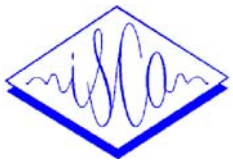

Speech Prosody 2002
Aix-en-Provence, France
April 11-13, 2002

The Synthesis of Cartoon Emotional Speech

Pierre-yves Oudeyer

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

Abstract

Recent years have been marked by the increasing development of personal robots such as small pets or humanoids, often having young and cartoon like personalities. A key feature they currently lack is the ability to speak in a emotional life-like manner. We present here a technology that makes this possible by using concatenative speech synthesis.

1. Introduction

Recent years have been marked by the increasing development of personal robots, either used as new educational technologies or for pure entertainment. Typically, these robots look like familiar pets such as dogs or cats (e.g. the Sony AIBO robot), or sometimes take the shape of young children such as the humanoids SDR3-X (Sony).

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

In this paper, we present the result of our research for means to express emotions vocally for a baby-like robot. Unlike most of existing work, we are dealing with cartoon-like meaningless speech, which has different needs and different constraints than 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 ([2]), but we use concatenative speech synthesis and our algorithm is simpler and completely specified. 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.

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 ([1]). 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 ([8]) 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 ([2]).

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 ([9]) who made experiments in which American people had to try 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 people were asked to try 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 of trying to detect 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 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 utter nonsense utterances, which is quite unnatural, but is confirmed by studies asking people to utter semantically neutral but meaningful sentences). The first result indicates that our goal of making 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 a perfect result, 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 state 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 MacGurk effect for emotions and that different contexts may lead people to interpret the same intonation as expressing different emotions for each context. These findings indicate that we shall not try to have our machine generate utterances that make fine distinctions; only the most basic affects should be investigated.

A number of experiments using computer based techniques

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

10.21437/SpeechProsody.2002-123

of sound manipulation have been conducted to explore which particular aspects of speech reflect emotions with most saliency. ([1], [11]) 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.

3. The generation of cartoon emotional speech

3.1. Goal

The goal we had in this research is quite different from most of existing work in synthetic emotional speech. Whereas traditionally (see [3], [5], [6]) the aim is to produce adult-like naturally occuring emotional speech, here the target was to provide a young creature with the ability to express its emotions in an exagerated/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 had to sound lively, not repetitive, and similar to infants babbling. Finally, we were willing that people from very different linguistic and cultural background could recognize easily the emotions of the creature.

Additionally, we wanted to have algorithms as simple as possible and to control as few parameters as possible: in brief, what is the minimum that allows 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 as a basis a concatenative speech synthesizer ([4]), 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 so far only five emotional states so far, corresponding to calm and one for each of the four regions defined by the two dimensions of arouseness 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 [3]) 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 not natural). Furthermore, the algorithms developped in this case are complicated and necessitate the control of many parameters, which renders their fine tuning quite impractical (see [3] for a discussion). Unlike these works, ([2]) has described a system which is very similar to ours: based on the work of ([3]), she made a system for her robot Kismet that allows it to produce meaningless emotional speech. Unfortunately, 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 as the basis. 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 Breazal'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 ([4]) have gained popularity in the recent years, and some have tried to use them to produce emotional speech. This is a challenge and 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 for example described in ([6]) 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 quite impractical if one wants to change the voice or add new emotions or even control the degree of emotions. The second approach consists (see for example [5]) in 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. Anyway, ([5]) showed that this method gave quite unsatisfying results (speech ends 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 exagerated/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 a more life-like signal quality.

3.3. A simple and complete algorithm

Our algorithm will consist 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 will only control the overall volume of sentences. Our program generates a file like in figure 2 which is fed into the MBROLA speech synthesizer.

```
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 5 200 300 300 80 378.0 100 401
```

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

The idea of the algorithm is to generate first 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 this sentence's pitch and duration informations are altered so as to yield a particular affect. Deformations consist in deciding that a number of syllables become stressed, and apply a certain stress contour on these syllables as well as some duration modifications. Also, all syllables are applyed 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 wether 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
                        (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 done 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 accent 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 exagerate 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).

	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

Now that we have described in details the algorithm, let us give (see table 1) examples of values of the parameters 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, then deducing some coherent initial value for the parameters and modifying them by hand, and trial and error until it gave a satisfaying result. Evaluation of the quality is given in next section.

3.4. Validation with human subjects

In order to evaluate the algorithm described in the precedent sections, 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 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 experiments 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)

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

	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 serie

	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 serie

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 ([9]) in which humans were asked to produce nonsence emotional speech, that 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. If we look closer at the results, we discover that the errors are most of the time not “bad” errors, especially about the degree of arouseness in the speech: happy is confused most often with anger (both are aroused), and calm is confused most often with sad and comfort (they are not aroused). In fact, less than 5 percent of errors are made about degree of arouseness. Finally, one can observe that many errors involve the calm/neutral affect. This led to a second unsupervised experiment, 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 exagerated. The results presented here are similar to those reported in ([2]) 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. Again the few errors are not “bad”. 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.

4. 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. We showed that concatenative speech synthesis could be used as successfully as formant synthesis. Further work will concentrate in extending the range of emotions spanned by this experiment.

5. 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 Lab and Sony Digital Creature Lab for his support during this research.

6. References

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

[2] Breazal C. 2000. Sociable machines: expressive social exchanges between humans and robots, PhD thesis, MIT AI Lab.

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

[4] Dutoit et al. 2000. Traitement de la Parole, Presses Romandes.

[5] Edgington M.D. 1997. Investigating the limitations of concatenative speech synthesis, in Proceedings of EuroSpeech’97, Rhode, Greece.

[6] IIda A., et al. 2000. A Speech Synthesis System with Emotion for Assisting Communication, ISCA Workshop on Speech and Emotion.

[7] Koike K.; Suzuki H.; Saito H. 1998. Prosodic parameters in Emotional Speech, in Proccedings of ICSLP 1998, pp. 679-682.

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

[9] 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.

[10] Vine D.; Sahandi R. 2000. Synthesizing Emotional Speech by Concatenating Multiple Pitch recorded Speech Units, ISCA Workshop on Speech and Emotion, Belfast 2000.

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