Available online at www.sciencedirect.com
SCIENCE DIRECT®

ELSEVIER

Int. J. Human-Computer Studies 59 (2003) 157–183

International Journal of
Human-Computer
Studies

www.elsevier.com/locate/ijhcs

# The production and recognition of emotions in speech: features and algorithms

Oudeyer Pierre-Yves

*Sony CSL Paris, 6, rue Amyot, 75005 Paris, France*

Received 28 September 2001; received in revised form 26 November 2002; accepted 30 November 2002

**Abstract**

This paper presents algorithms that allow a robot to express its emotions by modulating the intonation of its voice. 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 continuously control both the age of a synthetic voice and the quantity of emotions that are expressed. Also, we present the first large-scale data mining experiment about the automatic recognition of basic emotions in informal everyday short utterances. We focus on the speaker-dependent 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 very few examples are available. Furthermore, we describe a game to play with a personal robot which facilitates teaching of examples of emotional utterances in a natural and rather unconstrained manner.

*Keywords:* Emotions; Speech; Robots; Emotion production; Emotion 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

*E-mail address:* py@csl.sony.fr (O. Pierre-Yves).
*URL:* http://www.csl.sony.fr/py.

1071-5819/03/$ - see front matter © 2002 Elsevier Science Ltd. All rights reserved.
doi:10.1016/S1071-5819(02)00141-6

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 keyboards or dialog windows, and need to know how computers work to be able to use them. In contrast, 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 to express their own emotions. Indeed, emotions are not only 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 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 it 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 Iyengar, 1992), research dealing with the speech modality, both for automated production and recognition by machines, has only been active for very few years (Bosh, 2000). In this paper, we present the results of our research which provides robots with the means to express emotions vocally and to enable them to recognize basic emotional information in its caretaker's voice. Both of these research aspects are original. Unlike most existing work in production, we work with cartoon-like meaningless speech which has different needs and constraints than trying to produce naturally sounding adult-like normal emotional speech. For example we would like the emotions to be recognized by subjects of different cultural or linguistic background. Our work has similarities to Breazeal (2001), but we use concatenative speech synthesis and our algorithm is simpler and totally specified. As far as the recognition of emotions is concerned, we present here a large-scale data mining experiment in which we compare most of the standard machine learning algorithms and explore the value of 200 different features. As shown below, we found some new features which seem to be more efficient than the ones traditionally used in the literature. All the work presented here is based on the use of freely available software and thus can be reproduced with minor difficulties. A web site[1] containing some accompanying material such as sounds and graphs is also available.

The next section presents general information about the acoustic correlates of emotion in speech, which form the basis of our work. Section 3 presents our

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

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 number of researchers have already investigated this question (Banse and Sherer, 1996; Burkhardt and Sendlmeier, 2000). 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 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, which results in slow, low-pitched speech with little high-frequency energy (Breazeal, 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 were produced: (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 in the recognition 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 (Burkhardt 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 performance of the machine in relation to human performance. The fact that humans are not so good is mainly because 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 (Williams and Stevens, 1972; Murray and Arnott, 1993; Banse and Sherer, 1996; Burkhardt and Sendlmeier, 2000). Basically they 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 the existing work in synthetic emotional speech. Whereas traditionally (see Cahn, 1990; Iriondo and Gauss, 2000; Edgington, 1997; IIda et al., 2000) the aim is to produce adult-like naturally occurring 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 because we use this in experiments with robots to which we try to teach language: the use of intonation as a pre-linguistic ability 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, subjects from very different linguistic and cultural background should be able to recognize easily the creature's emotions.

Additionally, we want to have algorithms as simple as possible and to control as few parameters as possible: 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 Leich, 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 minimal 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 and comfort.

### 3.2. Existing work

As said above, previously existing work has concentrated on adult-like naturally sounding emotional speech, and most projects have tackled 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 modeling many co-articulation effects occurring in emotional speech. The drawbacks of formant synthesis are that quality of the produced speech remains unsatisfying (voices are still quite unnatural). Furthermore, the algorithms developed in this case are 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, 2001) 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 (e.g. 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). Also, Breazeal's work used formant synthesis, which does not correspond to our constraints (we cannot control as many parameters as she could, due to the use of concatenative speech synthesis).

Because of their very superior quality, concatenative speech synthesizers (Dutoit and Leich, 1993) have gained popularity in recent years, and have been used in some cases. 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 constraints 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 pre-recorded segments to be concatenated in the synthesis. This gives satisfying results but is not practical if one wants to change the voice or add new emotions or even control the degree of emotions. The second approach consists of

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

(see for example Edgington, 1997) making databases of human produced emotional speech, then compute the pitch and intensity contours and then apply them to sentences to be generated. This brings some problems of alignments, partially solved using syntactic similarities between sentences. 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 presented here is completely generative (it does not rely on the recording of human speech that would serve as input), and it uses concatenative speech synthesis as a basis. We will show that it allows to express emotions as efficiently as with formant synthesis, but with more simple 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, for the pitch we specify only one target per phoneme. 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 format looks like:

```
l 448 80 150 ;; means: phoneme “l” duration 448 ms,
             ;; try to reach 180 Hz at 80 percent of the phoneme duration
9~ 557 80 208
b 131 80 179
@ 77 80 200
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 80 200
```

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 information 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):

```
Choose the number of words of the sentence
      (random number between 2 and MAXWORDS);
Create the words:
 For each word, choose the number of syllables
             (random number between 2 and MAXSYLL), and
             decides with probability PROBACCENT whether
             the word is accented or not;
If the word is accented then choose randomly one
         of its syllables and mark it as accented;
Create the syllables :
 For each syllable
     choose whether this is a CV or a CCV syllable
                (CV syllable have probability 0.8);
 instantiate the C's and V by picking randomly a
                consonnant or vowel in the phoneme database;
 set the duration of each phoneme to MEANDUR + random(DURVAR);
 let e = MEANPITCH + random(PITCHVAR)
     set the pitch of consonnants to e - PITCHVAR
     set the pitch of vowels to e + PITCHVAR
     if the syllable is accented then
     add DURVAR to the duration of its phonemes;
     if DEFAULTCONTOUR = rising
         set the pitch of consonants to MAXPITCH - PITCHVAR
         set the pitch of the vowel to MAXPITCH + PITCHVAR
     if DEFAULTCONTOUR = falling
         set the pitch of consonants to MAXPITCH + PITCHVAR
         set the pitch of the vowel to MAXPITCH - PITCHVAR
     if DEFAULTCONTOUR = stable
         set the pitch of phonemes to MAXPITCH

Change the contour of the last word:
if not LASTWORDACCENTED
     let e = PITCHVAR/2
         if CONTOURLASTWORD = FALLING
         for each syllable in word
             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 easier expression of 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. The stochastic parts are crucial in the algorithm: on the one hand, it produces, for a given set of parameters, a different utterance each time (mainly because of 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 lines 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 30 000 or 35 000 Hz as compared to the 16 000 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 chosen.

Table 1
Parameter values for different emotions

| | 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 | 2 | |

Now that we have described in details the algorithm, let us give (see Table 1) examples of the parameters' values obtained for five affects: calm, anger, sadness, happiness, comfort. We obtained these parameters first by looking at studies describing the acoustic correlates of each emotion (e.g. Murray and Arnott, 1993; Burkhardt and Sendlmeier, 2000), then we deduced some initial value for the parameters and modified them by hand,by trial and error until it gave a satisfying result. Theses results give us a model of intonation contours for the different emotions:

- *Happiness*: The mean pitch of the utterance is high, has a high variance, the rhythm is rather fast, few syllables are accented, the last word is accented, and the contours of all syllables are rising.
- *Anger*: The mean pitch is high, has a high variance, the rhythm is fast, with little variance of phoneme durations, a lot of syllables are accented, the last word is not accented, the pitch contours of all syllables are falling.
- *Sadness*: The mean pitch is low, has a low variance, the rhythm is slow, with high variance of phoneme durations, very few syllables are accented, the last word is not accented, the contours of all syllables are falling.
- *Comfort*: The mean pitch is high but less than happiness, the rhythm is slow, with a high variance of phoneme durations, very few syllables are accented, the last syllable is accented, and the contours of syllables are rising.

One has to keep in mind that these models that we built certainly do not correspond to any existing human system of prosody variation to express emotion. This is because our goal was to build cartoon stylized emotional speech that can be recognized by subjects speaking different languages. These models predict the human response to this kind of speech, but not how humans express themselves emotionally. We think that if the goal is to build a system which makes human react emotionally, then one should not try to exactly copy humans but rather invent speech that we know is spoken by a nonhuman creature. Indeed, trying to copy humans explicitly provokes some expectations from users. These expectations are a drawback since they will not be matched quite often (because the task is just to hard for the machine), and lead to user frustration. The consequence of this approach is that solving the problem of cartoon emotional speech generation will not necessarily help us to solve the problem of recognition of emotions in human speech. As we will see in the second part of this paper, data mining techniques use features quite different to recognize emotions in human speech than the one we control to generate cartoon emotional speech.

### *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 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 chosen, and they were asked to make a choice between "Calm", "Anger", "Sadness", "Comfort" and "Happiness". They could hear each example only once. In a second experiment with different subjects, they were initially given four 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 words cited above. Eight naive adult subjects were present in each experiment: three French subjects, one English subject, one German subject, one Brazilian subject, and two 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 Table 2, we see that 76 percent of vocalizations intended to represent sadness were effectively perceived as sadness (Table 3).

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 nonsense

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

Table 2
Confusion matrix for the unsupervised series

| | Calm | Anger | Sadness | Comfort | Happiness |
|---|---|---|---|---|---|
| Calm | **38** | 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 3
Confusion matrix for the supervised series

| | 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** |

emotional speech, at best humans have 60 percent success, and most often less. Here we see that the mean rate of correct classification is 57 percent, which compares well to human performance. The errors are of two types: the most frequent one is related to 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 minimal confusion 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 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 exaggerated. The results presented here are similar to those reported in Breazeal (2001), which strongly suggests that using a concatenative synthesizer with a lot less parameters still allows conveyance of emotions (and in general provides more life-like sounds).

Examination of the supervised serie shows that with only a few vocalizations for their intended emotion, results improve very much: a 77 percent success rate is achieved. We see that confusions involving the neutral emotion and confusions between anger and happiness have nearly disappeared. Similarly, we made an experiment where the calm affect was removed, 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 combinations of color LED lights to express their "emotions". The present experiment shows that it would be enough to see visually the robot express emotions a few times while it is uttering emotional sentences, and then be able to recognize its intended emotion later just by listening to it.

### *3.5. Varying continuously the age of the voice and the degree of emotion*

Typically, robotic pets are initially "babies". They 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 previously been addressed in the literature. Generally, one has several voice databases to choose from (for the segments used by the speech concatenizer), 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. This is not acceptable in our case. On the other hand, only a limited number of databases are available (because it is not practical to make a lot of them and requires a lot of memory), which means that age cannot vary smoothly.

We found a solution to this problem (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 32 000 Hz (used in the validation experiment in the next section): if we want to make the signal sound "older", then we can override the sample rate (e.g. 28 000 Hz). We then use PSOLA to shorten the signal back to its original time length.

Furthermore, it would be useful if robotic pets could 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 corresponding to the "normal" degree of emotion (those described in the previous section), an associated set of similar parameters corresponding to the highest degree of emotion for a given emotion (e.g. very very happy). For example, to the parameter MEANPITCH (= 400) of (normal) happiness, we add a parameter MEANPITCH2 (= 50); then we define a variable *delta* taking values in $[-1; 1]$ which determines the degree of one emotion: 0 is normal, 1 maximum and $-1$ 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. In fact, we are making a local linear models of emotion degrees. This experimentally gives satisfying results, and requires the specification of 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 is older than the other. The re-sampling frequencies were taken in the range [25 000; 35 000] 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.

To evaluate the recognition of the degree of emotion, the same test was repeated except that the age was fixed (32 000 Hz), and pairs consisted of two utterances of the same emotion (but each time random), with a different random degree. Human subjects had to evaluate which utterance expressed a higher degree the emotion. Fifty pairs were presented to each subject. The mean rate of correct ranking was 85.1 percent, which is again satisfying.

## 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 use all the context and modalities, from lexic to facial expression and intonation. Unfortunately, this 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 current systems. 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 could go by using only prosodic information in the 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 a paragraph with emotions from 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 (Bosh, 2000). The first studies (e.g. Murray and Arnott, 1993; Williams and Stevens, 1972) did not try 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). Yet, to our knowledge, no large-scale study using the modern tools developed in the data mining and machine learning community has been conducted. Indeed, most often, either only one or two learning schemes are tested (e.g. Polzin and Waibel, 2000; Slaney and McRoberts, 1998; Breazeal, 2001) or very few and simple features are used (Polzin and Waibel, 2000; Slaney and McRoberts, 1998; Breazeal, 2001; Whiteside, 1998), or only small databases are used—less than 100 examples per speaker (as in Breazeal, 2001; McGilloway et al., 2000; Slaney and McRoberts, 1998) which means that the power of some statistical learning schemes may have been overlooked.

Only (McGilloway et al., 2000), using more than the traditional/standard set of features from the rest of the literature (mean, max, min, max–min, variance of the pitch and intensity distributions, lengths of phonemic or syllabic segments, or pitch rising segments), have tried using data mining. This work has some drawbacks: (1) only three kinds of learning schemes were used—support vector machines, gaussian mixtures and linear discriminants—which are far from being the best to deal with data in which there are possibly many irrelevant features. In particular this does not allow automatic derivation smaller set of features with optimal efficiency; (2) the feature set was explored by choosing one learning scheme and by 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 detection of the fitness of groups of features. Finally, their work is based on speech databases made by asking human subjects to read newspaper texts in an emotional manner. This does not correspond to our constraints. To our knowledge, only two research groups have tried to build automatic recognition machines of everyday speech (Breazeal, 2001; Slaney and McRoberts, 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 four basic emotions, seem impossible if there are several speakers. The great speaker variability has been described in (Slaney and McRoberts, 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 (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 et al., 2000) wherein 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 standard 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 and McRoberts (1998) or Breazeal (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/

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

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 themselves uttering 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, 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 up 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 has had the opportunity to have so many examples, even for one speaker, and to use the power of modern statistical learning algorithms. Another potential drawback of the database is that there might be a self-entertainment 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*

As per the work reported above, we collected two main measures related to intonation—pitch and intensity. For each signal, we also measured the intensity of its low- and high-passed version, the cutting frequency being chosen at 250 Hz (the particular value does not appear to be crucial). Finally, for the sake of exhaustivity, we made a spectral measure consisting of 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 0.01 s 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 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 four series: the series of its minima, the series of its maxima, the series of the durations between local extrema of the 10 Hz smoothed curve (which models rythmic 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 $5 \times 4 \times 10 = 200$ features.

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

### 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 not often equivalent: some are more efficient with certain types of class distributions than others, some are better at dealing with many irrelevant features (which is the case here, as seen a posteriori) or with structured feature sets (in which it 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 schemes, i.e. AdaBoostM1 (Witten and Frank, 2000), which allows generally significant 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 provides means like automatic cross-validation, or the search of feature spaces (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).

Table 4
Learning schemes

| Name | Description |
|---|---|
| 1-NN | 1 Nearest neighbor |
| 5-NN | Voted 2 nearest neighbors |
| 10-NN | Voted 10 nearest neighbors |
| 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 | Classification via M5PRime regression method |
| Naive Bayes | Naive Bayes classification algorithm |
| AdaBoostM1/C4.5 | Adaboosted version of C4.5 |
| AdaboostM1/PART | Adaboosted version of PART |

### 5.4.3. All features/all algorithms

For the 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-dependent recognition, with only three classes (but they were different than ours: approval, prohibition, attention). The best way to compare the different techniques 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 neighbors, 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 solving a problem about which we have few prior or intuitive knowledge. It is not surprising that the best results are obtained with decision trees and rules since these kinds of

Table 5
Using all features

| Name | Mean correct generalization rate across six 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 |

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 would like to naturally 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 4 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, Fig. 1 plots the database with axis being the 1st quartile and the 3rd quartile of the intensity distribution, and Fig. 2 being 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° 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.

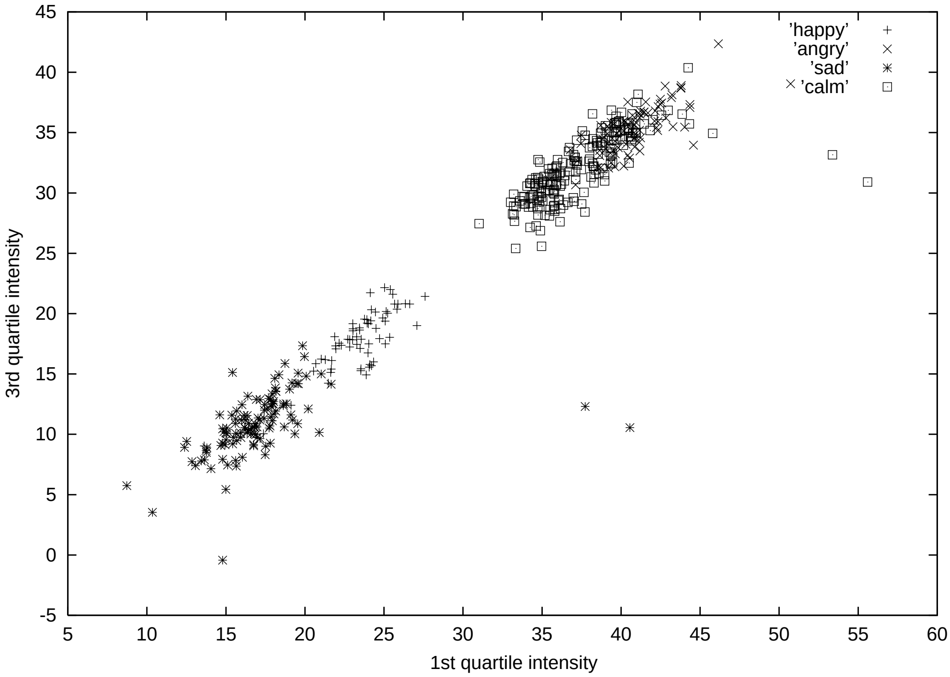

Fig. 1. Data points in speaker 1 database: 1st quartile of intensity distribution against 3rd quartile of intensity distribution.

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 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, 1995; Burkhardt and Sendlmeier, 2000; Williams and Stevens, 1972) or used in most application oriented research (Slaney and McRoberts, 1998; Breazeal, 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 1–6 related to low-passed signal intensity (LPF), with a feature set composed of the standard features (SF) used in (Breazeal, 2001) or (Slaney and McRoberts, 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 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 to when we use standard features, and the best result is obtained with the low-passed intensity related features (85.9 percent).

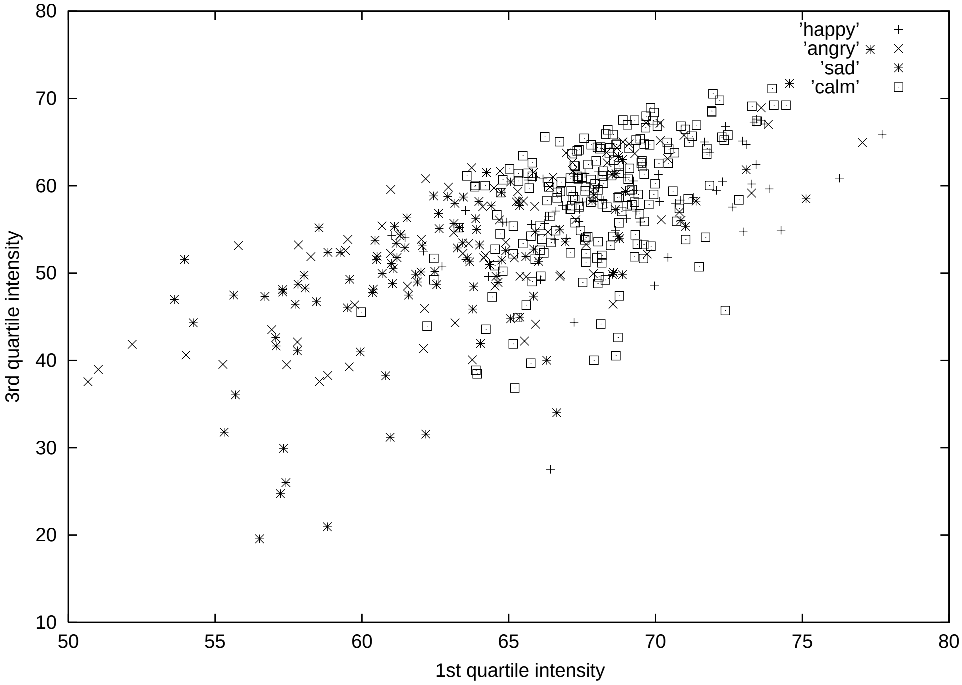

Fig. 2. The same data points from last figure, except that we plot here the 1st quartile of the low-passed signal intensity distribution against 3rd quartile of the low-passed signal intensity distribution.

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 in 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 two algorithms: Naive Bayes and 5-nearest neighbors (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, it was no surprise that 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 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 high results like the initial set but with more than 10 times less features. Moreover, and interestingly, the variation between learning schemes is less important and algorithms which performed badly like

Table 6
Information gain of 20 best features

| Feature | Information gain (mean across six 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 7
Comparing "standard" features and "low-passed signal intensity" features

| 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 8
Using the "optimal" feature set

| Name | Correct generalization rate (mean for six 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 |

nearest neighbors or Naive Bayes, behave now in a more satisfying manner (yet, for these two, this is not surprising since the feature 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 probably what happens with human babies and real pets, but humans tend to be more willing to spend 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 algorithm 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 neighbors or Naive Bayes. These results are rather comparable (and in fact slightly superior) to what is

Table 9
When very few training examples are provided

| Learning scheme | Mean across six 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 |

described in Breazeal (2001, except that in this case, learning was off-line with a larger database of several female speakers). What is important is that Breazeal 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 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 two 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 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 with a small game, in the spirit of language games used in robotic models of language acquisition (Steels, 1997; Steels and Oudeyer, 2000; Oudeyer, 2001a,b; Kirby, 1998). 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 emotions vocally (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 is continuously listening to what humans say, computing the intonation parameters of the sentences they hear, classifying them, and reacting to the detected emotion. For instance, if the robot detects a happy sentence, he utters a happy vocalization himself and modifies his 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 unappropriate 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 (e.g. "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 the 1-nearest neighbor algorithm learning scheme. In practice, rather robust guesses were possible, typically after six or seven examples of each class.

## 6. Conclusion

We have shown how one could generate life-like vocalizations with basic emotions recognizable by people from very different linguistic and cultural backgrounds. The algorithm presented has the advantage of being extremely simple (very few parameters need to be controlled) and completely specified. 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 large-scale modern data mining techniques allowed to find nonobvious features which were missed by precedent studies. In particular, it is interesting to see that the features put forward in the psychoacoustic literature are not the ones preferred by the machine learning algorithms. As precedent studies seemed to show that multi-speaker emotion recognition was a very difficult task in principle, the present work suggests 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 a quiet environment. The use of microphones embedded in real noisy robots might bring difficulties. The fact that speakers were professionals might not be so misleading since the target of this research is to recognize the emotional information of humans who talk to pet robots. 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 pursue this research.

## Acknowledgements

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.

## References

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

Banse, R., 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, Vol. 17. pp. 97–110.

Breazeal, 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 on 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, Los Altos, CA.

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 (1), 7–18.

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. Addison-Wesley, Reading, MA.

Halliday, M., 1975. Learning How to Mean: Explorations in the Development of Language. Elsevier, New York.

IIda, A., Campbell, N., Iga, S., Higuchi, F., Yassemana, M., 2000. A speech synthesis system with emotion for assisting communication. ISCA Workshop on Speech and Emotion.

Iriondo, I., Gauss, R., 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: Proceedings of the ISCA Workshop on Emotion and Speech.

Kirby, S., 1998. Syntax without natural selection: how compositionality emerges from vocabulary in a population of learners. In: Hurford, J., Studdert-Kennedy, M., Knight, C. (Eds.), Approaches to the Evolution of Language. Cambridge University Press, Cambridge.

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. In: ISCA Workshop on Speech and Emotion, Belfast, 2000.

McGilloway, S., Cowie, R., Cowie, E.D., Gielen, S., Westerdijk, M., Stroeve, S., 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. Int. J. Speech Communication 16 (4), 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), 1097–1108.

Oudeyer, P-y., 2001a. The origins of syllable systems : an operational model. In: Proceedings of the International Conference on Cognitive Science, COGSCI 2001. L. Erlbaum Edinburgh, Scotland, pp. 744–749.

Oudeyer, P-y., 2001b. Coupled neural maps for the origins of vowel systems. Proceedings of the International Conference on Artificial Neural Networks, ICANN 2001. Springer-Verlag, LNCS2130 Vienna, Austria.

Picard, R., 1997. Affective Computing. MIT Press, Cambridge, MA.

Polzin, T., Waibel, A., 2000. Emotion-sensitive Human–computer interface. In: Proceedings of the ISCA Workshop on Speech and Emotion.

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

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: Proceedings of the Seventh International Conference on Artificial Life. MIT Press, Cambridge, MA, pp. 382–391.

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.

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, Los Altos, CA.