow a percentage of correct classification in generalization of 94.4 percent for the speaker number four in the database. The striking fact is the repeated use of features related to the intensity of the low-pass signal.

To get another view of the feature set, one can also simply try to visualize it. Just to confirm the precedent intuition that low-passed intensity is crucial in the distinction of emotions, figure 1 plots the database with axis being the 1st quartile and the 3rd quartile of the intensity distribution, and figure 2 being

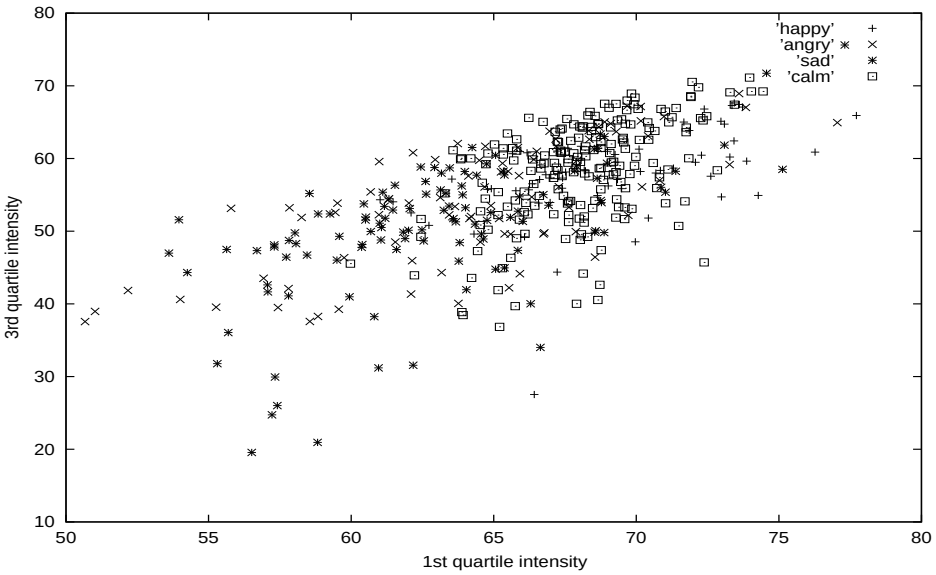

Fig. 1. Data points in speaker 1 database : 1st quartile of intensity distribution againts 3d quartile of intensity distribution

the same but for the intensity of the low-passed signal. This is for speaker 2. The same very striking effect happens also for the other speakers, but what is interesting is that the clusters are not situated at the same places (anger and happiness are 90 degrees rotated), which is an illustration of the great speaker variability that we presented earlier. The difference is not a scaling difference, but a qualitative difference that no learning schemes could learn with these features. Yet, it seems that the use of some well chosen features is very stable for each speaker.

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 Witten and Frank, 2000, 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

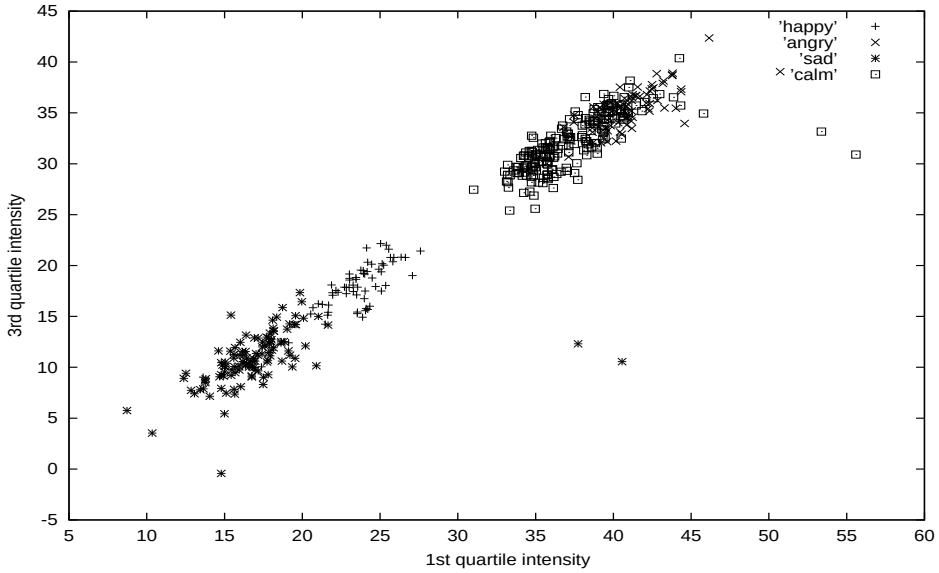

Fig. 2. The same data points than in last figure, except that we plot here the 1st quartile of the low-passed signal intensity distribution against the 3rd quartile of the low-passed signal intensity distribution

of the distribution of intensity values in the low-passed signals. It also shows 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 (Murray and Arnott 1996, Sendlmeier and Burkhardt 2000, Stevens and Williams 1972) or used in most of more application oriented research as in (Slaney et al. 1998, Breazal 2001).

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 set composed of the standard features (SF) used in (Breazal 2001) or (Slaney et 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 7 summarizes these

feature	information gain (mean across 6 speakers)
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 6

Information Gain of 20 best features

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 gets results extremely similar than when we use standard features, and the best result is obtained with the low-passed intensity related features (85.9 percent). 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

learning scheme	LPF (mean across speakers)	SF (mean across speakers)
1-NN	78.1	82.7
5-NN	84.1	81.9
10-NN	79.2	79.1
Decision Trees/C4.5	80.1	81.2
Decision Rules/PART	79.9	80.4
Kernel Density	85.9	79.1
Kstar	80.4	81.2
Linear Regression	63.1	64.1
LWR	75.6	72.9
Voted Perceptrons	51.2	60.4
SVM degree 1	63.1	65.7
SVM degree 2	71.2	70.1
SVM degree 3	76.8	76.4
SVM 4	85.1	79.4
VFI	79.1	76.0
M5Prime	85.5	82.3
Naive Bayes	82.1	80.7
AdaBoost M1/C4.5	82.1	82.8
AdaBoost M1/PART	83.2	82.9

Table 7

Comparing "standard" features and "low-passed signal intensity" features

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 exact genetic algorithm is the simple one described in (Goldberg, 1989). 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 the minimas of the pitch contour, but also features with relatively low individual information

name	correct generalization rate (mean for 6 speakers)
1-NN	92.1
5-NN	92.5
10-NN	91.4
Decision Trees/C4.5	92.9
Decision Rules/PART	94.1
Kernel Density	90.1
Kstar	86
Linear Regression	84.6
LWR	88.9
Voted Perceptrons	75.4
SVM degree 1	90.1
SVM degree 2	95.9
SVM degree 3	94.2
SVM 4	92.1
VFI	84.1
M5Prime	92.5
Naive Bayes	90.8
AdaBoost M1/C4.5	96.1
AdaBoost M1/PART	95.4

Table 8

Using the "optimal" feature set

gain : those related to the quartiles of the minimas of the unfiltered smoothed intensity curve. 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 unfiltered signal intensity). Results are summarized in table 8.

We observe that we get very similar highest 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 satisfying manner

(yet, for these two, this is not surprising since the fetaure set was selected using these algorithms as evaluators).

5.6 When only very few examples are provided

In last section, we used large training databases : this was crucial to explore feature and algorithmic spaces, but as we are dealing with a speaker dependent task, this is not directly applicable to a real world robotic pet. Indeed, it is not conceivable that the owner of such a robot would give hundreds of supervised examples to teach it how to recognize its way of expressing basic emotions (yet, this is what probably happens with human babies and real pets, but humans tend to be more willing at spending a lot of time with them than with robotic pets). Then it is natural to ask what the results will become if only very few training examples are given.

We made an experiment using the "optimal" feature set found earlier. We gave to each algorithms only 12 examples of each class, and tested them on the remaining items of the database. This was repeated 30 times with different sets of 12 examples and results were averaged (the standard deviation was rather low, typically around 1.1) Table 9 summarizes the experiment.

We see that some of the algorithms manage to keep a very reasonable level of performance (90.1 percent of success in generalization for adaboosted PART). Among them, examples of very cheap algorithms like 1-nearest neighbours or Naive Bayes. These results are rather comparable (and in fact slightly superior) to what is described in (Breazal, 2001) (except than in this case, learning was off-line with a larger database of several female speakers) : what is important

learning scheme	mean across 6 speakers
1-NN	85.1
5-NN	78.9
10-NN	69.4
Decision Trees/C4.5	79.1
Decision Rules/PART	80.1
Kernel Density	84.2
Kstar	75.6
Linear Regression	74.8
LWR	79.1
Voted Perceptrons	50.2
SVM degree 1	83.2
SVM degree 2	85.4
SVM degree 3	84.9
SVM 4	85.1
VFI	77.1
M5Prime	80.9
Naive Bayes	85.1
AdaBoost M1/C4.5	84.2
AdaBoost M1/PART	90.1

Table 9

When very few training examples are provided

is that Breazal conducted experiments and showed that this level of success is sufficient to develop interesting interactions with a robotic pet. Also, she showed how these results could be substantially improved when integrated into a larger cognitive architecture which is working in the real world. For example, linking this recognition module to an artificial lymbic/emotional system in which there is some kind of emotional inertia (one very rarely switches from anger to happiness in half a second) might give some additional information or tell the system there is uncertainty about the result. As a consequence, the robot may for instance take a posture showing it is not sure of what is happening and the human will often repeat his utterance with an even more

exaggerated intonation. This provides 2 samples instead of one, one of them being often very stylized.

5.7 Teaching a robot in the real world

In the previous paragraph, we saw that the use of adequate features and algorithms allows a reasonable rate of correct recognitions in the speaker dependent case. There remains the problem of providing these examples to the robot in a user-friendly manner : indeed, as stated at the beginning of the paper, we are to communicate with robotic pets in a relatively natural manner. This implies that it is not acceptable for instance to ask the user to connect its robot to a computer and use a windows/mouse based interface to record samples.

We have developed and experimented a small game, in the spirit of language games used in robotic models of the acquisition of language (Steels, 1997; Steels and Oudeyer, 2000; Oudeyer, 2001; Kirby, 2000). The idea is to regulate the interactions with both the use of a few simple keywords by the human and the robot's ability to express vocally emotions (see section 3). A continuous word-spotting module is implemented in the robot, as available for instance in the Sony AIBO robot or the NEC Papero robot. The robot continuously listen to what humans say, computing the intonation parameters of the sentences they hear, classify them, and react to the detected emotion. For instance, if the robot detects a happy sentence, he utters a happy vocalization itself and modifies its facial expression accordingly (here by changing the colors of the LEDS on its face), or if it hears a sad sentence, it gets sad also (and for calm/neutral sentences, it does nothing special of course). When the robot reacts in an un-

appropriate manner, then the human has to say a sentence with a key-word (or equivalent key-word) which was "bad guess" in our experiments. Then he has to say a sentence containing a key-word designating what was his intended emotion (for example "I was angry !'). Then the robot stores the intonation parameters of the last sentence he heard before the one containing the "bad guess", associated with the class corresponding to the second key-word. This gives it an example to put in its database. It is possible to use a key-word, like "well done" in our experiments, which means the robots reacted well to the last sentence, and shall add the intonation parameters of the last sentence to its database. Note that with robots like the Sony AIBO robot, it is possible to replace the "bad guess" and "well done" key-words by the information coming from the gentle/firm tap sensor which is on their head. Initially, the robot has an empty database, and we pre-programmed it to think everything is neutral initially. We used as a learning scheme the 1-nearest neighbour algorithm. In practice, rather robust guesses were possible typically after 6 or 7 examples of each class. To illustrate this mechanism, several videos showing this game between a virtual character (projected on a wall) and a human are available at [6].

6 Conclusion

We have shown how one could generate life-like vocalizations with basic emotions recognizable by people from very different linguistic and cultural background. The algorithm presented has the advantage of being extremely simple (very few parameters need to be controlled) and completely specified.

[6] www.csl.sony.fr/py

We showed that concatenative speech synthesis could be used as successfully as formant synthesis. Further work will concentrate in extending the range of emotions used in this paper. We also presented and validated techniques which allow the continuous control over the degree of emotion as well as ways to control smoothly the age of the voice.

As far as recognition is concerned, 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 literature are not ones preferred by 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 dependent recognition can reach very high scores, if adequate features and learning schemes are used. We also showed that with the right set of features, reasonable performance can be reached when only few examples are given, which might be the case in "real situation" robots. Yet, we have to remain prudent with these results since they were obtained with professional speakers, and we used high quality microphones in quiet environment. The use of microphones embeded in real noisy robots might bring difficulties. The fact that professional speakers might not be so biased since the target of this research is to recognize the emotional information of humans who talk to pet robors. Indeed, in this case they over-emphasize their intonation as professional speakers do (see Breazeal 2001). Also, one has to note that results should be improved if the algorithms presented here are embedded in a complete cognitive robot which can use other cues than intonation (vision, linguistic cues, semantic cues) to decide what is the emotional state of human beings.

This work should serve as a basis for necessary additional experiments with more databases including speakers of very different languages in more realistic settings. The use of only freely available softwares should allow other people who already possess these databases to help to pursue this research.

7 Acknowledgement

I would like to thank Mr. Tanaka and his colleagues at the Sony Digital Creature Lab in Tokyo for providing the databases, and Dr. Doi, President of Sony Computer Science Labs, Inc., and Sony Digital Creature Lab (Tokyo) for his support during this research.

8 References

Abelin A, Allwood J., (2000) Cross-linguistic interpretation of emotional prosody, in Proceedings of the ISCA Workshop on Speech and Emotion.

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

Boersma P. (1993) Accurate Short-Term Analysis of The Fundamental Frequency and the Harmonics-to-Noise Ratio of a Sampled Sound, in Proceedings of the Institute of Phonetic Sciences of the University of Amsterdam, 17, 97-110.

Breazal, C. (2001) Designing Social Robots, MIT Press, Cambridge, MA.

Burkhardt F., Sendlmeier W., (2000) Verification of Acoustical Correlates

of Emotional speech Using Formant-synthesis, in Proceedings of the ISCA Workshop n Speech and Emotion.

Bosh L.T. (2000) Emotions: What is Possible in the ASR framework ?, in Proceedings of the ISCA Workshop on Speech and Emotion.

Cahn J. (1990) The generation of Affect in Synthesized Speech, Journal of the I/O Voice American Society, 8:1-19.

Cauldwell R. (2000) Where did the Anger Go ? The Role of Context in Interpreting Emotions in Speech, ISCA Workshop on Speech and Emotion.

Druin A., Hendler J. (2000) Robots for Kids: Exploring new technologies for learning, Morgan Kauffman Publishers.

Dutoit et al. (2000) Traitement de la Parole, Presses Romandes.

Dutoit T., Leich H., (1993) MBR-PSOLA : Text-to-Speech Synthesis Based on an MBE Re-Synthesis of the Segments Database, Speech Communication.

Edgington M.D., (1997) Investigating the limitations of concatenative speech synthesis, in Proceedings of EuroSpeech'97, Rhode, Greece.

Ekman, P. (1982) Emotions in the human face, Cambridge University Press, Cambridge.

Fujita M., Kitano H. (1998) Development of an autonomous quadruped robot for robot entertainment, Autonomous Robots, 5.

Gobl C., Chasaide A.N. (2000) Testing Affective Correlates of Voice Quality through Analysis and Resynthesis, in Proceedings of the ISCA Workshop on Emotion and Speech.

Goldberg, D.E. (1989) Genetic Algorithms in Search, Optimization and Machine Learning, Reading, MA: Addison-Wesley.

Halliday M. (1975) Learning How to Mean: Explorations in the Development of Language, Elsevier, NY.

IIda A., et al. (2000) A Speech Synthesis System with Emotion for Assisting Communication, ISCA Workshop on Speech and Emotion.

Iriondo I., et al. (2000) Validation of an Acoustical Modelling of Emotional Expression in spanish using Speech Synthesis Techniques, in Proceedings of ISCA workshop on speech and emotion.

Kienast M., Sendlmeier W. (2000) Acoustical Analysis of Spectral and Temporal Changes in Emotional Speech, in Proccedings of the ISCA Workshop on Emotion and Speech.

Koike K., Suzuki H., Saito H. (1998) Prosodic parameters in Emotional Speech, in Proccedings of ICSLP 1998, pp. 679-682.

Kirby, S. (1998), Syntax without natural selection: how compositionnality emerges from vocabulary in a population of learners, in Hurford, J., Studdert-Kennedy M., Knight C. (eds.), Approaches to the evolution of language, Cambridge, Cambridge University Press.

Kusahara M. (2000) The art of creating subjective reality: an analysis of japanese digital pets, in Boudreau E., ed., in Artificial Life 7 Workshop Proceedings, pp. 141-144.

Massaro D., (2000) Multimodal Emotion Perception : Analogous to Speech Processes, ISCA Workshop on Speech and Emotion, Belfast 2000.

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.

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.

Murray I.R., Arnott J.L., (1993) Towards a simulation of emotion in synthetic speech: a review of the literature on human vocal emotion, JASA 93(2), pp. 1097-1108.

Oudeyer P-y. (2001) The origins of syllable systems : an operational model, to appear in the Proceedings of the International Conference on Cognitive Science, COGSCI 2001, Edinburgh, Scotland.

Oudeyer P-y. (2001) Coupled Neural Maps for the Origins of Vowel Systems, to appear in the Proceedings of the International Conference on Artificial Neural Networks, ICANN 2001, Vienna, Austria.

Picard R. (1997) Affective Computing, MIT Press.

Polzin T., Waibel A. (2000) Emotion-sensitive Human-computer Interface, in Proccedings of the ISCA Workshop on Speech and Emotion.

Reeves B., Nass C. (2000) The Media Equation, Cambridge University Press, 1996.

A. Samal, P. Iyengar (1992) Automatic recognition and analysis of human faces and facial expression: A survey. Pattern Recognition, 25(1):65–77.

Shigeno S. (1998) Cultural Similarities and Differences in the Recognition of Audio-Visual Speech Stimuli, in Proceedings of ICSLP 1998.

Slaney M., McRoberts G. (1998) Baby Ears : A Recognition System For Affective Vocalization, in Proceedings of ICASSP 1998.

Steels, L. (1997) The synthetic modelling of language origins. *Evolution of Communication*, 1(1):1-35.

Steels L., Oudeyer P-y. (2000) The cultural evolution of syntactic constraints in phonology, in Bedau, McCaskill, Packard and Rasmussen (eds.), Proceedings of the 7th International Conference on Artificial Life, pp. 382-391, MIT Press.

Tickle A. (2000), English and Japanese Speaker's Emotion Vocalizations and Recognition : A Comparison Highlighting Vowel Quality, ISCA Workshop on Speech and Emotion, Belfast 2000.

Vine D., Sahandi R., (2000) Synthesizing Emotional Speech by Concatenating Multiple Pitch recorded Speech Units, ISCA Workshop on Speech and Emotion, Belfast 2000.

Whiteside S.P. (1998) Simulated emotions : an acoustic study of voice and perturbation measures, in Proceedings of ICSLP 1998, pp. 699-703.

Williams U., Stevens K.N., (1972) Emotions and Speech : some acoustical correlates, JASA 52, 1238-1250.

Witten I., Frank E. (2000) Data Mining, Morgan Kaufflan Publishers.