Objective and Subjective Evaluation of an Expressive Speech Corpus

Ignasi Iriondo, Santiago Planet, Joan-Claudi Socoró, Francesc Alías

Department of Communications and Signal Theory
Enginyeria i Arquitectura La Salle, Ramon Llull University, Barcelona, Spain
{iriondo, splanet, jclaudi, falias}@salle.url.edu

Abstract

This paper presents the validation of the expressive content of an acted oral corpus produced to be used in speech synthesis. Firstly, objective validation has been conducted by means of automatic emotion identification techniques using statistical features obtained from the prosodic parameters of speech. Secondly, a listening test has been performed with a subset of utterances. The relationship between both objective and subjective evaluations is analysed and the obtained conclusions can be useful to improve the following steps related to expressive speech synthesis.

1. Introduction

There is a growing tendency towards the use of speech in human-machine interaction. Automatic speech recognition is used to consult information or to make managements. Speech synthesis let machines to communicate orally with users (automation of services or aid to disabled people). The incorporation of the recognition of emotional states or the synthesis of emotional speech can improve the communication by doing it more natural [1]. Therefore, one of the most important challenges in the study of the expressive speech is the development of oral corpora with authentic emotional content that enable robust analysis according to the task for which they have been developed. It is not the objective of the present work to carry out an exhaustive summary of the available databases for the study of emotional speech, since recently, complete studies have appeared in the literature. In [2], a new compilation of 48 databases is presented showing a notable increase of multimodal databases. In [3], the databases used in 14 experiments of automatic detection of the emotion are summarized. Finally, in [4] a revision of 64 databases of emotional speech is done, providing a basic description of each one and its application.

This paper describes the main aspects of the production of an expressive speech corpus in Spanish faced to synthesis and the objective and subjective evaluation of its emotional content. Section 2 introduces different aspects about the corpora of expressive speech. Section 3 explains the production of our corpus. Section 4 details the process of objective validation carried out through techniques of automatic identification of the emotion. The subjective evaluation by means of perception test is explained in Section 5, and finally, the conclusions (Section 6).

2. Building emotional speech corpora

According to [5], there are four main aspects to be considered in the development of an emotional speech corpus: *i)* the **scope** that covers the database (number of speakers, the language, dialects, genre of the speakers and types of emotional states); *ii)* the **context** in which a locution takes place (emotional significance perceived across the semantics, the prosody, the facial expression, gestures and posture); *iii)* the **descriptors** that allow to represent the linguistic, emotional and acoustic content of the speech; and *iv)* the **naturalness** of the locutions, which will depend on the strategy followed to obtain the emotional speech. With respect to the latter, the main debate is centred on the compromise between authenticity of the expressed emotion and the control on the recording. Campbell [1] and later Schröder [6] propose 4 emotional speech sources:

Natural occurrences. Spontaneous human interaction presents the most natural emotional speech although it has some drawbacks due to the lack of control on its content, the quality of sound, the difficulty of labelling, and finally, legal and ethical aspects (e.g. *The Reading-Leeds*, *The Belfast Naturalistic* and *The CREST* databases described in [5]).

Elicitation. The provocation of authentic emotions in people in the laboratory is a way of compensating some of the problems described previously, although the fullblown emotions would remain out of place [1]. In [6], five types of mood induction procedures are described.

Stimulated emotional speech. This method consists of the reading of texts with a verbal content adapted for the emotion to be expressed. The difficulty of comparing utterances with different texts should be counteracted with an increase of the corpus size so that statistical methods allow to generalize models [1]. This technique was followed in the creation of the *Belfast Structured Emotion Database* [5].

Acted emotional speech. The great advantage of this method is the control of the verbal and phonetic content of speech since all the emotional states can be produced using the same phrases. This allows direct comparisons of the phonetics, the prosody and the voice quality for the different emotions. The great objection that presents is the lack of authenticity of the expressed emotion [1].

Another important aspect to keep in mind is the purpose of the speech and emotion research. It is necessary to distinguish between processes of perception (*centred on the speaker*) and expression (*centred on the listener*) [6]. The objective of the former is to establish the relation between the speaker emotional state and quantifiable parameters of speech. Usually, they deal with the recognition of emotions from speech signal. According to [3], one of the challenges is the identification of oral indicators (prosodic, spectral and vocal quality) attributable to the emotional behavior and that are not simply own characteristics of conversational speech. The latter model the parameters of the speech with the goal to transmit a certain emotional state. The description of emotional states and the choice of speech parameters are key in the final result. There is a high consen-

This work has been partially supported by the European Commission, project SALERO FP6 IST-4-027122-IP.

sus in the scientific community for obtaining emotional speech by means of stimulated/acted speech for synthesis purposes [5] [2], although other authors argue in favour of constructing an enormous corpus gathered from recordings of the daily life of a number of voluntary speakers [7].

This work combines methods of both types of studies. On the one hand, the production of the corpus follows the guidelines of the studies *centred on the listener* since it is oriented to speech synthesis. On the other hand, we apply techniques of emotion recognition in order to validate its expressive content.

3. Our expressive speech corpus

We considered the development of a new expressive oral corpus for Spanish due to lack of availability of a corpus with the suitable characteristics within the framework of our research in expressive speech synthesis. This corpus had a double purpose: to learn the acoustic models of emotional speech and to be used as the speech unit database for the synthesizer. This section describes the steps followed in the production of the corpus.

3.1. Stimulated emotion and text design

For the recording of the present corpus, a female professional speaker has been chosen due to her capability to use the suitable expressive style to each text category (stimulted/acted speech).

For the design of texts semantically related to different expressive styles, we have made use of an existing textual database of advertisements extracted from newspapers and magazines. Based on a study of the voice in the audio-visual publicity [8], five categories of the textual corpus have been chosen and the most suitable emotion/style has been assigned to them: New technologies (neutral-mature), education (joy-elation), cosmetic (style sensual-sweet), automobiles (aggressive-hard) and trips (sad-melancholic).

A set of phrases has been selected from each category by means of a *greedy* algorithm [9] that has allowed to obtain a phonetic balance in each subcorpus. This type of algorithms take the locally optimum choice at each stage with the hope to find an adequate global solution. Therefore, the application of this algorithm to the raised problem will obtain a valid solution, although may be not the optimum one. In addition to looking for a phonetic balance, phrases that contain exceptions (e.g. foreign words, abbreviations) have been avoided due to they make difficult the automatic processes of phonetic transcription and labelling. Moreover, the selection of similar phrases to others previously selected has been penalized by the *greedy* algorithm.

3.2. Recording

The recording of the oral corpus has been carried out in a professional recording studio. Speech signals were sampled at 48 KHz and quantized using 24 bits per sample and stored in WAV files. Different recording sessions have been required and therefore a preestablished protocol has been followed in order to minimizing errors that can cause deficiencies in the corpus labelling. For the corpus segmentation in phrases, a semiautomatic process has followed by means of a forced alignment using Hidden Markov Models from the phonetic transcription and later a manual review and correction. This forced alignment also has been used to segment the phrases in phonemes. The recorded database has 4638 sentences and it is 5 h 12 min long.

4. Objective validation

The goal of the experiments described in this section was to validate the expressive content of the corpus by means of automatic emotion identification using different data mining techniques applied to statistical features computed over the prosodic parameters of speech. An exhaustive subjective evaluation of the full corpus (more than 5 hours of speech) would be a very tedious task and practically impossible. However, the whole corpus can be validated by means of these automatic techniques.

4.1. Acoustic analysis

Prosodic features of speech (fundamental frequency, energy, duration of phones and frequency of pauses) are related to vocal expression of emotion [10]. In this work, an automatic acoustic analysis of the sentences is performed using information of the previous phonetic segmentation.

4.1.1. F0 related parameters

The analysis of the fundamental frequency (F0) parameters is based on the result of the pitch marker described in [11]. This system assigns marks over the whole signal. The unvoiced segments and silences are marked using interpolated values from the neighboring voiced segments. For each phrase, three vectors of local F0 values are obtained (complete, excluding silences and unvoiced sounds, and only the stressed vowels). The information about the boundaries of voiced/unvoiced (V/UV) segments and silences is obtained from the corpus labelling. Notice that if the phonetic segmentation was not available, an automatic voice-activity detector (VAD) and a V/UV detector would be required [12]. Moreover, F0 has been calculated in both lineal and logarithmic scales.

4.1.2. Energy related parameters

For energy, speech is processed with 20-ms rectangular windows and 50% of overlap, calculating the power (linear and dBs) every 10ms. Following the same idea that for F0, three vectors per utterance are generated (complete, excluding silences, and only in the stressed vowels).

4.1.3. Rhythm related parameters

The duration of phones is an important cue for vocal expression of emotion. However, some studies omit this parameter by the difficulty to obtain it automatically [12]. In the present work we have incorporated this information (thanks to the labelling of the corpus) to generate datasets with and without this information in order to contrast its relevance. Z-scores have been employed for duration modeling in text-to-speech synthesis (TTS) to predict individual segment durations and to control the lengthening or the shortening of phones. As in [13], we take z-scores as a means to analyze the temporal structure of speech:

$$z\_score = \frac{dur(ms) - \mu}{\sigma} \quad (1)$$

where μ and σ are the mean and the standard desviation respectively of the corresponding phoneme. Therefore, the rhythm of an utterance is represented by a vector with the z-score of each phoneme. The version of this vector with only the stressed vowels is also computed.

Moreover, two pausing related parameters are added for each utterance: the number of pauses per time unit and the percentage of silence respect to the total time.

4.2. Statistical analysis and datasets

The prosody of an utterance is represented by some sequencies of values by phoneme such as F0 (lineal and logarithmic), energy (lineal and dB) and normalized durations (z-score). For each sequence, the first and the second derivative are calculated. For all these resulting sequences, the following statistics are obtained: mean, variance, maximum, minimum, range, skew, kurtosis, quartiles, and interquartilic range. Finally, 464 parameters by utterance are calculated, considering both parameters related to the pausing.

This set of parameters has been divided into different subsets according to different strategies to reduce the dimensionality (see the diagram of the figure 1). A first criterion to reduce it has been to omit the second derivative (from Data1 to Data2) in order to valorate the significance of this function. Secondly, preliminary experiments have shown that the use of the logarithmic versions of F0 and energy obtain better results. For this reason, two new datasets have been generated without the linear versions of both F0 and energy. Each one of these datasets (Data1L and Data2L) has been divided in two new sets considering all the phonemes or only the stressed vowels. Moreover, an automatic reduction of both initial datasets (with and without the 2nd derivative) has been carried out by means of the simple genetic algorithm (GA) implemeted in Weka [14] (Data1G and Data2G). This reduction is independent of the later classification algorithm and therefore all the techniques have been tryed with these datasets. Finally, two similar datasets to *Navas et al. (2006)* [12] have been generated to test the significance of omiting the timing parameters (Data1N and DATA1NG).

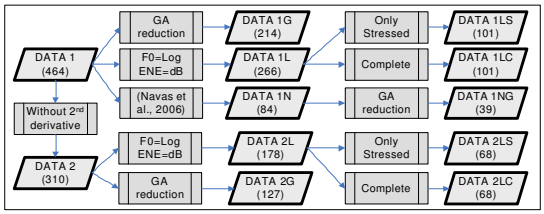

Figure 1: Generation of different datasets

4.3. Experiments and results

Numerous schemes of automatic learning can be used in a task such as classifying the style/emotion from the acoustic analysis of the speech. The objective evaluation of expressivity in our speech corpus is based on [15], where a large-scale data mining experiment about the automatic recognition of basic emotions in short utterances was conducted. After different preliminary experiments, the set of machine learning algorithms showed in table 1 has been selected in order to be tested with the different datasets. Some algorithms have been completed with their *boosted* versions that achieve better results although they present a greater computacional cost. All the experiments have been carried out using Weka software [14] by means of ten-fold cross-validation. Both tryed versions of SMO (Support Vector Machine of Weka) obtain the best results so much on average as in maximum value (see table 1). SMO algorithms achieve the highest results with Data1G, showing that the dimensionality reduction based in GA helps to these systems, although differences with Data1L and Data1LC are minimum. However, other algorithms(i.e. J48, IB1 and IBk) work better with datasets genererated by two consecutive reductions (without 2nd derivative and latter GA reduction). And finally, we can observe that there is a third group of algorithms that work better if the linear/logarithic redundance of F0 and energy is removed. Also we can observe that the boosted versions improve significantivelly the results respect to their corresponding algorithms. Figure 2 shows a comparison between different datasets depending on the algorithm. Notice that Data1LC obtains almost the same results than Data1G and Data1L, but with less than the half of parameters. The same effect is presented in the datasets without de 2nd derivative. Results experiment a slight loss when timing parameters are removed (Data1N and Data 1NG). However, results worsen significantly when parameters are calculated only in the stressed vowels (Data1LS and Data2LS). Table 2 shows the confusion matrix with the average results for the eleven classifiers with Data2G, that has achieved the best mean percentage of identification (97.02 % $\pm$ 1.23).

Table 1: Learning Algorithms used for the automatic recognition experiment

Name	Description	mean(95%CI)	max(Data)
J48	Decision tree based on C4.5	93.4 $\pm$ 2.0	96.4 (2G)
B.J48	Adaboosted version of J48	96.4 $\pm$ 1.4	98.3 (1L)
Part	Decision Rules (PART)	94.2 $\pm$ 2.0	96.9 (2L)
B.Part	Adaboosted version of PART	96.7 $\pm$ 1.3	98.4 (1G)
DT	Decision Table	88.7 $\pm$ 2.6	92.3 (1L)
B.T	Adaboosted version of D. T.	93.4 $\pm$ 1.6	96.1 (1L)
IB1	Instance-based (1 solution)	93.3 $\pm$ 2.8	97.5 (2G)
IBk	Instance-based (k solutions)	94.0 $\pm$ 2.3	97.9 (2G)
NB	Naive Bayes with discretization	94.6 $\pm$ 1.9	97.8 (1L)
SMO1	SVM with 2nd degree pol. Kernel	97.3 $\pm$ 1.2	99.0 (1G)
SMO2	SVM with 3rd degree pol. Kernel	97.1 $\pm$ 1.5	98.9 (1G)

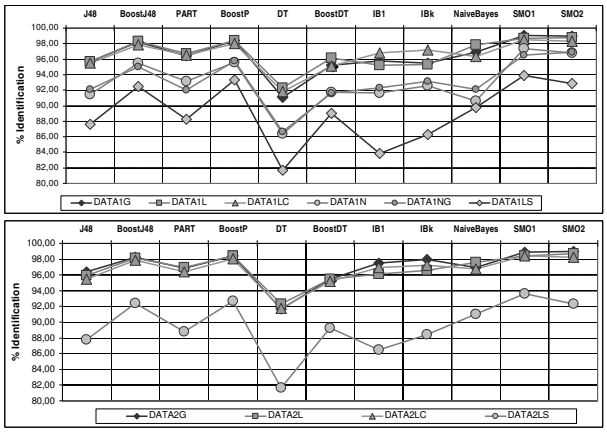

Figure 2: Identification percentage for the ten tested datasets

5. Subjective evaluation

A subjective evaluation is a tool that allows to validate the expressivity of acted speech from a point of view of the users. An exhaustive evaluation of corpus would be excessively tedious (the corpus has 4635 utterances). For each style, 96 utterances have been chosen, having done a total of 480. This test set has been divided in 4 subsets, having 120 utterances each one. An ordered pair of subsets has been assigned to each subject. Therefore, 12 different tests have been generated. The allocation of ordered pairs tries to compensate the fact that the second test could be easier to evaluate due to the previous training.

A forced answer test has been designed with the question *¿What emotional state do you recognize from the voice of the*

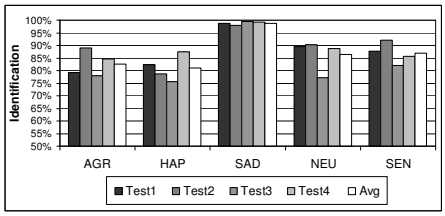

Figure 3: Percentage of identification depending on the test

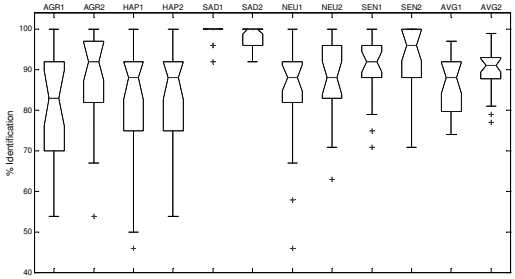

Figure 4: Boxplots of sorted pairs first-second round depending on the style and the average

speaker in this phrase?. The possible answers are the 5 styles of the corpus plus one more option *Don't know / Another*, with the objective of avoid insecure or erroneous answers for the confusing cases. Adding this option has the risk that some evaluators use excessively this answer to accelerate the end of the test [12]. However, this effect has not been considerable in this test.

The process of evaluation has been carried out on a web platform developed for this type of tests, that permits to leave the test and to resume it subsequently. The evaluators belong mainly to the staff of *Enginyeria i Arquitectura La Salle* with a quite heterogeneous profile. Only the results of the 26 volunteers who finished the two tests have been taken into account. The results of the subjective test show that all the styles achieve a high percentage of identification. The figure 3 shows the percentage of identification by style and test, being the sad style the best rated (98.8% of average), followed by sensual (86.8%) and neutral (86.4%) styles, and finally the aggressive (82.7%) and happy (81%) ones.

The confusion matrix (table 3), shows that the main errors are in the agressive style (14.2% identified as happy) and the happy one (15.6% identified as aggressive). Moreover, neutral style is confused slightly with all and there is certain confusion of sensual with sad (5.7%). If we compare these results with the confusion matrix for the best average rated dataset (table 2), we can conclude that the algorithms confuse mainly sensual with neutral, however subjects show confusions between happy and aggressive. This difference is due to the lack of voice quality parameters because sadness and neutral have similar prosody, but sensual voice is most whispered than neutral, a difference which is clearly noticed by the subjects. Also, the influence of order has been studied. In average, the second round obtains better results than the first, especially for neutral, sensual, and aggressive styles (see figure 4).

6. Conclusion and future work

In this paper, the production of an oral corpus oriented to expressive speech synthesis has been presented. We have performed subjective (listening test) and objective (automatic emotion identification) evaluation in order to validate its expressive content showing good results. The advantage of the automatic experiments is that they are performed over the whole corpus, while the listening test comprises a subset of utterances.

In future, we will introduce voice quality parameterization in addition to prosody to minimize the confusion between sensual and neutral styles. Moreover, this work should serve to analyze the bad classified utterances in order to eliminate them and to improve the latter modelling and synthesis processes.

Table 2: Average confusion matrix for the automatic identification experiment with Data2G and the eleven algorithms

	Agr	Hap	Sad	Neu	Sen
AGR	**99.1%**	0.8%	0.1%	0.0%	0.0%
HAP	1.6%	**97.1%**	0.0%	1.2%	0.2%
SAD	0.2%	0.1%	**99.3%**	0.4%	0.1%
NEU	0.2%	0.9%	0.4%	**93.9%**	4.5%
SEN	0.0%	0.1%	0.2%	4.9%	**94.8%**

Table 3: Average confusion matrix for the subjective test

	Agr	Hap	Sad	Neu	Sen	Dk/A
AGR	**82.7%**	14.2%	0.1%	1.8%	0.1%	1.1%
HAP	15.6%	**81.0%**	0.1%	1.9%	0.2%	1.2%
SAD	0.0%	0.0%	**98.8%**	0.5%	0.6%	0.1%
NEU	5.3%	1.3%	0.7%	**86.4%**	3.6%	2.7%
SEN	0.0%	0.1%	5.7%	4.7%	**86.8%**	2.6%

7. References

[1] N. Campbell, "Databases of emotional speech," *Proceedings of the ISCA Workshop on Speech and Emotion*, pp. 34–38, September 2000.

[2] R. Cowie, E. Douglas-Cowie, and C. Cox, "Beyond emotion archetypes: databases for emotion modelling using neural networks," *Neural Networks*, vol. 18, pp. 371–388, 2005.

[3] L. Devillers, L. Vidrascu, and L. Lamel, "Challenges in real-life emotion annotation and machine learning based detection," *Neural Networks*, vol. 18, pp. 407–422, 2005.

[4] D. Ververidis and C. Kotropoulos, "Emotional speech recognition: Resources, features, and methods," *Speech Communication*, vol. 48, no. 9, pp. 1162–1181, September 2006.

[5] E. Douglas-Cowie, N. Campbell, R. Cowie, and P. Roach, "Emotional speech: towards a new generation of databases," *Speech Communication*, vol. 40, pp. 33–60, April 2003.

[6] M. Schröder, "Speech and emotion research: An overview of research frameworks and a dimensional approach to emotional speech synthesis," Ph.D. dissertation, PHONUS 7, Research Report of the Institute of Phonetics, Saarland University, 2004.

[7] N. Campbell, "Developments in corpus-based speech synthesis: Approaching natural conversational speech," *IEICE - Trans. Inf. Syst.*, vol. E88-D, no. 3, pp. 376–383, 2005.

[8] N. Montoya, "El papel de la voz en la publicidad audiovisual dirigida a los niños," *Zer. Revista de estudios de comunicación*, no. 4, pp. 161–177, 1998.

[9] H. François and O. Boëffard, "The greedy algorithm and its application to the construction of a continuous speech database," in *Proceedings of LREC*, vol. 5, May 29-31 2002, pp. 1420–1426.

[10] R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz, and J. G. Taylor, "Emotion recognition in human computer interaction," *IEEE Signal Processing*, vol. 18, no. 1, pp. 33–80, January 2001.

[11] F. Alías, C. Monzo, and J. C. Socoró, "A pitch marks filtering algorithm based on restricted dynamic programming," in *International Conference on Spoken Language Processing (ICSLP)*, pp. 1698–1701.

[12] E. Navas, I. Hernáez, and I. Luengo, "An Objective and Subjective Study of the Role of Semantics and Prosodic Features in Building Corpora for Emotional TTS," *IEEE Trans. on Audio, Speech and Language Processing*, vol. 14, no. 4, pp. 1117–1127, July 2006.

[13] A. Schweitzer and B. Möbius, "On the structure of internal prosodic models," in *Proceedings of the 15th International Congress of Phonetic Sciences (Barcelona)*, 2003, pp. 1301–1304.

[14] I. H. Witten and E. Frank, *Data Mining: Practical Machine Learning Tools and Techniques*, 2nd ed. San Francisco: Morgan Kaufmann, 2005.

[15] P.-Y. Oudeyer, "The production and recognition of emotions in speech: features and algorithms," *Int. Journal of Human Computer Interaction*, vol. 59, no. 1-2, pp. 157–183, 2003, special issue on Affective Computing.