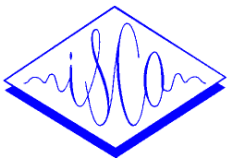

ITRW on
Speech and Emotion
Newcastle, Northern Ireland, UK
September 5-7, 2000

VALIDATION OF AN ACOUSTICAL MODELLING OF EMOTIONAL EXPRESSION IN SPANISH USING SPEECH SYNTHESIS TECHNIQUES

Ignasi Iriondo[†], Roger Guaus[†], Angel Rodríguez[††], Patricia Lázaro[††], Norminanda Montoya[††], Josep M[a] Blanco[††], Dolors Bernadas[††], Josep Manel Oliver[††], Daniel Tena[††] and Ludovico Longhi[††]

[†]Department of Communications and Signal Theory. La Salle School of Engineering. Universitat Ramon Llull. Barcelona (Spain)

[††]Department of Audio-Visual Communication and Publicity. Universidad Autonoma de Barcelona. (Spain)

ABSTRACT

This paper describes the methodology used for validating the results obtained in a study about acoustical modelling of emotional expression in Castilian Spanish.

We have obtained a set of rules that describes the behaviour of the most significant parameters of speech related with the emotional expression.

The validation of the results of the study has been achieved by the use of synthetic speech that has been generated following the different rules we have obtained for each emotion.

1. INTRODUCTION

Synthesised speech is mainly distinguished by a lower intelligibility, a not natural prosody and lack of expressiveness. These are important drawbacks for computer human speech communication.

Our work comprises the systematic study of speech with emotional expression to model the effects of emotion on speech. Next step has been the validation of the model using speech synthesis techniques. The implementation of this model will improve the naturalness of voice in text to speech systems.

In Section 2 of the paper, we explain the methodology to model acoustically the emotional expression in Castilian Spanish. Also, the values of the most relevant parameters for each emotion are shown.

Section 3 explains the process we are followed to validate the obtained results in the previous study. We have used synthetic speech matching the most outstanding acoustical parameters for each studied emotion. There are two different ways to generate emotional synthesised speech. The first methodology is based on a text-to-speech converter that can generate speech from carrier sentences. The emotional expression in this speech is achieved modifying its prosody appropiately. Also, the obtained results with this method are shown in this section. The other methodology uses recorded speech with neutral prosody to turn it into emotional speech with the corresponding prosody.

In the last section, we expose the conclusions and the future steps to do.

2. MODELLING EMOTIONAL EXPRESSION OF SPEECH

2.1 Introduction to the modelling

In the study of emotion in speech, it is supposed the hypothesis that voice suffers acoustical changes caused directly by the physiological alterations of the human body when a person has a strong feeling **[1, 2, 3]**. These changes also depend on the used language. In spite of this supposition, we think that is convenient to do a study without difference between linguistic and no linguistic processes, considering emotional speech like a united system that comprises simultaneously the cultural influence of the language and the physiological mechanism of emotion. On the other hand, we have considered that emotions suffer a dynamic evolution in the time with a variable duration **[4]** and the acoustical features that determinate each emotional state have not to fit into the own characters of the language. Then, to do the study, we have analysed supra-segmental voiced patterns, because they hold jointly the features of the language and the acoustical features depending of each emotional state.

The number of parameters to characterise acoustically an emotion can be extensive, therefore we have to use different acoustic parameters (fundamental frequency, tone contour, duration, rhythm, spectrum, etc). From our point of view, the analysis of short speech segments is not sufficient to find the specific acoustic features of emotions and the influence of a language.

Therefore, an efficient modelling of emotional expression of speech has to be based on full discourses.

2.2 Methodology

The study comprises different perceptual tests and the analysis of emotional speech. With the purpose of building a speech corpus with all the emotions, we recorded 336 discourses. These discourses were recorded by eight actors who simulated seven basic emotions: joy, desire, fury, fear, surprise, sadness and disgust **[5]**. These oral discourses were judged by 1054 experimental people who made a perceptual test. The best valued 34 discourses were selected to be analysed and modelled.

To sum up, the perceptual tests were used to select the discourses that surely had the emotional acoustic patterns we were looking for.

Building the corpus

The speech corpus was built by the interpretation of two carrier texts by 8 actors (4 men and 4 women) simulating the seven 'basic' emotions. Each text was repeated with 3 degrees of emotional intensity. Therefore, 336 discourses were recorded (2 texts x 8 actors x 7 emotions x 3 intensities).

The greater part of studies about emotions agree that there is a reduced number of emotions, but they disagree about the exact number and the words to name them. This lack of definition is an added difficulty to obtain homogeneous results, as from the point of view of the actor when has to interpret a emotion, as from the point of view of the listener when has to judge it.

The corpus was built in the audio studies of Faculty of Communication Science of the Universidad Autonoma de Barcelona (UAB), using a recording equipment that was calibrated in the same conditions for all sessions.

Developing the perception tests

Next step was the building of the definitive corpus to start the acoustic analysis of emotional speech. A perceptual test was carried out to choose the most representative interpretations of each emotion. Each emotional discourse (with duration from 20 to 40 ms) was listened by two groups of more than 30 people. Each group valued 30 interpretations answering three questions: 1) to mark which emotion or emotions you recognise in each speech 2) to mark a level of credibility of the speaker 3) to specify if you had felt emotion and in which degree. A total of 1054 listeners participated in the test. Almost all the listeners were students.

This test was used to decide objectively which interpretations really had acoustic information about emotions. The best four or five valued interpretations of each emotion were chosen following the highest percentage of identification and the highest level of credibility.

Acoustic analysis

A protocol has been designed to do a systematic study of the 34 chosen discourses. This protocol analyses these parameters:

1) Fundamental Frequency (F0): mean, range and variability
2) Sound Pressure: mean, range and variability
3) Timing Parameters: complete time of the discourse, not silent time, time of silences, number of phonic groups, number of silences, duration of each phonic group, duration of each silence, average duration of phonic groups, average duration of pauses, percentage pause/ loudness, number of syllables per second.

In addition to these parameters, two graphical representations were calculated:

1) Total representation of the discourse that includes the waveform view, pitch contour and sound pressure level display.
2) Partial representation from 1 to 3 phonic groups with the same displays than the total representation.

Results

Once the acoustic analysis has been done, we have to define a measurement to compare the different values of the analysed parameters. Therefore, we have defined a measure, named average state, as the mean of the data for each parameter of each speaker. Therefore, it has been established a method to measure the acoustical deviations of different voices with regard to their average state. This measure is used to compare voices form different speakers.

It has been observed that emotional speech has in general these features:

- The characteristic prosodic structure of emotion (pitch and energy contours) is able to appear only in some phonic groups of the discourse.
- This partial structure is sufficient to identify an emotional state.
- The timing structure related to an emotion tends to appear along the whole discourse.
- The intonation in Spanish is characterised by a sawtooth shape with peaks in stressed vowels and valleys in the other ones. The variability of this structure depends on the emotional state.

With reference to the acoustical modelling for each emotion, we have obtained valid results in six of the seven studied basic emotions. Next, the most important features of each acoustical-emotional modelling are summed up:

1) Joy
   - Increase of the average tone (10 - 50 %)
   - Increase of tone variability (120%)
   - Fast inflections of tone
   - Stable intensity
   - Decrease of the duration of silences (20%)
2) Desire
   - Decrease of the average tone (10%)
   - Decrease of tone variability (5-10%)
   - Slow inflections of tone
   - Regular fall of the intensity (max 25 dB)
   - Strong exhalation at the end of each phonic group

- Decrease of the duration of phonic groups (10-30%)
- Increase of the discourse fragmentation (20%) and the total time

3) Fury

- Variation of the intonation structure (20-80 Hz)
- Raising intensity from the begin to the end (5-10 dB)
- Reduction of the number of silences (25%)
- Increase of the duration of silences (10%) and the total time of the discourse
- The variation of the timbre seems the most characteristic feature of fury. Increase of energy in 500-636 Hz and 2000-2500 bandwidths (10-15 dB)

4) Fear

- Increase of the average tone (5-10%)
- Decrease of tone variability (5 %)
- Raising intensity (10 dB)
- Decrease of the duration of phonic groups (20-25%)
- Decrease of the duration of silences (10%)

5) Surprise

- Increase of the average tone (10-15%)
- Increase of tone variability (15-35 %)
- High inflections of intonation
- Increase of the average intensity (3-5 dB)
- Decrease of the duration of phonic groups (10%)

6) Sadness

- Decrease of the average tone (10-30%)
- Decrease of tone variability (30-50%)
- Null inflections of intonation
- Decrease of the average intensity (10-25%)
- Increase of the discourse fragmentation (10%)
- Increase of the duration of silences (50-100%)

7) Disgust

- It has not been able to be modelled because this emotion was only identified by 50 percent of the listeners.

3. VALIDATING THE MODEL USING SYNTHETIC SPEECH

In the described study we have proposed a new methodology to model the speech variability depending on emotional expression that comprises three different steps:

- Construction of a corpus with emotional speech
- Validation and selection of the most representative discourses using perception tests.
- Acoustical analysis of these discourses and modelling of emotions

Then, the results will be integrated in a speech synthesis system to validate definitely these results.

In this section, we describe the used methodology to generate synthetic emotional speech and its later validation.

There are two possibilities to generate synthesised speech with emotional prosody:

- The first method is based on a text to speech (TTS) converter. Emotional discourses are generated by concatenative synthesis of speech units which prosodic parameters may be previously calculated to match the desired emotion. Also, it is important to have a prosody edition tool to adjust these parameters.
- The second method consists of a prosody modification of 'neutral' discourses through speech analysis and synthesis techniques. This method will be useful to modify only the prosody of a discourse and to turn into an emotional discourse.

In this study, the first method has been tested, but the second one has hardly used. Our next work will be to test this second method. In spite of this, we explain the theoretic basis and the obtained results with the first method in subsections 3.1 and 3.2, and a brief description of the second one in subsection 3.3.

3.1 Text-to-speech conversion

A text-to-speech system can generate any oral message from text. Speech synthesis is achieved through concatenation of speech units that previously have been recorded, analysed and stored adequately. First, phonetic transcription and prosodic information are obtained from text. Both kinds of information are the input of the synthesis block, where speech signal will be generated using speech unit concatenation and prosodic modification.

To do this work, it has been used the computer tool EMOVS, developed in the Communications and Signal Theory Department of Enginyeria La Salle (Universitat Ramon Llull). Some features of this tool are:

- Text-to-speech conversion based on TD-PSOLA **[6]** with automatic calculation of prosody without emotional expression.

- Edition of the energy and pitch contours and the duration of phones, phonic groups, sentences or the total discourse
- Elimination an insertion of phones
- Zoom functions
- Playback of the synthetic speech

We have to notice that the emotional natural speech corpus was recorded in Castilian Spanish. However, the tool EMOVS has a Catalan Spanish speech database. Nevertheless, the hypothesis about the trans-linguistic character of the emotional features and the closeness between both of these languages were the reason to considerate that the obtained model and the synthesis tool were compatible.

The database is formed by 1026 speech units (diphones and triphones). These units are suitable segmented and labelled.

TD-PSOLA technique is based on a pitch synchronous analysis of speech, using 2T-length windows (T as the fundamental period) **[6]**. Synthesis is performed doing overlap and add (OLA) of the required units in the database. Pitch variation is done joining or separating the 2T frames before the OLA process. To modify the duration, some frames have to be repeated or removed. Intensity is modified increasing or decreasing the amplitude of the speech waveform.

3.2 Generation of emotional speech

The generation of emotional speech using TTS conversion has been done following these stages:

- Construction of a text corpus of carrier sentences.
- TTS conversion of each carrier sentence using EMOVS, repeating it so times as the desired emotional models.
- Acoustical edition of prosody following the parameters of each model. Every sentence has to be able to express all studied emotions.

In next paragraphs, we explain some practical considerations that have been appeared when prosody has been modified to match the expected emotion.

General considerations

The task of generating artificially the acoustic model of the emotional natural speech corpus has two main objectives:

- The validation of the emotional model.
- To improve the naturalness and the expression of the speech synthesis tool EMOVS.

This work of acoustical edition using a synthesis system has revealed some new important aspects, which would be very difficult to discover using only an acoustical analysis of natural speech:

- The type of symmetry in the intonation (sawtooth shape) that appears in emotional speech is a fundamental acoustic feature to model oral emotional expressions.
- The type of correspondence in time between maximums of energy and pitch is also an essential feature to recognise an emotion.
- The correlation between the time evolution of pitch and energy, rising (R) or falling (F), are outstanding in oral expression. There are both direct relations (R-R and F-F) and invert ones (R-F and F-R), depending on the expressed emotion.

Fear

In Figure 1, it is shown the energy and pitch contours of a sentence with only one phonic group that expresses fear. It can be observed that the sawtooth has fast variations of pitch (from 60 to 100 Hz) in only 20 or 30 milliseconds. The rise of pitch is very much slower than the fall because the abrupt jumps and the rising plateau structure of the higher part of each "tooth". This asymmetrical structure produces the typical sound of strangled voice when a person is frightened. It is important to observe that the energy is globally rising and the pitch-energy relation is synchronous, R-R and F-F.

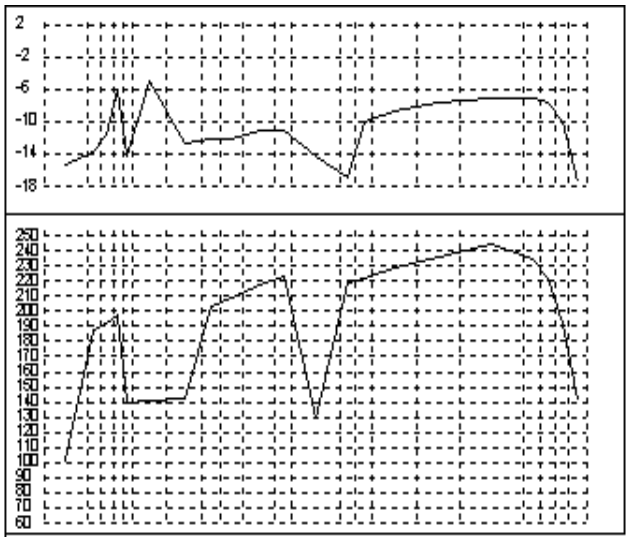

Figure 1: Energy and pitch contours for fear.

Fury

The example of fury has the same text as the preceding one. The variability of the sawtooth is practically the same than in fear emotion. The pitch-energy relation is also synchronous, R-R and F-F. The main difference between both emotions is the inverse symmetry of the sawtooth. In fury, the pitch attack is very faster than the fall (see Figure 2), because de plateau of the "tooth" is falling. Another difference is the width of the "teeth".

In fury, they are narrower (monosyllabic) than in fear (bi- or tri-syllabic). The violent and repeated rise of intonation that is related with the energy maximums produces the typical sensation of furious beating of this speech.

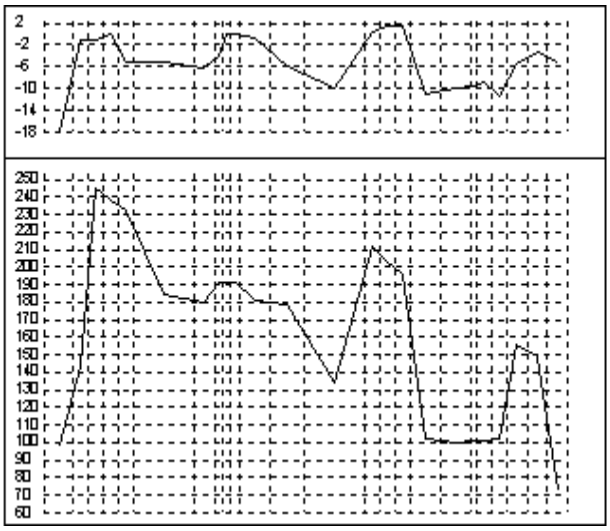

Figure 2: Energy and pitch contours for fury

Sadness

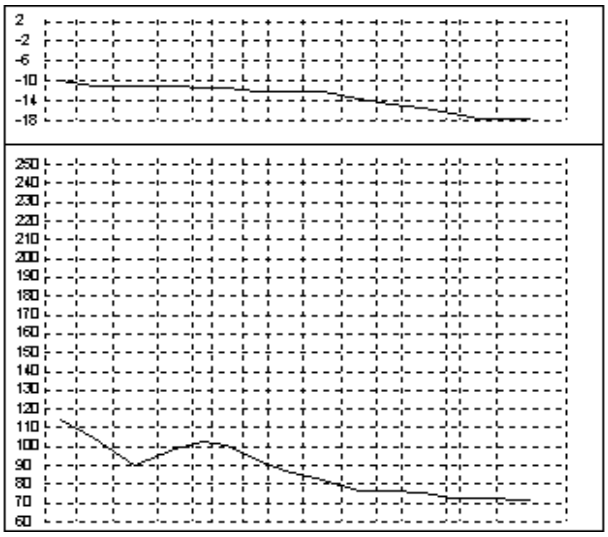

Figure 3: Energy and pitch contours for sadness

In this example, the text is also the same. The most significant feature of sadness is the minimum variability of energy and pitch (see Figure 3). Whereas the pitch variability in fear or fury can exceed 140 Hz, it does not exceed 30 Hz in sadness. The fall of intonation starts in a very low pitch. The sawtooth structure is not noticeable, and therefore, there is not intra-phonetic intonation like in fury or fear. These features and the slowness of the discourse produce a typical monotonous voice of sadness. The pitch-energy relation is synchronous.

Happyness

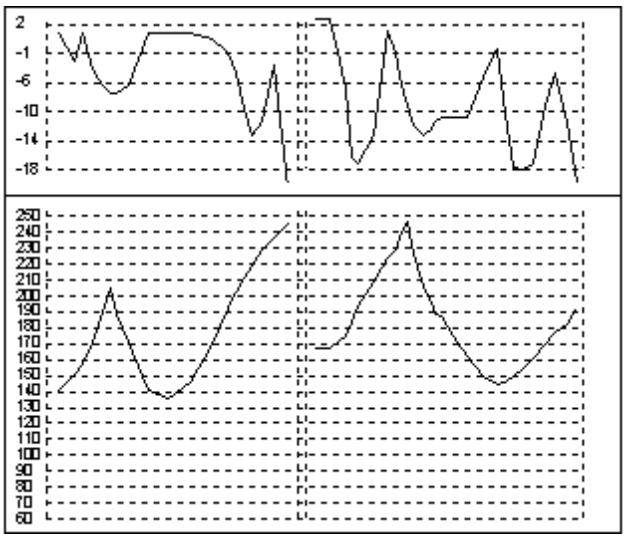

Figure 4: Energy and pitch contours for happiness

In this example, the text is different and it has two phonic groups. In happiness, the pitch variability is also very high, like in fear and fury, but the sawtooth is symmetrical, with similar rise and fall times. Another important difference is the absence of a plateau in the highest values of pitch. However, the most outstanding feature is the kind of relation between pitch and energy maximums. This pitch–energy relation is asynchronous (see Figure 4). Peaks in energy are advanced with regard to pitch ones, with a gap from 100 to 150 milliseconds. This gap between energy and pitch makes alike speech sound to acoustic structure of laugh.

Desire

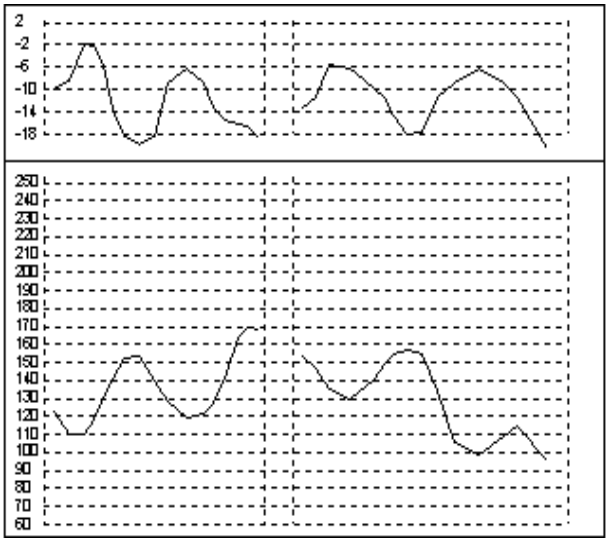

Figure 5: Energy and pitch contours for desire

In this example, the text is the same than in happiness. In desire, the sawtooth structure is symmetrical, although its variability is lower and inflections are very much smooth. In this case, there is also a gap between pitch and energy peaks, with a reversed structure R-F and F-R (see Figure 5). The smoothness of the contours and the opposite structure to violent sound are the soft and sensual features of this kind of voice.

3.3 Voice transformation

In the previous paragraphs, a method to generate emotional speech using text-to-speech has been described. The method of this subsection is an alternative to the use of a text-to-speech system. The emotional speech is obtained through voice transformation. To modify the emotional features of a voice signal, it is necessary a speech analysis-synthesis system that allowed the modification of prosody.

It has been developed a speech analysis system based on a decomposition of voice in a harmonic part and a stochastic part **[7]**. This system allows us to generate synthesised speech modifying these prosodic parameters:

- Pitch: variation of harmonic frequencies and their amplitude and phase
- Duration: repeating or removing frames
- Intensity: varying the amplitude of harmonics
- Timbre: varying the amplitude of harmonics in some bandwidths
- Energy of unvoiced part

This kind of analysis has more possibilities to modify prosody than TD-PSOLA technique. This part is under development and there are not valid results yet.

4. CONCLUSIONS

We considerate that the obtained results about acoustic models of fear, fury and sadness are applicable to a synthesis system. However, we have not finished the perception test yet. Therefore, the conclusions about these results are not definitive.

On the other hand, we have some important conclusions about the construction of emotional expressions using synthesised speech:

- In emotional speech, there is an intra-phonetic intonation. The tool EMOVS only allows a lineal interpolation of pitch and energy between phones. Then, to synthesise intra-phonetic inflections, it is necessary to repeat the same vocalic unit from 2 to 5 times.
- It is necessary to record para-phonetic sounds to express emotions as fear or desire. These units would be "unvoiced vowels" and vowel-exhalation and exhalation-vowel diphones.
- It would be necessary that the variability of the system were enlarged. TD-PSOLA technique achieves from 20 to 30 % of pitch variation with a low distortion of the speech signal. However, pitch and energy variability can exceed 60 % in emotional speech. In consequence, we should use a synthesis system with greater variability to synthesise some kind of emotions.

To improve the modelling of desire, happiness and surprise we considerate necessary to include the above features to the synthesis system and to perform new experiments.

5. REFERENCES

1. SCHERER,K.R.: *"Methods of research on vocal communication: paradigms an parameters"*. in K.R. SCHERER and P.EKMAN Eds., Handbook of Methods in non verbal behavior research, Cambridge University Press, 1982.
2. SCHERER,K.R.: "Vocal affect signaling: A comparative approach" in J.S ROSENBLATT, C. BEER, M.C. BUSNEL and P.J.B. SLATER Eds., *Advances in the Study of Behavior* (Vol. 15). New. York, Academic Press, 1985.
3. SCHERER,K.R.: "Vocal affect expression: a review an a model for future research". *Psichological Bulletin*, 99, 143-165, 1986.
4. REEVE, J.M.: *Motivación y emoción*, McGraw-Hill/Interamericana de España, S.A, Aravaca (Madrid). 1994
5. Rodríguez, A.; Lazaro,P.; Montoya, N.; Blanco, J.M.; Bernadas, D.; Oliver, J.M.; Longhi, L.: "Modelización acústica de la expresión emocional en el español". *Procesamiento del Lenguaje Natural*, nº 25, Lérida (España), septiembre del 1999, issn: 1135-5948, (pp. 159-166)
6. Moulines, E. and Charpentier F. Pitch-Synchronous waveform processing techniques for text-to-speech synthesis using diphones. *Speech Communication 9,* 453-467. 1990.
7. Stilianou, Y. *Harmonic plus noise models for speech, combined with statistical methods, for speech and speaker modifications.* PhD thesis, École Nationale des Télécommunications, Paris – France, 1996. sous la direction d'Eric Moulines.