Expressive Synthesis of Read Aloud Tales

Virginia Francisco and **Pablo Gervás** and **Mónica González** and **Carlos León** [1]

Abstract.

An important challenge for text-to-speech is to get a synthesized voice that sounds as similar as possible to human voice. However, nowadays the voice generated by synthesizers sounds artificial and this is the main cause of rejection by users. In this paper we propose a solution for modeling emotions in the FESTIVAL synthesizer by controlling the parameters of the system. We have chosen Fairy Tales as the domain for the synthesized text, because emotions play a fundamental role in the speech of such stories. We also present an evaluation process for the resulting voices of our prototype, and we show the results we have obtained in our first experiments, as well as the conclusions for those results.

1 Introduction

Often it is not possible to read text written on a screen. Some users who might not be able to access a particular textual system (children, or blind people) could access the information stored in a computer, if it was spoken, and all users may experiment a better user experience. However, nowadays most of the information in the computers is stored as text, and this impedes the retrieval of the content.

From this point of view, translating from written text to phonetical sounds that can be listened and understood by humans can be very useful. This process is known as *text-to-speech* (TTS), but it still presents many problems. One of the main challenges for *text-to-speech* is to get a synthesized voice that sounds as similar as possible to the human voice. Nowadays the voice generated by synthesizers sounds artificial. This is the main cause of rejection of this kind of systems by humans. To make the TTS systems more *user friendly* and, in this way, more useful for people, we have to generate voices that can express emotions, just like humans do.

There is much information in the way we speak that is not present in written text. It is very important for the generation of emotional voice to generate clear emotions, so that there will be no confusion for the listener. However, there is an important lack of emotions in the usual *text-to-speech* systems. In some domains, this might not be very crucial, but when narrating a fairy tale, for instance, synthesized voice must express emotions, because these emotions carry much information that should not be ignored.

This project arises to explore the possibility of modeling emotions through control parameters in an existing synthesizer when reading tales aloud. There are many theories which try to define emotional scales, and the choice of a specific scale determines the emotions that we try to distinguish. Another important challenge is to analyse the acoustic characteristics of human voice production at different emotional states in order to try to reproduce the same characteristics in the synthesizers.

To obtain the parameters that must be passed to the synthesizing system, we have to carry out an analysis of the emotional components of significant chunks of audio to create a model of that emotional speech. Once we have this model, the next task is to test the results.

The work presented in this paper extends previous work carried out with a different synthesizer [10]. The main goal of the present work is to change the synthesizer engine used in the previous system for another one which generates a more natural voice and allows us to control more voice parameters than the previous synthesizer. In the Conclusions of this paper we compare the results obtained in the present work with the results of the previous one.

2 Previous Work

The first studies about emotional speech were written by Fairbanks and Pronovost [9]. Even though this line of work gave rise to a great amount research and published articles, there is still a lot of important aspects to cover. The complexity of affective speech starts with the concept of emotion. Nowadays there are many theories about emotions, each of them with a different interest. Sometimes these theories are contradictory and it is difficult to integrate all of them in a single one.

Research on expressive synthesized speech has been carried out by Cahn[4], Murray and Arnott [18] for the English language, Burkhardt for German [2], Mozziconacci [17] for French and Montero [15] for Spanish.

2.1 Meaning of the Word "Emotion"

Emotions are defined as a flexible mechanism for the adaptation to a changing environment [21]. There are mainly two types of emotions [6]:

- Extreme emotions: This term denotes an emotion fully developed, which is intense and incorporates most of the aspects which are considered relevant in a emotion.
- Underlying emotions: It denotes the type of emotional colouring which is part of most of the mental states.

2.2 Clasification of Emotions

For the study of emotional speech we need to decide which emotions we are going to model, and how we are going to represent them. There are different methods in order to represent emotions [6]:

- *Emotional categories*. It is the most common method for the description of emotions. The method of Emotional categories uses

[1] Departamento de Inteligencia Artificial e Ingenieria del Software, Universidad Complutense de Madrid,Spain, email: virginia@fdi.ucm.es, pgervas@sip.ucm.es, monica.glez.jenal@gmail.com, cleon@fis.ucm.es

emotion-denoting words, or category labels for indicating emotions. Several approaches have been proposed in the literature for reducing the number of emotion-denoting adjectives:

  - *Basic emotions*. There is general agreement that some full-blown emotions are more basic than others. The number of basic emotions is usually small so it is possible to characterize each emotional category in terms of its intrinsic properties [7].
  - *Super ordinate emotion categories*. Some emotional categories have been proposed as more fundamental than others on the grounds that they include the others. Scherer [22] and Ortony suggest that an emotion A is more fundamental than another emotion B if the set of evaluation components of the emotion A are a subset of the evaluation components of the emotion B. Cowie and Cornellius [7] give a short overview of recent proposals of such lists.
  - *Essential everyday emotion terms*. A pragmatic approach is to ask for the emotion terms that play an important role in everyday life. The approach is exemplified by the work of Cowie [8], who proposed a Basic English Emotion Vocabulary.

- *Descriptions based on psychology*. The appraisal of a stimulus determines the significance of stimulus for the individual, and triggers an emotion as an appropriate response [1].
- *Descriptions based on evaluation*. Emotions are described from the point of view of the evaluations involved [19].
- *Circumflex models*. Emotional concepts are represented by means of a circular structure [20] such that two emotional categories being close in the circle represents the conceptual similarity of these two categories.
- *Emotional dimensions*. Emotional dimensions [6] represent the essential aspects of emotion concepts. Evaluation (positive/negative) and activation (active/passive) are the main dimensions; sometimes they are augmented with the power dimension (dominant/submissive). This approach is very useful because it allows measurement of the similarity between different emotional states. Another important property of this method is the relative arbitrarity in naming the dimensions.

2.3 Obtaining Prosodic Rules for Emotions

There are a lot of researches to obtain the prosodic rules which take part in the generation of emotional voice. These rules are obtained in different ways:

- Extracting it from the existing literature [3, 18].
- Analizing a corpus [16].
- Obtaining the optimum values from the systematic variation of the parameters in the synthesis [2, 17].

In the present work we are going to combine these three methods in order to obtain better results in the hope that the weaknesses of each individual approach are reduced by their combination.

2.4 Data Sources for Emotional Voice

The identification of the prosody associated to each emotion must be obtained empirically. There are different sources that have been used in the past in order to generate an emotional voice data base:

- *Actors*. The oldest and the most frequently used technique is to obtain recorded data from actors. The main advantage of that method it is that all the emotions can be reproduced using the same sentence [17] or the same pseudo-sentence composed of words with no sense [14]. This way the phonetics, prosody and voice quality can be compared in the same sentence with different emotions. Another advantage of this method is the facility of obtaining extreme emotions. A disadvantage of this technique is that the actor can reproduce a stereotype of the emotion which do not correspond with the emotion obtained spontaneously.
- *Expressive reading of emotional material*. It is a variant of the previous method suggested by Nick Campbell [5]. Campbell proposed to have readers that read texts with an appropriate verbal content with the emotion which is expected to be transmitted.
- *Production of emotions*. Subjects are urged to cause an emotion by means of the so-called MIPS (*Mood Induction Procedures*) [11].
- *Natural occurrences*. A research of Klaus Scherer, Bob Ladd and Kim Silverman [22] deals with the spontaneous generation of emotions.

Each one of these methods varies with respect to the control on the voice signal, from more to less control. These methods can be ordered in the following way: *actors*, *expressive reading of emotional material*, *production of emotions* and *natural occurrences*. Each of these is better or worse depending on the domain of the study. For the researching of extreme emotions the most appropriate is the use of *actors*. On the other hand for the researching of underlying emotions the best method is the observation of *natural occurrences*. In the case of studies centered on the speaker, the best choice is the *production of emotions*.

2.5 Prosody and Emotions

In all researches the global parameters of the prosody, like the base frequency, the scale of the base frequency and the speech rate, are treated like universals, at least when the number of emotional categories is small. The most interesting acoustic variables for voice synthesis are the ones that can be controlled through a voice synthesis system.

For modeling a system able to generate an emotional voice it is necessary to have a correspondence between the emotions and the values of the characteristics of the voice.

2.6 PRAAT

PRAAT[2] is a free, stable, scriptable and user-friendly scientific software program, designed and continuously developed by Paul Boersma and David Weenink at the Institute of Phonetic Sciences of the University of Amsterdam. It can run on a great variety of operating systems and allows to perform a great variety of tasks, which is why it is used in a wide range of situations such as phonetics classes, pronunciation improvement teaching and emotional voice synthesis research.

PRAAT does not only allow speech analysis but also speech synthesis, including articulatory synthesis. It can be used to manipulate speech as well as to create high-quality representations that show the parameters of the analyzed voice. These outputs can be spectrograms, intensity contours or even pitch and formants graphics. PRAAT's seemingly endless possibilities also include functions for learning algorithms, segmentation, labeling and listening experiments, filters, sound recording and a lot of other functionalities that are continually expanded by its users. This is why PRAAT is among

[2] http://www.fon.hum.uva.nl/praat

the most popular free downloadable speech analysis software packages and the reason why we chose to use it for our research.

For the present work the main advantage of PRAAT is the generation of high quality graphs in which pitch, spectrogram, intensity, formants, record pulses ... can be visualized.From these graphics and analyzing different records we can establish how to change the voice characteristics in order to express emotions.

2.7 FESTIVAL

The synthesizer employed for our emotional story teller is FESTIVAL 3. FESTIVAL is a speech synthesis system that offers full text-to-speech through a number of APIs, such as the Scheme API, the Shell API, the Server/client API, the C/C++ API and the Java and JSAPI. It uses the UniSyn residual excited LPC diphone synthesizer, the CMU lexicon and letter-to-sound rules trained from it. The intonation was trained from the Boston University FMRadio corpus and the duration for this voice also comes from that database 4. It is multilingual and includes many voices. For our research we chose to use the default voice *kal diphone*, which is an American English male speaker.

The system is written in C++ and uses the Edinburgh Speech Tools for low level architecture and has a Scheme (SIOD) based command interpreter for control that we used to transform our SABLE marked up texts into audio files. We employed the FESTIVAL 1.95-beta version of the system and made it run on Cygwin[3], which is a Linux-like environment for Windows.

2.8 SABLE

SABLE[4] is an XML (Extensible Markup Language)/SGML (Standard Generalized Markup Language) based markup scheme for text-to-speech synthesis. It was developed to address the need for a common, system-independent TTS control paradigm. The aim of the Sable Consortium is to merge the STML (Spoken Text Markup Language) standard, developed by Bell Labs and the Edinburgh University, and Sun's JSML (Java Speech Markup Language). There are different groups involved or interested in this project, such as the Edinburgh University, Bell Laboratories, AT&T, Sun Microsystems and the Carnegie Mellon University. The 0.2 version of the Sable specification was released in March 1998 and FESTIVAL contains a basic implementation of it in its standard distribution since its 1.3.0 version. Although we found that not all tags have been implemented yet in the FESTIVAL 1.95-beta version we used for this research, the specification has been a useful guideline. The set of text description and speaker directives tags we finally used to mark up our texts are a subset of those implemented by the FESTIVAL 1.95-beta version that allowed us to modify the voice parameters our previous research efforts had proved to be relevant.

2.9 Evaluation Paradigms

There are several paradigms of evaluation for the emotional voice, the three most used are:

- *Forced choice*: This type of evaluation has been used in a lot of researches of generation of emotional voice [3, 15, 17]. The procedure is to give to evaluators a finite set of possible answers which includes all the emotions that have been modeled. The advantages of this approach are that it is easy to carry out, it provides a simple recognition measurement and it allows to compare different researches. On the other hand it has a disadvantage because it does not provide information about the quality of the stimulus from the point of view of naturalness and veracity.
- *Free choice*: The answer it is not restricted to a close set of emotions [18, 23]. It is very useful when the aim of the evaluation is to find unexpected phenomena.
- *Free choice modified*: Murray and Arnot [18] and then Stallo [25] introduced some modifications to the previous paradigm: introduced distraction categories, the "others" category, neutral texts with emotional texts. The difference between the recognition of the neutral text and the emotional text is taken as a measurement of the impact of prosody in the perception.

2.10 The Previous System: EMOSPEECH 0.1

The details of the previous system can be consulted in [10]. This system used FREETTS[5] as synthesizer, which is a voice synthesizer engine written entirely in Java, based on FLITE[6], and derived from FESTIVAL and FESTVOX[7]. FREETTS allows variations in the following voice parameters: pitch, pitch range, volume and rate.

It does not allow modifications of the parameters half way through a sentence, nor different assignations of parameters to different part of a sentence. This was found to be an important disadvantage for further work. For this reason, we have developed a new system, which will be explained in later sections.

3 Our Proposal

Fairy tale narration has been chosen as the domain of the application, because it is considered to be an environment where emotions clearly take part in the communication effort. Tales try to summarize the emotions that most of the children experiment in their way to maturity: happiness, sadness, anger, fear, envy... When reading a tale, one tends to exaggerate. The voice of the person or persons who read the tale will be as important a tool as the words themselves for a child to infer the emotions of the characters. Therefore we can affirm that the emotions expressed in a tale are extreme emotions, not underlying ones.

3.1 Emotions in the Fairy Tale Narration

In order to explain the voice markers which make our tale lively and personalized, we are going to use Scherer's research [21]. By means of the personality markers the speaker externalize some characteristics and the listener perceives it and assigns these characteristics to the speaker.

The classification of the emotions expressed in speech that satisfies best the requirements of a fairy tale teller is the basic emotions classification, because when a tale is being told, we usually exaggerate the emotions, so a small set of extreme emotions is enough. We have selected five basic categories in order to model the emotions: *happiness*, *sadness*, *fear*, *anger* and *surprise*.

In the tales synthesized by our story-teller there is only a narrator speaking and there are no dialogues. So we have decided that emotions are related to fragments of the tale and we have selected the sentences as emotional units. The narrator will try to impress the

[3] http://www.cygwin.com/
[4] http://www.cstr.ed.ac.uk/research/projects/sable/sable_spec2.html
[5] http://freetts.sourceforge.net
[6] http://www.speech.cs.cmu.edu/flite/
[7] http://www.festvox.org

emotion with which the sentence has been tagged, so that his voice will transmit sadness when he is reading a sad sentence and happiness when reading a happy one.

3.2 Tales Marked Up with Emotions

The input of our system are tales marked up with basic emotions. In order to get tales tagged at the same time as we generate them, an existing module for automatic story generation [12] has been modified. This module generated a conceptual representation of fairy tales and its corresponding text by means of natural language generation techniques. The input of the module are the actions which take part in the story plot and the semantic information about characters, locations, attributes and relations involved in the actions. From this input the story is generated automatically.

The marking up of tales in our generator is carried out in the *lexicalization* stage of the natural language generation process, where it is decided which specific words and phrases should be chosen to express the domain concepts and relations which appear in the messages. Given the basic linguistic structures used by the generation module, the mark up is done by phrases. The result of the *lexicalization* stage is a list of messages with their correspondent lexical forms and the emotion they are going to be marked up with. A final stage of *surface realization* assembles all the relevant pieces into linguistically and typographically correct text.

Two elements of the tales are taken into account when deciding the emotion associated to each sentence: characters and actions in which the characters are involved.

3.2.1 Emotions Associated to Characters

Using the traditional distinction between good and evil, the characters in our stories are supposed to be involved in good, bad or neutral situations. For each case, one of the basic emotions is associated to the character. For the tale "Cinderella" the emotions in Table 1 have been considered for the main characters.

	Good	**Bad**	**Neutral**
Cinderella	Happy	Sad	Neutral
prince	Happy	Sad	Neutral
father	Neutral	Neutral	Neutral
mother	Neutral	Neutral	Neutral
stepmother	Angry	Happy	Neutral
stepsisters	Angry	Happy	Neutral

Table 1. Emotions associated to characters

As they are the villains of the tale, for the "stepmother" and "stepsisters" the emotion assigned for the good situations is *angry*. For the hero and victim, the assignment is just the opposite.

3.2.2 Emotions Associated to Actions

The actions are considered as good, bad or neutral situations. When choosing the emotion associated to the message representing the action, the characters involved in it are taken into account. There is a type of action that must be treated in a special way. These are the surprising actions, that are always assigned the *surprise* emotion, not taking into account their arguments. The information about the type of action is specified in the story plan received by the generation module as input.

3.3 Voice Parameters and Emotions

In order to obtain the parameters of the prosody we have analyzed recorded material of tales read by actors. This is how we have identified the relation between the parameters of the voice and the different emotions. We have used actors because we are going to deal with extreme emotions and the employment of actors is the best choice when the aim of the research are extreme emotions.

We have used PRAAT in order to analyze the tales read by actors. With PRAAT we have obtained the pitch base line, the pitch range and the rate related to each of the emotions which take part in the tale.

The aspects of the voice that act as emotional identifiers are: pitch, volume, voice quality and rate. For this research we have assumed that the voice aspects that are necessary for modeling the different emotions are: pitch baseline, pitch range, volume and rate, we have not used the voice quality. The answer to the question "Is voice quality fundamental for the generation of voice with emotions?´´ is not unanimous. We have assumed that the voice quality is not fundamental for the generation of emotional voice.

To obtain the values of these parameters for every emotion, we have consulted Scherer [21] and the results of the analysis of emotional material generated by actors. Finally we have obtained the optimal values through the systematic variation of the parameters during synthesis. Table 2 shows the rules of the synthesizer for the basic emotions.

	Volume	**Rate**	**Pitch Baseline**	**Pitch Range**
Anger	$+10\%$	$+21\%$	$+0\%$	$+173\%$
Surprise	$+10\%$	$+0\%$	$+25\%$	$+82\%$
Happiness	$+12\%$	$+25\%$	$+35\%$	$+27\%$
Sadness	-30%	-10%	-12%	-40%
Fear	$+10\%$	$+12.5\%$	$+75\%$	$+118\%$

Table 2. Configuration Parameters for Emotional Voice Synthesis

3.4 EMOSPEECH 0.2

In the new version of our system, EMOSPEECH 0.2, we have changed some design and implementation aspects. We have used SABLE as a language for the control of the TTS engine, this way the configuration of the different voice parameters for the expression of emotions is independent of the synthesizer engine use. In the previous system we have no control language, we made all the changes directly in FreeTTS.

The input file of our system is a XML file in which every sentence is marked up with the five basic emotions or with the neutral emotion. The file is generated automatically with the modified system for automatic story generation mention before. A sample part of a marked tale is given in Table 3.

Based on the XML file we generate a SABLE file. In order to make this transformation we apply the rules for the synthesizer, which we have obtained from the Scherer researches, the analysis of emotional material and the systematic variation of the voice parameters, to the XML file. For each of the sentences of the previous tale

```
...
<Neutral> Gretel ate the window. </Neutral>
<Surprise> The witch came out of the house. </Surprise>
<Fear> The witch locked Hansel up. </Fear>
<Neutral> Brave Hansel was locked
          in the cage. </Neutral>
<Angry> Hansel made Gretel work
        very hard. </Angry>
<Surprise> She tricked the witch. </Surprise>
<Neutral> The witch was locked in the oven. </Neutral>
<Neutral> Gretel released Hansel. </Neutral>
<Neutral> She was pretty. </Neutral>
<Surprise> Hansel came out of the cage. </Surprise>
<Happy> Pretty Gretel found the treasure. </Happy>

...
```

Table 3. Fragment of a Marked Up Tale

we apply these rules and we generated automatically the SABLE file given in Table 4.

```
...
<MARKER MARK="Neutral">
Gretel ate the window.
</MARKER>

<MARKER MARK="Surprise"> <VOLUME LEVEL="+10%">
<RATE SPEED="+0%"> <PITCH BASE="+25%" RANGE="+82%">
The witch came out of the house.
</PITCH> </RATE> </VOLUME> </MARKER>

<MARKER MARK="Fear"> <VOLUME LEVEL="+10%">
<RATE SPEED="+12.5%"> <PITCH BASE="+75%" RANGE="+118%">
The witch locked Hansel up.
</PITCH> </RATE> </VOLUME> </MARKER>

<MARKER MARK="Neutral">
Brave Hansel was locked in the cage.
</MARKER>

<MARKER MARK="Angry"> <VOLUME LEVEL="+10%">
<RATE SPEED="+21%"> <PITCH BASE="+0%" RANGE="+173%">
Hansel made Gretel work very hard.
</PITCH> </RATE> </VOLUME> </MARKER>

<MARKER MARK="Surprise"> <VOLUME LEVEL="+10%">
<RATE SPEED="+0%"> <PITCH BASE="+25%" RANGE="+82%">
She tricked the witch.
</PITCH> </RATE> </VOLUME> </MARKER>

<MARKER MARK="Neutral">
The witch was locked in the oven.
</MARKER>
...
```

Table 4. Fragment of SABLE file obtained from the mark up tale.

The SABLE file is processed by the FESTIVAL text-to-speech system which returns an audio file in which the tale is read aloud taking into account the different emotions marked in the input text.

4 Evaluation

It is not easy to evaluate this kind of systems, because there is not "good" or "bad" output. In order to evaluate our work we carried out two different tests with two different types of audio files. Fifteen evaluators have taken part in this experiment.

With our evaluation we have tried to measure two main aspects: how well the emotions modeled are recognized by the evaluators and how much the meaning of the reading text influences the emotional recognition.

In order to obtain these measurements we have made some distinctions about the texts that are going to take part and the type of tests that are going to be performed.

The texts that have taken part in the evaluation are two:

- *"Hansel and Gretel"* tale. We have selected this tale because it is a tale that has all the modeled emotions and because it is the tale with most emotional sentences from all the tales generated.
- Sentences without emotional content read aloud with each of the five modeled emotions were reproduced by PRAAT without articulating any word, and these intonated sentences were marked up by the evaluators.

So we measured the two main aspects commented before: on the one hand how well the the evaluators recognized the emotions modeled and on the other hand how much the meaning of the reading text influences the emotional recognition.

We have carried out two types of tests:

- *Test of free choice*: Evaluators can assign to each of the sentences any emotion they consider that best suits the voice they are listening.
- *Test of force choice*: Evaluators have to choose one of the six modeled emotions (five basic emotions and the neutral emotion).

We have made this distinction in order to determine if the emotions are only well distinguished among the five basic emotions or are well distinguished among all the range of emotions.

4.1 Free Choice Results

The graph in Figure 1 shows the percentage of sentences marked with the correct emotion, in each of the tests carried out with the two types of audio files (tale and intonated sentences), grouped by emotions. This figure seems to indicate that the emotions with a high success percentage are Neutral (54%), Fear (46%) and Sad (45%) in the case of *"Hansel and Gretel"* tale and Surprise (38%) in the case of the intonated sentences.

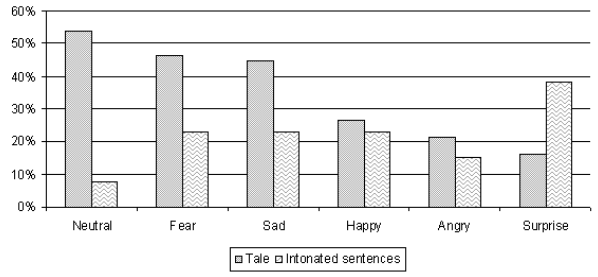

Figure 1. Percentage of sentences marked up with the correct emotion in the free choice tests.

In Tables 5 and 6 the main confusions can be seen.

	Neutral	**Sad**	**Surpr.**	**Worry**	**Hysteria**	**Bored**
Angry	×					
Fear				×	×	
Happy			×			
Neutral		×				
Sad	×					×
Surprise	×					

Table 5. Confusion between emotions in the tale in free choice test.

	Ne.	Sa.	Su.	Ca.	An.	Bo.	Ex.	Hy.
Angry	×			×			×	
Fear			×		×			×
Happy			×		×			
Neutral		×		×		×		
Sad					×	×		
Surprise	×	×				×		

Table 6. Confusion between emotions in the intonated sentences in free choice test. The headings correspond, from left to right, to: Neutral (Ne), Sad (Sa), Surprise (Su), Calm (Ca), Anger (An), Bored (Bo), Excited (Ex), Hysteria (Hy)

The graphs in Figures 2 and 3 show the percentage of sentences marked up with an emotion different from the one that the synthesizer is trying to express in the case of the *"Hansel and Gretel"* tale and the intonated sentences.

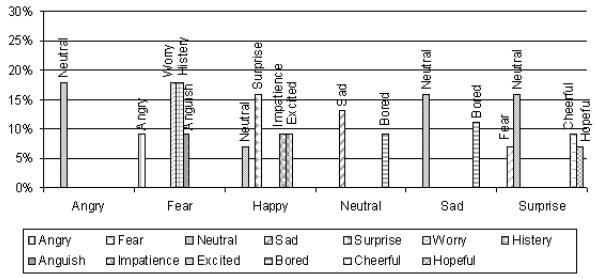

Figure 2. Percentage of sentences marked up with a wrong emotion in the Tale group by emotions.

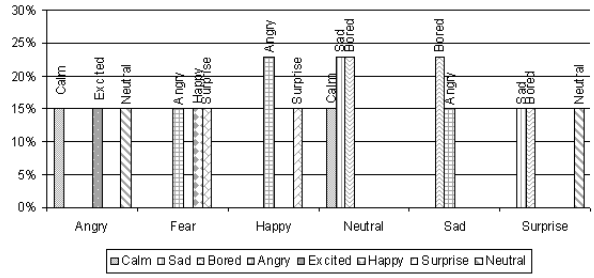

Figure 3. Percentage of sentences marked up with a wrong emotion in the Intonated Sentences group by emotions.

4.2 Force Choice Results

The graph in Figure 4 shows the percentage of sentences marked with the correct emotion. In each of the tests carried out with the two types of texts (tale, and intonated sentences), grouped by emotions. This figure seems to indicate that the emotion with a high success percentage are Sad (77%, 69%), Neutral (65%, 69%) and Fear (64%, 54%) in all the tests.

The graphs in Figures 5 and 6 show the percentage of sentences marked up with a different emotion from the one that the synthesizer is trying to express in the case of the tale and the intonated sentences.

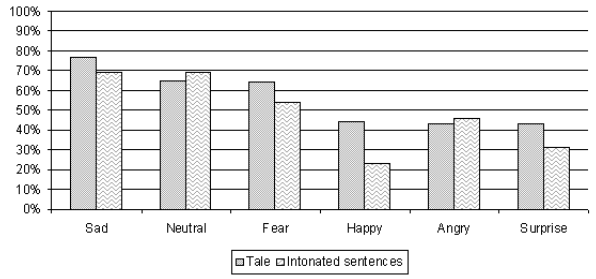

Figure 4. Percentage of sentences marked up with the correct emotion in the force choice tests grouped by emotions.

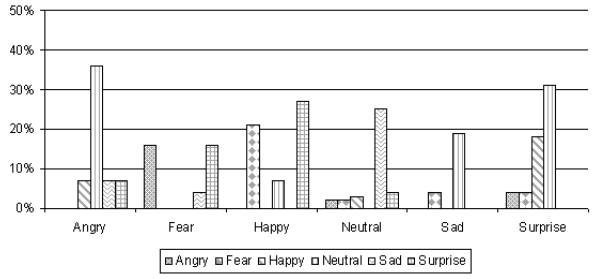

Figure 5. Percentage of sentences marked up with a wrong emotion group by emotions.

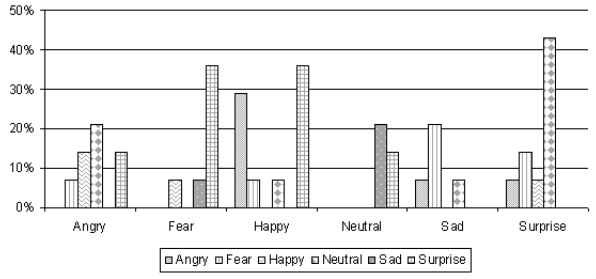

Figure 6. Percentage of sentences marked up with a wrong emotion group by emotions.

In Tables 7 and 8 the main confusions can be seen .

4.3 Conclusions of the Tests

4.3.1 Free Choice

In the case of intonated sentences Neutral is the emotion less recognized. That is because the voice base is very serious and has a low pitch base and pitch range so it confused mainly with the emotion which has these characteristics. Happy has more or less the same results in the two audio files, so we can conclude that the meaning of the text does not influence in this emotion. Surprise has better results in the case of the tale, so we can conclude that the meaning of the text influences in this emotion.

Angry is confused with the excited emotion. Excited is a type of anger, so we can consider that the angry sentences confused with ex-

	Angry	Fear	Happy	Neutral	Sad	Surprise
Angry				×		
Fear	×					×
Happy		×				×
Neutral					×	
Sad				×		
Surprise			×	×		

Table 7. Confusion between emotions in the tale in force choice test.

	Angry	Fear	Happy	Neutral	Sad	Surprise
Angry			×	×		×
Fear						×
Happy	×					×
Neutral					×	×
Sad		×				
Surprise		×		×		

Table 8. Confusion between emotions in the tale in force choice test.

cited are correct. I this way the percentage of correct angry sentences increases to 30%. Fear is confused with worry and hysteria emotions. These two emotions are types of fear so we can consider these sentences as correctly marked and increase the percentage of fear sentences correctly marked to 57%. Happy is confused with Surprise emotion. A surprise can be good or bad. In the case of good surprises the result is a happy emotion. The same occurs with the surprise sentences confussed with the sad emotion. Neutral sentences are confused with sad, calm and bored emotions. This indicates how the base voice is perceived by the evaluators.

If we compare the results of the tales and the intonated sentences in terms of confusion with other emotions, we can see that in both cases the following confusion are presented:

- Angry - Neutral
- Neutral - Sad.
- Surprise - Neutral.
- Happy - Surprise.
- Sad - Bored.

4.3.2 Forced choice

If we compare the percentage of sentences correctly recognized in the tale and the intonated sentences we can see that the percentage decreases in the case of happy and surprise. We can conclude that in these two emotions the meaning influence in a good way. The three emotions more recognized (Sad, Neutral and Fear) are common in the two cases and they are the same as the more recognized in the tale of the free choice test.

If we compare the results of the tales and the intonated sentences in terms of confusion with other emotions, we can see that in both cases the following confusion are presented:

- Angry - Neutral
- Neutral - Sad
- Surprise - Neutral
- Happy - Surprise
- Fear - Surprise

In the case of the sad sentences we can see that they are confused with neutral sentences only in the tale, which indicates that the meaning of the sentences plays a main role. The same occurs with the fear sentences which are confused with angry sentences only in the tale's test.

There are no sentences confused with the happy emotion.

4.3.3 General

If we compare the results of the two types of tests (Free choice and Force choice) we can conclude that the following confusion are present in both cases:

- Angry - Neutral.
- Neutral - Sad.
- Surprise - Neutral.
- Happy - Surprise.

5 Conclusions and Future Work

Expressive characters need an intuitive and simple interface which makes the interaction with the user easy. Communication through the voice is the best solution for this problem. Nowadays the voice generated by synthesizers sounds artificial and this is the main cause of rejection by the users. The success of expressive characters in the everyday life depends on the overcoming of this rejection. In order to obtain a lively synthesizer it is important to generate voice with different emotional states. The generation of emotional voice tries to get emotions clear enough to avoid confusion in the listener.

In this fist approach Sad and Neutral are highly recognized, around a 70% of sentences are recognized, Fear has around a 55% of sentences correctly recognized. Happy and surprise need to be improved because they have a low percentage of recognition, around 30%. These results confirm the ones obtained by [26], [24] and [13]; in general, emotions which can be considered negative are better recognized than emotions which can be considered positive. This is particulary true with the happy emotion, which is the worse recognized in the whole research.

If we compare this approach with a previous one [10], based on the FREETTS synthesizer, we can conclude that:

- As in the previous approach the results obtained in the case of the tale are better than those with the intonated sentences. Is in both cases meaning influences in a positive way the recognition of emotions.
- Surprise and Happy are the emotions less recognized in both approaches.
- Fear is better recognized in the present approach with a 55% of sentences well recognized, as opposed to the 50% of the previous approach.
- Sad is better recognized in the previous approach in which the percentage of sad sentences correctly tagged is around a 100% against the 70% of the previous approach.

We need to explore the new parameters that can be modified with FESTIVAL in order to improve the results. We have to explore the characteristics of the sad emotion in the previous approach in order to apply these characteristics to the previous approach and return to the 100% percentage of success.

There is much work that has to be done, and we are working on different approaches. In future versions we will consider a finer granularity for emotional units. We are considering the use of shallow

parsing techniques to determine the different blocks of the sentences, and assigning different emotions to each of this blocks.

We also plan to use the knowledge acquired about the use of emotions in the generation of narrations. In this way, we expect to create more interesting stories, ready to be expressed with emotions.

ACKNOWLEDGEMENTS

This research is funded by the Spanish Ministry of Education and Science (TIN2006-14433-C02-01 project), Complutense University of Madrid and the G.D. of Universities and Research of the Community of Madrid (UCM-CAM-910494 research group grant).

REFERENCES

[1] K. Alter, E. Rank, S.A. Kotz, U. Toepel, M. Besson, A. Schirmer, and A.D. Friederici, 'Accentuation and emotions - two different systems?', in *Proceedings of the ISCA Workshop on Speech and Emotion*, pp. 138–142, Northern Ireland, (2000).

[2] F. Burkhardt, *Simulation emotionaler Sprechweise mit Sprachsyntheseverfahren*, Ph.D. dissertation, TU Berlin, 2000.

[3] J. Cahn, 'The generation of affect in synthesized speech', *Journal of the American Voice I/O Society*, (July 1990).

[4] J.E. Cahn, 'Generation of affect in synthesized speech', in *Proceedings of the 1989 Conference of the American Voice I/O Society.*, pp. 251–256, (1989).

[5] W.N. Campbell, 'Databases of emotional speech.', in *ESCA Workshop on Speech and Emotion.*, pp. 34–37, Belfast, (2000).

[6] R. Cowie and R.R. Cornelius, 'Describing the emotional states that are expressed in speech', in *Speech Communication Special Issue on Speech an Emotion*, pp. 5–32, (2003).

[7] R. Cowie and R.R. Cornelius, 'Describing the emotional states that are expressed in speech', in *Speech Communication Special Issue on Speech and Emotion*, (2003).

[8] R. Cowie, E. Douglas-Cowie, and A. Romano, 'Changing emotional tone in dialogue and its prosodic correlates', in *Proc ESCA International Workshop on Dialogue and prosody*, Veldhoven, The Netherlands, (1999).

[9] G. Fairbanks and W. Pronovost, *An experimental study of the pitch characteristics of the voice during the expression of emotion.*, 87–104, Speech Monograph, 1939.

[10] Gervás P. Hervás R. Francisco, V., 'Analisis y síntesis de expresión emocional en cuentos leídos en voz alta', in *In Proceedings of Sociedad Española para el Procesamiento del Lenguaje Natural.*, volume 35, (2006).

[11] A. Gerrards-Hesse, K. Spies, and F. W. Hesse, 'Experimental inductions of emotional states and their effectiveness: A review.', *British Journal of Psychology.*, **85**, 55–78, (1994).

[12] P. Gervás, B. Díaz-Agudo, F. Peinado, and R. Hervás, 'Story plot generation based on CBR', in *12th Conference on Applications and Innovations in Intelligent Systems*, eds., Anne Macintosh, Richard Ellis, and Tony Allen, Cambridge, UK, (2004). Springer, WICS series.

[13] M. Guidetti, 'L'expression vocale des émotions: approche interculturelle et développementale.', in *L'Année Psychologique*, pp. 383–396, (1991).

[14] Lea Leinonen, Tapio Hiltunen, Ilkka Linnankoski, and Maija L. Laakso, 'Expression of emotional-motivational connotations with a one-word utterance', *J Acoust Soc Am*, **102**(3), 1853–1863, (September 1997).

[15] J.M. Montero, *Estrategias para la mejora de la naturalidad y la incorporación de variedad emocional a la conversión texto a voz en castellano*, Ph.D. dissertation, Escuela técnica superior de ingenieros. Universidad Politécnica de Madrid., 2003.

[16] J.M. Montero, J. Gutiérrez-Ariola, S. Palazuelos, E. Enríquez, S. Aguilera, and J. M. Pardo, 'Emotional speech synthesis: From speech database to tts.', in *In Proceedings of the 5th International Conference of Spoken Language Processing*, volume 3, pp. 923–926, Sydney, Australia, (1998).

[17] S. J. L. Mozziconacci, *Speech Variability and Emotion:Production and Perception*, Ph.D. dissertation, Technical University Eindhoven, 1998.

[18] I.R. Murray and Arnott J.L., 'Implementation and testing of a system for producing emotion-by-rule in synthetic speech', *Speech Commun.*, **16**(4), 369–390, (1995).

[19] A. Ortony, G. L. Clore, and A. Collins, *The Cognitive Structure of Emotion.*, Cambridge University Press.

[20] J.A. Russell, 'A circumflex model of affect', *Journal of Personality and Social Psychology*, **39**, 1161–1178, (1980).

[21] K. R. Scherer, *Personality markers in speech*, Cambridge University Press, Cambridge, 1979.

[22] K.R. Scherer, *On the nature and function of emotion: A component process approach*, Scherer and K.R. and Ekman P and editors, Erlbaum, Hillsdale, NJ, 1984.

[23] Marc Schröder, 'Can emotions be synthesized without controlling voice quality?', *PHONUS*, **4**, 37–55, (1999).

[24] H. Siegwart, 'The differential perception of linguistic and emotional prosody: A neuropsychological study.', in *Colloque CERE (Coordination Européenne des Recherches sur l'Emotion)*, Paris, (1990).

[25] J. Stallo, *Simulating emotional speech for a talking head.*, Ph.D. dissertation, School of Computing, Curtin University of Technology., 2000.

[26] H.G Wallbott and K.R. Scherer, 'Cues and channels in emotion recognition.', *Journal of Personality and Social Psychology*, **51**, 690–699, (1986).