Modeling and Synthesizing Emotional Speech for Catalan Text-to-Speech Synthesis

Ignasi Iriondo, Francesc Alías, Javier Melenchón, and M. Angeles Llorca

Enginyeria i Arquitecura La Salle, Ps. Bonanova 8, 08022 Barcelona, Spain
{iriondo,falias,jmelen,st06156}@salleURL.edu

Abstract. This paper describes an initial approach to emotional speech synthesis in Catalan based on a diphone concatenation TTS system. The main goal of this work is to develop a simple prosodic model for expressive synthesis. This model is obtained from an emotional speech collection artificially generated by means of a copy-prosody experiment. After validating the emotional content of this collection, the model was automated and incorporated into our TTS system. Finally, the automatic speech synthesis system has been evaluated by means of a perceptual test, obtaining encouraging results.

1 Introduction

Nowadays, human-computer interaction (HCI) systems tend to incorporate both vision and speech because they are the natural channels of human communication. For this reason, HCI should be bidirectional [1], since i) the machine could *understand* the user's message using both speech recognition and computer vision techniques [2] and ii) the machine could *answer* by means of audiovisual synthesis [3]. Moreover, the interaction would become more efficient and user-friendly if emotional expressions could be recognized [4][5] and synthesized [6].

The present work is our first approach to automatic emotional speech synthesis in Catalan with the purpose of including emotional expressivity in the output channel of an HCI system [7] [8]. Catalan is the native language of Catalonia, the Valencian Country and the Balearic Islands (central east and north east part of Spain), which is spoken by more than 6 million people. Nevertheless, these are bilingual areas where Spanish is also spoken, being, in fact, the dominant language used for communication (news, TV, radio, ...). Thus, Catalan is a minority language in front of Spanish influence. Like other languages, Catalan has several varieties (Central, North-occidental, Valencian, Balearic and others) that are spoken in different areas of the territory. We have focused our studies in the Central variety which is spoken near the city of Barcelona.

This paper is organized as follows: Section 2 presents a brief description of the related literature and our previous work in this field. Section 3 summarizes the method chosen to model and synthesize emotional speech, which is fully described in sections 4 and 5. Section 6 presents a discussion about the different topics related to the approach introduced. Finally, the last section presents the conclusion of this work.

E. André et al. (Eds.): ADS 2004, LNAI 3068, pp. 197–208, 2004.

2 Background

Emotional speech synthesis is a complex process that involves the integration of different knowledge areas. On the one hand, *Psychology* tries to describe emotions and the corresponding human actions to express and perceive them [9]. According to the review presented by Bartneck [10], emotions can be described as discrete categories or as regions of a multidimensional space. On the other hand, *Psychoacoustics* analyzes the effect of emotions on speech, proving that voice suffers acoustic changes due to physiological alterations [11]. Therefore, these variations have to be considered in order to obtain an acoustic model of emotional speech. Usually, the modeling process involves choosing the most relevant parameters and their behavior representation. In this sense, an emotional speech corpus becomes essential to define the corresponding acoustic model [12]. As a final point, the *Speech Technology* research collects this knowledge in order to synthesize emotional speech, incorporating the defined models and corpora. In [6], a full review of different approaches to emotional speech synthesis is presented. These approaches can be classified into three main groups: i) rule-based synthesis, such as HAMLET [13] and the Affect Editor [14], ii) diphone concatenation synthesis [15] [16], and iii) corpus-based synthesis [17]. To date, the most complete study of emotional speech synthesis in Castilian Spanish has been described by Montero [18] [19].

2.1 Our Previous Work

This work is based on the previous investigations of [20], who built an acoustic model for basic emotions in Castilian Spanish following these steps:

- *Generation of a database of emotional speech.* The speech corpus was recorded by 8 actors (4 males and 4 females), performing a set of carrier sentences with the seven basic emotions (fear, happiness, anger, sadness, surprise, desire and disgust). Each text was pronounced with 3 levels of emotional intensity. Therefore, 336 different discourses were collected (2 texts x 8 actors x 7 emotions x 3 intensities).
- *Perceptual test.* A perceptual test was carried out to choose the most representative recordings for each emotion. Each emotional locution (with duration ranging from 20 to 40 seconds) was listened to by two groups of 30 people. The five best ranked utterances of each emotion were selected for the final database, according to the highest percentage of both identification and level of credibility. The identification score of all the selected utterances exceeded 90%, except for disgust, where all utterances scored less than 50%. For this reason, disgust was not acoustically modeled and therefore the model was only defined for six emotions.
- *Acoustic analysis.* A systematic analysis of the selected utterances was developed in terms of fundamental frequency (mean, range and variability), sound pressure (mean, range and variability) and timing parameters (duration and number of silences, duration and number of phonic groups and syllables per second).

Later, this acoustic model was validated using concatenative speech synthesis [21], concluding that the model was valid to simulate sadness, anger and fear only by means of the prosodic modification. However, the results obtained for happiness, surprise and desire showed that this methodology was inadequate to achieve a sufficient level of identification of these emotions.

3 Our Approach

The main objective of this work is to incorporate emotional speech into our Catalan TTS system, with the purpose of achieving better expressiveness.

Our approach starts from the hypothesis that the modification of the prosodic parameters of the TTS system is the first step towards emotional speech synthesis. To date, we only considerer melody, rhythm and energy modeling by means of the modification of pitch, energy and duration of the phones and the duration of pauses. Initially, only four basic emotions have been modeled (fear, happiness, anger and sadness) with the goal of exploring the possibilities of this method. As a result, a prosodic model for each emotion will be obtained and implemented.

One significant aspect of this approach is the use of a previous developed resource, a Spanish database, instead of developing a new Catalan database. We believe that a Spanish database would suffice because two languages are very similar and almost all Catalan people speak perfectly both languages. At the begining of this work we dispose of a Spanish database perfectly validated by means of a rigorous perceptual test [20]. The recording and validation of a new database in Catalan would suppose an expensive task that we could not carry out. Therefore, the data generation in Catalan was performed by means of a copy-synthesis experiment with translated texts of the Spanish database.

3.1 Summary of the Methodology

Below, we present a summary of the steps followed during this work (see Figure 1), which are detailed in sections 4 and 5:

1. The emotional speech corpus in Spanish presented in [20] was chosen as the source data for the acoustic modeling. From this corpus, four different utterances containing the same text (one per emotion) were selected, representing the corresponding emotion clearly. Each utterance contained 11 sentences which are segmented into phones and labeled (pitch, energy and duration). The duration of the pauses between sentences was also annotated.
2. Then, these sentences were translated into Catalan and their phonetic transcriptions were generated. Moreover, the prosody associated to each phone was manually adjusted from corresponding phone of the Spanish database.
3. Next, a small data collection of emotional speech in Catalan was generated after TTS synthesis.
4. A perceptual test of emotion identification was developed in order to validate this synthetic speech collection.

5. A prosodic model adapted to our TTS system was obtained by comparing the annotated emotional speech with the default output (neutral) when synthesizing the same text.
6. The emotional prosodic model was incorporated to the Natural Language Processing module of our TTS system
7. And finally, a new perceptual test was performed to evaluate the degree of emotion identification when this model was automated.

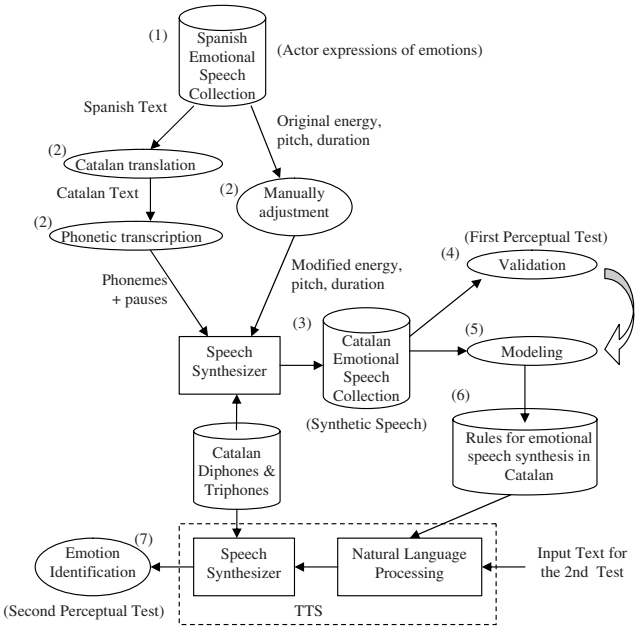

Fig. 1. Flowchart that summarizes the seven steps followed during the definition and the validation of the prosodic model for emotional speech synthesis in Catalan

4 Prosodic Modeling Oriented to Expressive Catalan TTS

This section describes the process followed to obtain a prosodic model for emotional speech synthesis adapted to our TTS system in Catalan. This process follows three steps: firstly, the generation of the emotional speech data collection, secondly, its validation, and, finally, the definition of the prosodic model.

4.1 Stimuli Generation

The emotional speech data collection in Catalan is generated using our TTS system with an input of phonetic and prosodic information manually adjusted. The phonetic transcription was obtained by translating the original text into Catalan. The prosodic information was adjusted following the same patterns of intonation, rhythm and intensity of the utterances selected from the Spanish corpus. Note that we assume the Spanish prosodic pattern can be valid for Catalan due to the similarities between the phonetics and the prosody of both languages.

As a result, we had at our disposition a synthetic male speech collection corresponding to four different emotions in Catalan of the same text (see Appendix A). In addition, the speech collection was completed with neutral utterance synthesizing the same text by our TTS system. This neutral utterance is used as the reference pattern for the prosodic modeling of the considered emotions.

4.2 Validating the Emotional Speech Data

A perceptual test involving 10 listeners was conducted to validate the emotional speech generated artificially. Before the beginning of the test, neutral synthetic speech was played to the listener in order to become familiarized with the synthetic speech of our TTS system. Next, each subject listened to four utterances randomly selected in two consecutive rounds. They had to choose between fear, happiness, anger, sadness or uncertain emotion (forced-test) after listening to each utterance only once. As depicted in Figure 2, the second round presents a better degree of identification than the first one. The reason for this result is that the subjects have listened to all stimuli once (they already disposed of a comparison criterion among the four emotions).

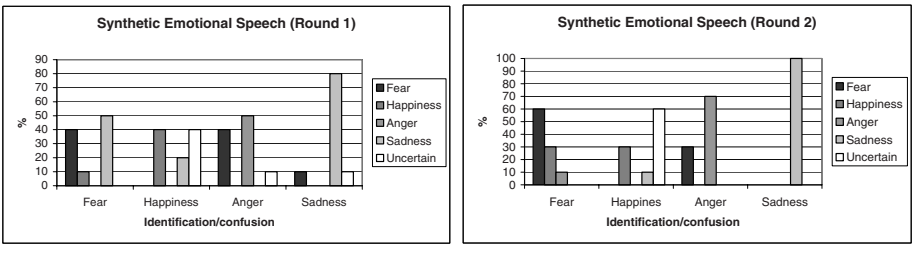

Fig. 2. Identification/confusion percentages between emotions on the synthetic speech obtained with the prosody manually adjusted

Moreover, Figure 2 shows that sadness is the best identified emotion in both rounds, with 100% of identification in the second one. Moreover, anger and fear improve their percentages after the first round due to the fact of establishing differences between emotions. Nevertheless, happiness is the only emotion that

presents worse results in the second round, decreasing from 40% to 30% of identification. This experiment corroborates the general results which denote that happiness is a difficult emotion to simulate [16],[18] and [21]. Moreover, some listeners expressed the increased difficulty of identifying happiness in an utterance with no positive intention.

4.3 Generating the Prosodic Model

The main goal of this modeling is to obtain a general behavior of the prosodic parameters of speech related to the four considered emotions. However, the emotional prosodic model developed in this work is simple because of the constraint of being incorporated in a concatenative speech synthesizer. In this approach, the prosodic parameters calculated from the male speech collection are grouped into three categories: pitch, timing and energy, which are described as follows:

Table 1. Relative percentage of mean variation of pitch parameters with respect to neutral style for each emotion

Relative mean variation	Fear	Happiness	Anger	Sadness
Average pitch	+52%	+13%	+33%	-7%
Pitch Range	-3%	-10%	+30%	-60%

Pitch. The pitch parameters describe several features of the fundamental frequency (F0). The average pitch and the pitch range of each utterance are calculated. Table 1 summarizes the percentage of the average variation of both parameters with respect to the corresponding values of the neutral style. Notice that the average pitch presents high increments in fear and anger, and a slight reduction in sadness. On the other hand, the pitch range presents higher variations in anger (positive) and sadness (negative). However, happiness presents non-representative variations of both parameters in terms of their mean values. Moreover, Figure 3 presents the statistical representation of the results obtained for both parameters in the emotional speech data collection.

Timing. The timing parameters describe the features related to speech rate. In this paper, the duration of pauses and the average duration of phonic group are calculated for each utterance. The duration of pauses parameter represents the increment or the decrement of the mean duration of pauses for each emotion with respect to the neutral speech (see Table 2).

The duration of phonic groups is calculated as the relative percentage of the mean phone durations of an emotional utterance with respect to its mean value in the neutral utterance (see Table 3). Note that the last phonic groups of the

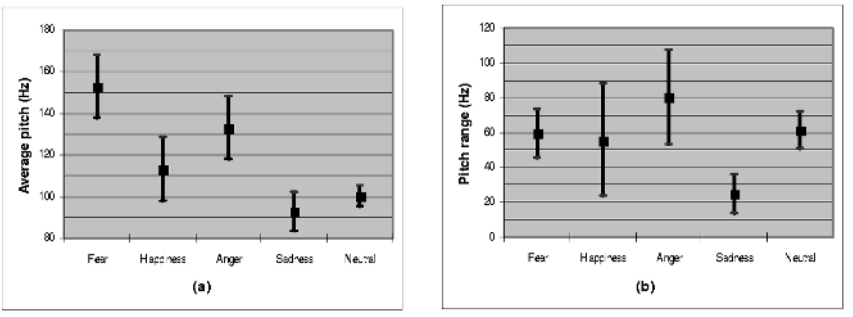

Fig. 3. (a) Mean and standard deviation ($\mu \pm \sigma$) of average pitch. (b) Mean and standard deviation ($\mu \pm \sigma$) of pitch range.

Table 2. Relative percentage of mean variation of pause duration with respect to the neutral style

Relative mean variation	Fear	Happiness	Anger	Sadness
Duration of pauses	+38%	+3%	-15%	+128%

sentences are treated separately because they present a different behavior. Table 3 also shows that the speech rate is accelerated for anger and it is clearly slowed for sadness.

Table 3. Relative percentage of mean variation of phone duration with respect to neutral style

Relative mean variation	Fear	Happiness	Anger	Sadness
Duration of last phonic groups	+9%	+0.2%	-7%	+25%
Duration of the other phonic groups	+6%	+12%	-4%	+23%

Energy. The energy parameters describe features of the amplitude of the speech signal. We have calculated the average energy and the energy range of each phrase of the corpus. The Table 4 summarizes the calculated values as an increment or decrement of the average energy with respect to the neutral emotion and its energy range. All these values are expressed in decibels (dB).

5 Automation of the Prosodic Model

The automation of the prosodic model involves the definition of a set of rules describing the results presented in Section 4. These rules are defined as a modi-

Table 4. Relative variation of energy parameters with respect to the neutral style in dB

Relative variation	Fear	Happiness	Anger	Sadness
Average Energy	-0,16	+0,29	+1,13	-1,46
Energy Range	+13	+11,1	+14,3	+10,4

fication of the prosodic parameters generated automatically by the TTS system for the neutral style. As pitch and energy parameters have two degrees of freedom (average and range), the adjustment of their corresponding values for each phone of a sentence follows different steps:

1. From the text, the prosodic parameter values for each phone are calculated (neutral style), p_0 in equations (1) and (2).
2. Normalize p_0, subtracting its mean value $\bar{p}_0$
3. Adjust the normalized values to the desired range following equation (1), where $\Delta\bar{R}$ is the mean range correction.
4. The final parameter values, p_f, are obtained adding the desired average parameter to the values obtained after step 3. In equation (2), $\Delta\bar{A}$ is the mean average correction.

$$\hat{p} = \Delta\bar{R} \cdot (p_0 - \bar{p}_0) \tag{1}$$

$$p_f = \hat{p} + \Delta\bar{A} \cdot \bar{p}_0 \tag{2}$$

On the other hand, the duration parameter adjustment consists of multiplying the values generated by the TTS system by the mean duration correction. The new speech rate is obtained after applying the corresponding duration corrections to pauses and phones. As described in Section 4.3, the model defines a particular modification for the final phonic group.

5.1 Evaluating the Automatic Results

A perceptual test was conducted in order to evaluate the emotional speech generated automatically. Ten non-expert subjects (students of Engineering) listened to four utterances with a different emotion synthesized from the same text in two consecutive rounds. The listener had to choose between fear, happiness, anger, sadness or uncertain (forced-test).

Figure 4 shows the percentages of the conducted emotion identification/confusion test. Sadness is the emotion with the highest percentage of identification followed by fear, which presents an acceptable result. Anger is confused with happiness in a 30%. Happiness obtains a result only slightly higher than the baseline.

6 Discussion

After analyzing the obtained results, we wish to discuss some important topics related to the followed methodology and the resources employed in this work.

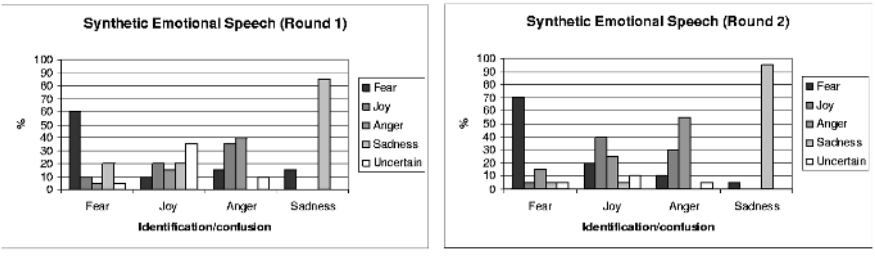

Fig. 4. Perceptual test of identification/confusion of emotion from our TTS system when the automated prosodic model is incorporated

6.1 Emotion Description for Speech Synthesis Applications

In our approach, the emotions are treated as distinct categories, and in this particular case, we have taken into account only four basic emotion categories such as anger, happiness, fear and sadness. According to [22], an alternative approach starts with a different representation of emotional states, as emotion dimensions rather than discrete categories. Emotion dimensions are a simplified representation of the essential properties of emotions. Usually, Evaluation (positive/negative) and Activation (active/passive) are the most important dimensions. The resulting synthesis system of this dimensional approach is by design highly flexible and very efficient when complementary sources of information, such as verbal content, the visual channel, and the situational context, are present in an HCI application. In this sense, authors will further pursue research in order to incorporate different emotional states or dimensions.

6.2 Emotional Speech Corpus

The development of a new database of emotional speech in any language is a difficult and expensive task. In [12], the main issues that need to be considered when developing this kind of corpus are presented. If the database is oriented to emotional speech synthesis, it has to be designed to involve both the acoustic modeling of emotions (off-line process) and the synthesis process (on-line). Currently, there is a tendency to use unit selection synthesizers in order to minimize the prosodic modification of the acoustic units at synthesis time [17]. Consequently, the resulting speech is natural-sounding for the specific categories which were recorded without a previous modeling of the acoustic properties of emotions. The most important limitation of this kind of approach is the enormous database (one corpus per emotion) involved in the speech synthesis process.

6.3 Modeling Speech and Emotion

One of the most critical aspects of the acoustic modeling of emotional expressions is the temporal behavior of the voice parameters, which suffer changes

depending on the emotional state of the talker. These changes are only present in certain moments of the speech message [20]. Thus, both the frequency of apparition of these changes and their time position in the discourse are key issues to be explored in future investigations. For instance, in this work, we have not obtained good results modeling and synthesizing happiness because our model has only been based on average values of the analyzed parameters. Therefore, this approach seems inadequate when the considered parameters present a high deviation with respect to their mean value. We believe that the right control of the parameter changes over the time would improve the acoustic modeling of this kind of emotion.

Another important issue to be taken into account is the development of analysis and synthesis tools to process the whole set of relevant speech parameters involved in emotional expressions [14]. Currently, we are working on a parameterized concatenative speech synthesis system to allow a higher degree of control of prosodic and acoustic modification of the speech units [23]. This system is a hybrid method based on TD-PSOLA and the harmonic plus noise model, which incorporates a novel method to jointly modify pitch and time-scale. Moreover, it is able to control separately the energy of both the harmonic and the noise components of speech.

Moreover, prosodic modeling of emotions would have to be improved to achieve a major level of credibility for some emotions. On the one hand, the method for modeling duration should consider the differential elasticity of the different phoneme classes instead of stretching uniformly all synthesized phonemes [24]. On the other hand, there are problems associated with energy modeling that we have to solve. For instance, the change in voice quality parameters related to vocal effort.

7 Conclusion

This work is our first approach to emotional speech synthesis in Catalan by means of modeling average variation of prosodic parameters with respect to the neutral prosody generated by our TTS system. An emotional speech collection in Catalan has been generated artificially and it has been validated via a perceptual test. The analysis of this speech collection has resulting in the definition of a model that converts neutral prosody into emotional prosody automatically. A second perceptual test has been performed in order to evaluate the identification/confusion percentages of the automatic system obtaining encouraging results.

References

1. Massaro, D.W., Light, J., Geraci, K., eds.: Auditory-Visual Speech Processing (AVSP 2001), Aalborg (Denmark) (2001)
2. Petajan, E.D.: Automatic lipreading to enhance speech recognition. In: Proceedings of the Global Telecommunications Conference, IEEE Communication Society (1984) 265–272

3. Bailly, G., Bérar, M., Elisei, F., Odisio, M.: Audiovisual speech synthesis. International Journal of Speech Technology (2003) 331–346
4. Cowie, R., Douglas-Cowie, E., Tsapatsoulis, N., Votsis, G., Kollias, S., Fellenz, W., Taylor, J.G.: Emotion recognition in human computer interaction. IEEE Signal Processing **18** (2001) 33–80
5. Batliner, A., Fischer, K., Huber, R., Spilker, J., Nöth, E.: How to find trouble in communication. Speech Communication **40** (2003) 117–143
6. Schröder, M.: Emotional speech synthesis: A review. In: The Proceedings of Eurospeech 2001 - Scandinavia. Volume 1., Denmark (2001) 561–564
7. Melenchón, J., Alías, F., Iriondo, I.: Previs: A person-specific realistic virtual speaker. In: IEEE International Conference on Multimedia and Expo (ICME'02), Lausanne, Switzerland (2002)
8. Melenchón, J., De la Torre, F., Iriondo, I., Alías, F., Martínez, E., Vicent, L.: Text to visual synthesis with appearance models. In: IEEE International Conference on Image Processing (ICIP), Barcelona, Spain (2003)
9. Cowie, R., Cornelius, R.R.: Describing the emotional states that are expressed in speech. Speech Communication **40** (2003) 5–32
10. Bartneck, C.: Affective expressions of machines. Master's thesis, Stan Ackerman Institute, Eindhoven (2000) http://www.bartneck.de/work/aem.pdf.
11. Scherer, K.R.: Vocal affect expression: a review and a model for future research. Psichological Bulletin **99** (1986) 143–65
12. Douglas-Cowie, E., Campbell, N., Cowie, R., Roach, P.: Emotional speech: towards a new generation of databases. Speech Communication **40** (2003) 33–60
13. Murray, I.R., Arnott, J.L.: Implementation and testing of a system for producing emotion-by-rule in synthetic speech. Speech Communication **16** (1995) 369–390
14. Cahn, J.E.: Generating expression in synthesized speech. Master's thesis, Massachusetts Institute of Technology (1989)
15. Heuft, B., Portele, T., Rauth, M.: Emotions in time domain synthesis. In: Proceedings of ICSLP'96, Philadelphia, USA (1996)
16. Bulut, M., Narayanan, S., Syrdal, A.: Expressive speech synthesis using a concatenative synthesizer. In: Proceedings of ICSLP'02. (2002)
17. Iida, A., Campbell, N., Higuchi, F., Yasumura, M.: A corpus-based speech synthesis system with emotion. Speech Communication **40** (2003) 161–187
18. Montero, J.M., Gutiérrez Arriola, J., Colás, J., Macías Guarasa, J., Enríquez, E., Pardo, J.: Development of an emotional speech synthesiser in spanish. In: The Proceedings of Eurospeech 1999. (1999) 2099–2102
19. Montero, J.M.: Estrategias para la mejora de la naturalidad y la incorporación de variedad emocional a la conversión texto a voz en castellano. PhD thesis, Universidad Politécnica de Madrid (2003) (In Spanish).
20. Rodríguez, A., Lázaro, P., Montoya, N., Blanco, J., Bernadas, D., Oliver, J., Longhi, L.: Modelización acústica de la expresión emocional en el espaol. Procesamiento del Lenguaje Natural (1999) 159–166 (In Spanish).
21. Iriondo, I., Guaus, R., Rodríguez, A., Lázaro, P., Montoya, N., Blanco, J., Bernadas, D., Oliver, J., Tena, D., Longhi, L.: Validation of an acoustical modelling of emotional expression in Spanish using speech synthesis techniques. Proceedings of the ISCA Workshop on Speech and Emotion (2000) 161–166
22. Schröder, M.: Speech and Emotion Research: An overview of research frameworks and a dimensional approach to emotional speech synthesis. PhD thesis, Institute of Phonetics, Saarland University, Germany (2003)

23. Iriondo, I., Alías, F., Sanchis, J., Melenchón, J.: Hybrid method oriented to concatenative text-to-speech synthesis. In: the 8th European Conference on Speech Communication and Technology (EUROSPEECH), Geneve, Switzerland (2003)
24. Brinckmann, C., Trouvain, J.: The role of duration models and symbolic representation for timing in synthetic speech. International Journal of Speech Technology **6** (2003) 21–31

Appendix

Initial text in Spanish:

"La casa apareció al final del camino. Empezaba a ponerse el sol, pero la fachada del edificio aun se veía con claridad. Unas figuras pasaban por detrás de las ventanas del piso superior. Me acerqué poco a poco. Nadie me vio, nadie me esperaba, nadie me recibió, entre sin hacer ruido. Subí las escaleras con agilidad. Las voces me guiaron hasta la gran habitación y lo vi todo."

Translated text into Catalan:

"La casa aparegué al final del camí. Començava la posta de sol però, la façana de l'edifici encara es veia amb claredat. Unes figures passaven per darrera de les finestres del pis superior. Em vaig apropar a poc a poc, ningú em veié, ningú m'esperava, ningú em rebé. Vaig entrar sense fer soroll. Vaig pujar les escales amb agilitat. Les veus em guiaren fins a la gran habitació i ho vaig veure tot."