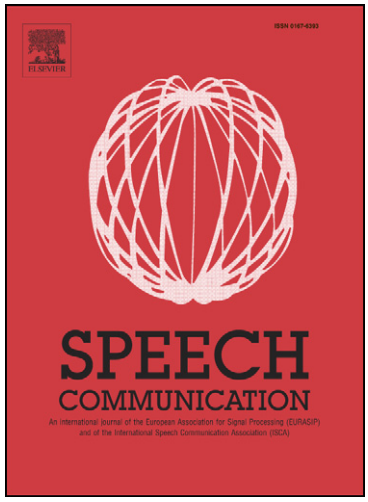
Automatic refinement of an expressive speech corpus assembling subjective perception and automatic classification

Ignasi Iriondo, Santiago Planet, Joan-Claudi Socoró, Elisa Martínez, Francesc Alías, Carlos Monzo



Automatic refinement of an expressive speech corpus assembling subjective perception and automatic classification

Ignasi Iriondo, Santiago Planet, Joan-Claudi Socoró, Elisa Martínez, Francesc Alías, Carlos Monzo

PII: S0167-6393(08)00181-7
DOI: 10.1016/j.specom.2008.12.001
Reference: SPECOM 1767

To appear in: *Speech Communication*

Received Date: 1 December 2007
Revised Date: 2 December 2008
Accepted Date: 4 December 2008

Please cite this article as: Iriondo, I., Planet, S., Socoró, J-C., Martínez, E., Alías, F., Monzo, C., Automatic refinement of an expressive speech corpus assembling subjective perception and automatic classification, *Speech Communication* (2008), doi: 10.1016/j.specom.2008.12.001

Automatic refinement of an expressive speech corpus assembling subjective perception and automatic classification

Ignasi Iriondo*, Santiago Planet, Joan-Claudi Socoró, Elisa Martínez, Francesc Alías, Carlos Monzo

GPMM - Grup de Recerca en Processament Multimodal
Enginyeria i Arquitectura La Salle, Universitat Ramon Llull
C/ Quatre Camins 2, 08022 Barcelona (Spain)

Abstract

This paper presents an automatic system able to enhance expressiveness in speech corpora recorded from acted or stimulated speech. The system is trained with the results of a subjective evaluation carried out on a reduced set of the original corpus. Once the system has been trained, it is able to check the complete corpus and perform an automatic pruning of the unclear utterances, i.e. with expressive styles which are different from the intended corpus. The content which most closely matches the subjective classification remains in the resulting corpus. An expressive speech corpus in Spanish, designed and recorded for speech synthesis purposes, has been used to test the presented proposal. The automatic refinement has been applied to the whole corpus and the result has been validated with a second subjective test.

Key words: Expressive speech databases, Expression of emotion, Speech technology, Expressive speech synthesis

1. Introduction

There is a growing trend towards the use of speech in human-machine interaction. In this field, the inclusion of automatic emotion recognition or expressive speech synthesis can improve communication by making it more natural. One of the most important challenges in the study of expressive speech is the development of oral corpora with authentic emotional content. The naturalness of the locutions depends on the strategy used to obtain the emotional speech. The main debate centres on the relation between authenticity and the degree of control during the recording. Campbell (2000) and Schröder (2004) proposed four emotional speech sources: *i)* natural occurrences obtained from spontaneous human interaction which present the most natural emotional speech, although this approach does have some drawbacks due to the lack of control of its content, i.e. insufficient quality of sound and the difficulty in labelling; *ii)* elicitation of authentic emotions in the laboratory is a way of compensating some of the problems detected in natural situations, although full-blown authentic emotions are rarely achieved; *iii)* stimulated emotional speech created by reading texts with verbal content related to the target emotion. The difficulty of comparing utterances with different texts can be counteracted by increasing the corpus size so that statistical methods can be used to generalize acoustic models of emotion; and *iv)* acted emotional speech by reading the same sentences with different emotions which allows direct comparisons of prosody and voice quality. However, the greatest obstacle in this case is the lack of authenticity in the expressed emotion. The databases used for conducting emotional speech synthesis are usually based on acted speech (Douglas-Cowie et al., 2003) where a professional speaker reads a set of texts (with or without emotional content) and simulates the emotions to be produced.

Another important aspect is the purpose of the

*Correponding author. Tel.: +34 932902452; fax: +34 932902470.

Email address: iriondo@salle.url.edu (Ignasi Iriondo)

URL: http://www.salle.url.edu/∼iriondo (Ignasi Iriondo)

speech and emotion research. It is necessary to distinguish between the processes of perception (centred on the speaker) and those of expression (centred on the listener) (Schröder, 2004). The objective of the former is to establish the relationship between the emotional state of the speaker and the quantifiable parameters of speech that usually deal with the recognition of emotions from the speech signal. The latter are focused on shaping the speech parameters in an effort to transmit a certain emotional state. According to Devillers et al. (2005), one of the main challenges is the identification of oral indicators which are attributable to the emotional content and not simple characteristics of conversational speech.

The paper proposes an expressive speech corpus from stimulated speech which provides enhanced expressiveness. Although the corpus has been designed for use in different tasks related to research in expressive speech synthesis, this work combines techniques commonly applied in processes of perception and expression. On one hand, the production of the corpus follows the guidelines of listener-centred studies given the fact that it is oriented to speech synthesis. On the other hand, we apply techniques of emotion recognition in order to validate its expressive content and refine the recorded speech database. The refinement is conducted by an automatic system designed to emulate the subjective criteria used in the identification of emotions from speech. The system is trained with the results of a subjective evaluation, which is carried out on a small part of the corpus. Once the system has been trained, it is able to check the complete corpus content and perform an automatic pruning of the unclear utterances, i.e. with expressive styles that are different from the intended. The proposed refinement is performed on the corpus described in the paper and the final results are validated through a second subjective test.

A review of Devillers et al. (2005); Cowie et al. (2005); Ververidis and Kotropoulos (2006) provides a deep insight into databases and feature extraction methods covering recent advances in speech and emotion research. These tutorials review previous studies such as those of Murray and Arnott (1993); Cowie et al. (2001); Scherer (2003); Douglas-Cowie et al. (2003); Schröder (2004). To the best of our knowledge, the automatic refinement of an expressive speech corpus in order to obtain a content that closely correlates with subjective classification has not yet been proposed.

The remainder of this paper is organized as follows: Section 2 presents the design, recording and segmentation of a stimulated expressive speech corpus oriented to speech synthesis in Spanish. Section 3 describes the subjective evaluation of a reduced but significant set of the corpus by means of a listening test. Section 4 is devoted to an initial objective validation which is carried out by using techniques of automatic emotion identification. Section 5 describes the proposed method to refine the content of the recorded speech database. There is a general discussion in Section 6 and, finally, Section 7 provides conclusions and outlines future work.

2. Speech corpus development

A new expressive oral corpus oriented to expressive speech synthesis in Spanish has been developed with a twofold purpose: firstly, to be used in the acoustic modelling (prosody and voice quality) of emotional speech, and secondly, to be the speech unit database for the synthesizer.

The *Spanish Emotional Speech* (SES) database (Montero et al., 1998) and the *Interface Emotional Speech Synthesis Database* (IESSDB) (Nogueiras et al., 2001; Hozjan et al., 2002) are two examples of existing acted speech oriented to emotional speech synthesis in Spanish. Both databases were validated on the basis of subjective tests whose results were considered sufficient to guarantee the expressiveness of the recordings. The SES database included four emotions (sadness, happiness, anger and surprise) as well as a neutral expression. In the subjective listening test almost 90% of the emotions were correctly identified, except in the case of happiness where the figures dropped to 74%. The IESSDB composed of six emotions and a neutral expression was also recorded in French, English and Slovenian. To evaluate the authenticity of the emotional content, a subjective test was also carried out. Subjects could select two emotions for each utterance presented. The global identification was about 80% for the first option and it reached 90% when the second option was also considered.

2.1. Corpus design

When creating an oral corpus, the first stage is the definition of the associated tasks and the design of the key elements which will determine the final quality. Corpus design depends on the targets set and the limitations which may be applied. In this

section two aspects are described: the theoretical objectives defined for the production of an oral corpus oriented to expressive speech synthesis and the practical approach considered in the process design.

2.1.1. Theoretical objectives

There are four theoretical objectives that should be covered by the oral corpus in order to be useful for expressive speech synthesis:

a. Naturalness and audio quality. Ensuring naturalness (understood as the characteristic of transmitting the emotional state of the speaker) is the most important condition of an expressive speech corpus. Although spontaneous utterances are considered the most natural kind of speech (Campbell, 2000), they present two important drawbacks when used in speech synthesis: the content can not be predefined and the recording environment conditions can affect audio quality. Therefore, the quality of the recording is a priority, although it should not be an obstacle in achieving the authenticity required to simulate expressive speech.

b. Expressive coverage. An expressive speech corpus should cover a wide range of emotions, attitudes and moods of one or more speakers. This objective can be achieved by creating a very large corpus (over 1000 hours) by recording daily situations (Campbell, 2002), thus allowing a subsequent synthesis near to natural speaking (Campbell, 2005). However, it may take several years just to obtain a sufficient number of speech utterances. Nevertheless, a smaller corpus with the appropriate coverage may be enough for synthesis purposes and would consequently reduce costs and development time.

c. Phonetic and prosodic coverage. Corpus based speech synthesis systems require speech databases with a large number of phonetic units and variations of the linguistic features to be reproduced in the synthesis stage (François and Boëffard, 2002). As in the previous objective, while it is important to control the size of the resulting database, an adequate coverage of phonetic units and prosodic variability must be ensured. Semiphonemes, diphonemes and triphonemes are the most frequently used units in concatenative speech synthesis systems. While the former are not useful in high quality speech synthesis, it is difficult to consider the latter to achieve a total coverage (Bozkurt et al., 2003). Thus, it is essential to cover the most usual diphonemes and some triphonemes and to consider the phonetic frequency distribution for a specific language. The prosodic coverage will be achieved by guaranteeing a specific variety of intonation patterns by selecting declarative, interrogative and exclamation expressions (Iriondo et al., 2007c).

d. Suitable semantic content. To create stimulated expressive speech, the text should include good phonetic and prosodic coverage as well as a suitable semantic content which helps to express the desired styles. An increase in the number of styles makes the design of the text more difficult (Navas et al., 2006) and therefore, it is always preferable to start with a rich textual corpus.

2.1.2. Practical approach

In order to achieve the theoretical objectives described in Section 2.1.1 we sought the help of experts from the Laboratory of Instrumental Analysis (LAICOM) of the Autonomous University of Barcelona (UAB).

The corpus described in this article was constructed with reading texts whose semantic content helped to express the desired style (stimulated speech). The texts for each expressive style were read by a professional female speaker in different recording sessions. It is assumed that this strategy diminishes the possibilities of modelling informal spontaneous speech utterances, while guaranteeing the control of the recording environment, the style definition and the text design.

To select the texts semantically related to different expressive styles, we made use of an existing textual database of advertisements extracted from newspapers and magazines. This database was previously classified into subject categories. Based on a prior study on audiovisual advertising (Montoya, 1998), some of these categories were considered to be indicative of promoting different expressive styles.

It is important to mention that the speaker had previously received training in the vocal patterns of each style. The phonetic features (segmental and supra-segmental) for these vocal patterns were defined by the experts of LAICOM. The use of texts from an advertising category aimed to help the speaker to maintain the desired style through the whole recording session. Therefore, the intended style was not performed according to the speaker's criteria for each sentence, but all the utterances of the same style were consecutively recorded in the same session following the previously learned pattern. Thus the speaker was able to keep the required expressiveness even with texts whose seman-

tic content were not coherent with the style. Moreover, an expert supervision was required through the recording in order to avoid possible deviations from the predefined style.

Five subject categories were selected from the advertising corpus and their assignment to expressive speech styles was the following:

- New technologies: a neutral style (NEU) which transmits certain maturity.
- Education: a happy style (HAP) which generates a feeling of extroversion.
- Cosmetics: a sensual style (SEN) based on a sweet voice.
- Automobiles: an aggressive style (AGR) which transmits hardness.
- Travel: a sad style (SAD) which seeks to express melancholy.

The definition of these five styles aimed to provide a sufficient expressive diversity in order to advance our research in different topics related to expressive speech. There was not any category suitable for sad style and, although its assignment to travel category was slightly artificial, it was accepted because sadness is the easiest style to simulate.

A set of phrases for each category was selected by means of a greedy algorithm (François and Boëffard, 2002) which permitted the selection of phonetically balanced sentences from each subcorpus. In addition to looking for a phonetic balance, sentences that contain exceptions (foreign words and abbreviations) were discarded because they make the automatic processes of phonetic transcription and labelling more difficult. Moreover, the selection of sentences which were similar to those previously selected was penalized by the greedy algorithm. To optimize the selection process, the required phonemes were sorted according to occurrence rate, allowing the greedy algorithm to start by selecting sentences that contain less probable phonemes. This occurrence rate was studied from Pérez (2003). Tables 1 and 2 show the phonetic distribution of Spanish vowels and consonants respectively, compared with the average of the five studies presented in Pérez (2003).

Table 1: Frequency distribution of Spanish vowel phonemes in the designed corpus compared to the average of the five studies presented in Pérez (2003)

	/a/	/e/	/i/	/o/	/u/
Designed Corpus	12.74	13.56	6.13	9.24	2.74
Average 5 studies	13.27	13.13	6.32	9.71	2.32

The phonetic symbols are in SAMPA notation (Wells, 1993).

2.2. Recording

The recording of the oral corpus was carried out in the recording studio of Enginyeria i Arquitectura La Salle, Universtitat Ramon Llull (EALS-URL). This studio has two zones: the control room, which has mixing and production equipment, and the recording room. Both of them are acoustically conditioned to offer the required conditions and high level of isolation. The recording room is a non-square surface of 20 m^2 and is 3.5 meters high. The response time of the room is 0.8 seconds but the position of the speaker and the microphone ensures that there are not audible echoes.

A high quality condenser microphone (AKG C-414) was used. It has a flat response (2 dB in 20-20000 Hz range) and its signal-noise ratio is 80 dBA SPL. The recording was stored in a hard disk through the Pro Tools 5.1 digital platform in a Mac G5 computer using a Yamaha 02R digital console. The speech signal is digitalized in 48 kHz-24 bits WAV files.

Four people were involved in the protocol established for the recording sessions in order to optimize subsequent processes: the professional speaker; an audio engineer who was responsible for adjusting the recording platform including the position of the microphone and the speaker; an audiovisual communication expert who trained the speaker and corrected her deviations from the required expressive model (see Section 2.1.2); and a control technician to guarantee that the speaker read the sentences properly.

2.3. Segmentation

Firstly, the recorded corpus must be manually processed in order to obtain the best version of each utterance in the event of some of them being repeated by the speaker. Secondly, the master audio file is divided in blocks in order to be automatically processed. The speech recognition tool HTK[1] was

[1] http://htk.eng.cam.ac.uk/

Table 2: Frequency distribution of Spanish consonant phonemes in the designed corpus compared to the average of the five studies presented in Pérez (2003)

	/p/	/t/	/k/	/b/	/d/	/g/	/n/	/m/	/J/
Designed Corpus	2.70	4.82	3.84	2.67	4.59	0.99	6.27	3.44	0.22
Average 5 studies	2.66	4.66	4.02	2.66	4.58	1.02	5.3	2.73	0.28
	/s/	/x/	/C/	/T/	/r/	/R/	/l/	/L/	/f/
Designed Corpus	7.51	0.85	0.24	1.83	5.72	0.92	4.99	0.32	0.81
Average 5 studies	8.72	0.65	0.34	1.89	4.48	0.69	4.86	0.57	0.74

The phonetic symbols are in SAMPA notation (Wells, 1993).

used for this purpose. A forced time alignment was conducted to segment each block into utterances by means of phonetic transcription and using Hidden Markov Models. An subsequent manual review is required to correct badly segmented sentences, something that can occur due to the lack of consistency between the pausing made by the speaker and the punctuation marks. This review process consists in adding the corresponding punctuation mark to both the text and the phonetic transcription when there is a misplaced silence in the audio file. Alternatively, the punctuation mark can be removed from the text and the phonetic transcription when the speaker does not make the pause.

The final recorded corpus has 4,638 sentences and is 5 hours 27 minutes long. More specifically, 833 sentences are of neutral style (50 minutes), 916 of happy style (56 minutes), 841 of sensual style (51 minutes), 1,000 of sad style (86 minutes), and 1,048 of aggressive style (84 minutes). An extension of the neutral, happy and sensual styles is planned in order to balance the length of all the sub-corpora.

The previously described time alignment was also used to segment the sentences into phonemes. The sequence of marks that delimit the phonemes is the base for the acoustic analysis of the corpus allowing parameterization at segmental level (see Section 4.1).

3. Subjective evaluation

Subjective evaluation enables us to validate the expressiveness of acted speech from the user's viewpoint (e.g. Montero et al. (1998); Hozjan et al. (2002); Navas et al. (2006); Morrison et al. (2007)). When a large speech corpus is compiled, we must make sure that all the utterances are consistent with the expressive category definition. However, for this kind of corpus, an exhaustive evaluation of the whole corpus would be extremely costly, and would involve many evaluators. Therefore, we propose a partial evaluation of the corpus through a listening test which aims to obtain a representation of subjective perception. These results provide us with a first glance of the expressive quality of the recorded corpus and will also be used in the final approach for validating the whole corpus (Section 5).

3.1. Test design

The recorded corpus is longer than 5 hours and we have only considered almost ten percent of the utterances to be evaluated in the listening test. 96 utterances have been randomly chosen from each expressive style, which provides a total of 96*5=480. The listeners are 25 volunteers from EALS-URL. The evaluation of 480 utterances by each person would be too hard. To resolve this problem, this test set was divided in four subsets with each containing 120 utterances. An ordered pair of subsets was assigned to each subject, generating 12 different combinations (perhaps some listeners had to evaluate firstly the 3rd subset and after the 2nd one while others did the 2nd subset firstly and after the 3rd one). The distribution in ordered pairs aims to prevent second round tests being easier than the first round tests due to the previous training received. The results of each test will come from both those who did it in the 1st round and others that took it in the 2nd round.

A forced answer test was designed with the question *"What emotion do you recognize from the voice of the speaker in this utterance?"*. The possible answers were the 5 styles of the corpus (see Section 2.1.2) plus the additional option of *Don't know/Another* (Dk/A) to avoid biasing the results in the case of confusion or doubts between two options. The risk of adding this option is that some evaluators may use it excessively to accelerate the test (Navas et al., 2006). However, this effect was negligible in this test (see right column of Table 3).

3.2. Evaluation process and results

The subjective test was carried out on a web platform (Planet et al., 2008) designed for this type of experiments. The platform permits the user to stop the test at any given moment and resume it at a later time, thus minimizing the effects of tiredness. Evaluators with a reasonably heterogeneous profile assessed an ordered pair from the four subsets. The results of the 25 volunteers who finished the two assigned tests have been reported.

As a general result, the subjective test shows that all the expressive styles achieve a high percentage of identification (87.1% on average). Figure 1 shows the percentage of identification by style and test, with SAD being the most highly rated (98.8%), followed by SEN (86.8%) and NEU (86.4%) styles, and finally AGR (82.7%) and HAP (81%). The confusion matrix shown in Table 3 reflects the misclassifications. It reveals that the main errors are produced in AGR (14.2% identified as HAP) and HAP (15.6% identified as AGR). Moreover, NEU is slightly confused with all the options and there is a certain level of confusion of SEN with SAD (5.7%) and NEU (4.7%). The DK/A option was hardly used, although it was more present in NEU and SEN than in the rest of the styles.

The influence of the order of the tests has been also studied. On average, the second round obtains slightly better results (90%) than the first one (86%), especially for neutral, sensual, and aggressive styles, while the sad style is well recognized in both of them, as can be observed in Figure 2. This slight difference (4%) reveals firstly, that both versions of each test are required to evaluate the perception from the user's viewpoint and secondly, that evaluator fatigue does not contribute to overall results.

Figure 3 shows two histograms based on the global percentages of (a) correctly classified utterances and (b) the use of the *Don't know/Another* option. From these results, it is difficult to establish which utterances could be considered unclear from an expressiveness viewpoint. However, these histograms are used to define a simple heuristic rule in order to decide if a sentence was correctly performed by the speaker, as described in Section 4.

4. Automatic speech style identification

As previously stated, some utterances may lack the desired expressiveness when made by a professional speaker (stimulated speech) instead of recording spontaneous utterances. If these unclear utterances remain in the corpus, some drawbacks may occur in both the acoustic modelling and the speech synthesis process. An exhaustive manual review of the whole corpus would be very time-consuming given its length. It would therefore be of great use to develop an automatic system able to validate all the recorded sentences in terms of expressiveness.

The experiments described in this section have the initial goal of validating corpus expressiveness by using automatic emotion recognition techniques which involve the application of machine learning (ML) algorithms to statistics calculated from the acoustic parameterization. Firstly, the acoustic analysis of the speech corpus is detailed and secondly, a style identification experiment by combining different datasets and ML algorithms is explained.

Table 3: Average confusion matrix for the subjective test

Answer →	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%

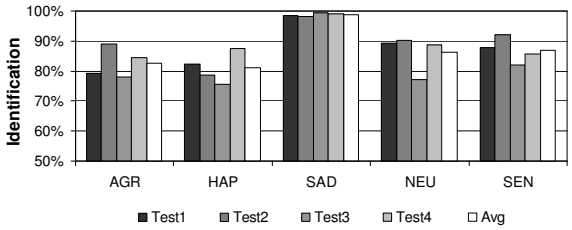


Figure 1: Percentage of identification per style for the four subsets in the listening test and their average.

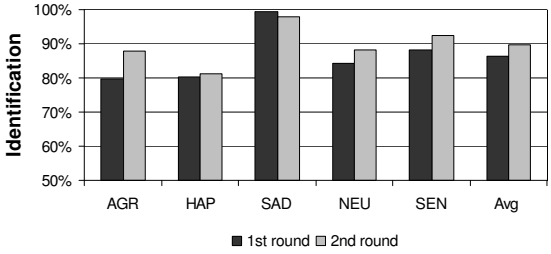
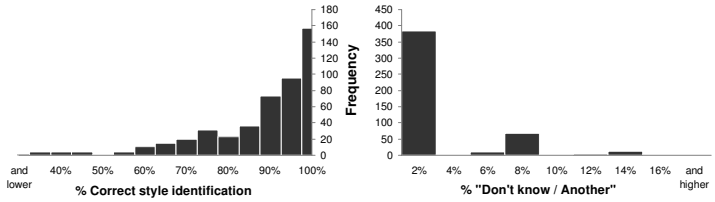


Figure 2: Percentage of identification in the first and second rounds of the listening test according to the style and the average of the five styles.



(a) (b)

Figure 3: Histograms for the correct identification and *Don't know/Another* percentages in the subjective test.

4.1. Acoustic analysis

Prosodic features of the speech signal (fundamental frequency, energy, duration of phonemes and frequency of pauses) and voice quality (VQ) parameters are related to vocal expression of emotion (Cowie et al., 2001). In this section, both parameterizations are explained. Initially, only prosodic parameters are taken into account (Sections 4.2 and 5.1), but afterwards VQ parameters are introduced in an effort to discriminate expressive styles with prosodic similarity more precisely (Section 5).

4.1.1. Prosody parameters

The prosodic parameters measured in this work are related to fundamental frequency (F0), energy and rhythm, as described below.

F0 related parameters: F0 analysis is based on the result of the pitch marker described in Alías et al. (2006). Three sequences of local F0 values are computed for each utterance (complete, excluding silences and unvoiced sounds and considering only the stressed vowels). To obtain information about the voiced/unvoiced segments and also about silences and stressed vowels phonetic labelling is used. Furthermore, F0 is computed in linear and logarithmic scales.

Energy related parameters: Speech is processed with 20 ms rectangular frames every 10 ms, calculating the mean energy in linear and logarithmic scale (dB). Three sequences have been generated per sentence, as in the F0 analysis (complete, excluding silences and unvoiced sounds, and considering only the stressed vowels).

Rhythm related parameters: Although the duration of phonemes is an important aspect of vocal expression of emotion (Cowie et al., 2001), it is not considered in some studies because of the complexity of obtaining it automatically (Navas et al., 2006). Modelling of duration in text-to-speech conversion systems tends to be based on z-score measure (Campbell and Isard, 1991) to predict the individual duration of segments while controlling the speech speed. This measure of each phoneme is calculated to analyze the temporal structure of speech, as in Schweitzer and Möbius (2003):

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

where μ and σ are the mean and the standard deviation, respectively. Both parameters are estimated for each phoneme from the full corpus. Therefore, the rhythm of a sentence is represented by means of a vector with the z-score values of each phoneme. Furthermore, another version of this vector is generated considering only the stressed vowels. In addition, two pausing related parameters are measured for each sentence: number of pauses per second, and percentage of silence time with respect to the duration of the sentence. Both parameters represent the frequency and duration of pauses.

4.1.2. Voice quality parameters

Some voice quality (VQ) parameters have been calculated by means of Praat analysis software[2] from the audio recordings without using either invasive transducers or extra hardware according to the proposal of Drioli et al. (2003) and the implementation described in Monzo et al. (2007):

Jitter: It measures the cycle-to-cycle variations of the fundamental period averaging the magni-

[2] http://www.praat.org/

tude difference of consecutive fundamental periods, divided by the mean period.

Shimmer: It measures the cycle-to-cycle variations of amplitude by averaging the magnitude difference of the amplitudes of consecutive periods, divided by the mean amplitude.

GNE (*Glottal-to-Noise Excitation Ratio*): Ratio between the excitation due to vocal chord vibrations and the excitation due to turbulent noise. Unlike other parameters such as HNR (Harmonic-Noise Ratio) or NNE (Normalized Noise Energy), GNE is almost independent from Jitter and Shimmer (Michaelis et al., 1997).

HammI (*Hammarberg Index*): It is defined as the difference between the maximum energy in the 0-2000 Hz and 2000-5000 Hz frequency bands.

Do1000: is a linear approximation of the spectral tilt above 1000Hz, calculated using the least squares method.

In Monzo et al. (2007), these VQ parameters were used to discriminate among the five expressive speech styles described in this paper and acceptable results were obtained. However, it was concluded that prosodic information was essential in order to improve the results.

4.2. Preliminary objective evaluation

This first stage of the investigation was centred on the application of a selection of data mining techniques on different datasets for automatic expression classification. These datasets consisted of several statistics related to prosodic parameters from each utterance. The details of this study can be found in Iriondo et al. (2007b), but a brief description follows.

As described in Section 4.1, the prosodic information of an utterance is represented by the sequences of values per phoneme of F0 (linear and log), energy (linear and log) and normalized duration (z-score). There are three versions for each sequence of F0 and energy values and two versions for duration. Moreover the first and second discrete derivatives are calculated. For each sequence, the following statistics are measured: mean, variance, maximum, minimum, range, skew, kurtosis, quartiles and interquartilic range. Therefore, the result of the total number of parameters per utterance rises to 464 when the pausing parameters are included (see Table 4). For example, for the F0 row, there are 198 parameters per utterance resulting from the product of two units (linear and log), 3 sequences (complete, excluding silences and unvoiced sounds, and considering only the stressed vowels), 3 functions (original, first and second discrete derivatives) and 11 statistics.

Table 4: Detail of the different kind of parameters used to represent the prosody of each utterance for the initial dataset named Data1

	Unit	Sequences	Functions	Statatistics	Total
F0	2	3	3	11	198
Energy	2	3	3	11	198
Duration	1	2	3	11	66
Pausing	2	-	-	-	2
TOTAL					464

The original set of parameters, Data1, is divided into smaller subsets in order to reduce the dimensionality of the problem. The datasets obtained and the strategy followed in their generation is represented in Figure 4. Data2 is Data1 without all the parameters related to the second derivative. For the remaining datasets, G indicates a reduction in dimensionality based on genetic algorithms[3] (Goldberg, 1989), L shows log versions of F0 and energy, N is a subset of parameters similar to the one presented in Navas et al. (2006), C illustrates the complete sequence, i.e. with all the phonemes of the sentence and finally S datasets only consist of the parameters computed for the stressed vowels.

The preliminary objective validation of the corpus expressiveness is based on Oudeyer (2003) who experimented with a large number of ML algorithms and datasets to recognize basic emotions in short utterances. As in the cited work, Weka implementations of ML algorithms (Witten and Frank, 2005) were used in adherence with a 10-fold cross-validation strategy. Some of them were also considered in their boosted versions in an effort to improve the results despite greater computational costs (Duda et al., 2001). Table 5 shows the results of the experiments carried out on the best al-

[3]The search for the optimal subset of attributes is performed by a genetic algorithm using a correlation-based method for evaluating the worth of them. This combination selects subsets maximizing the correlation with the class but with a low redundancy between the attributes of the subset. This allows choosing the most relevant attributes independently of the classification schema.

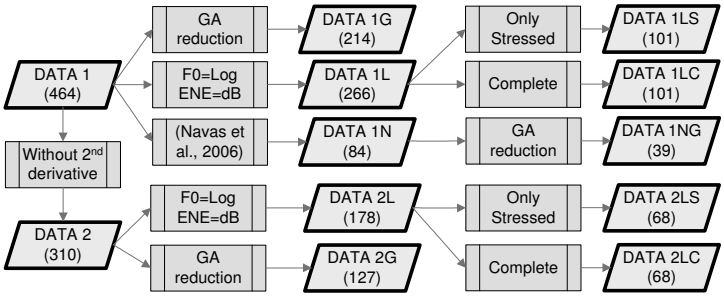


Figure 4: Generation of different datasets.

Table 5: Accuracy mean and maximum values for the learning algorithms used in the experiment of automatic expression identification from different speech datasets

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

gorithms. The third column represents the global average identification percentage of each algorithm with a confidence level of 0.95. The fourth column represents the best result and the dataset.

Figure 5 compares the identification percentage obtained by different datasets depending on the classification algorithm. Both versions of Support Vector Machines (SMO in Weka) obtained the best average results, and also the best identification percentage with Data1-G. Other algorithms such as the boosted versions of J4.8 and PART, and also Naïve Bayes, provide a suitable solution to this problem. The fact that Data1-LC gives almost the same good results as Data1-G and Data1-L, but with less than half of parameters is noteworthy. A similar effect can be observed in datasets without the second derivative, i.e. Data2-LC achieves practically the same results as Data2-G and Data2-L. A more exhaustive analysis of these results can be found in Iriondo et al. (2007b).

Table 6 shows the confusion matrix for the results from the eleven classifiers with the best dataset on average, i.e. Data2-G (97.02% ± 1.23). A certain confusion between neutral and sensual styles can be observed, as well as between happy-aggressive and neutral-happy, although to a lesser extent. Qualitatively, the same pairs were confused by human subjects in the subjective test (see confusion matrix shown in Table 3). However, unlike the automatic system, where the main confusion is in the sensual-neutral pair, on a quantitative basis, people were prone to confuse happy with aggressive styles.

Table 6: Average confusion matrix for the automatic identification with Data2-G

Answer →	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%**

5. Automatic refinement according to subjective criteria

If we consider the automatic emotion recognition, the results shown in Section 4.2 could be considered to be excellent. However, the proposed objective was to validate the authenticity of the recorded cor-

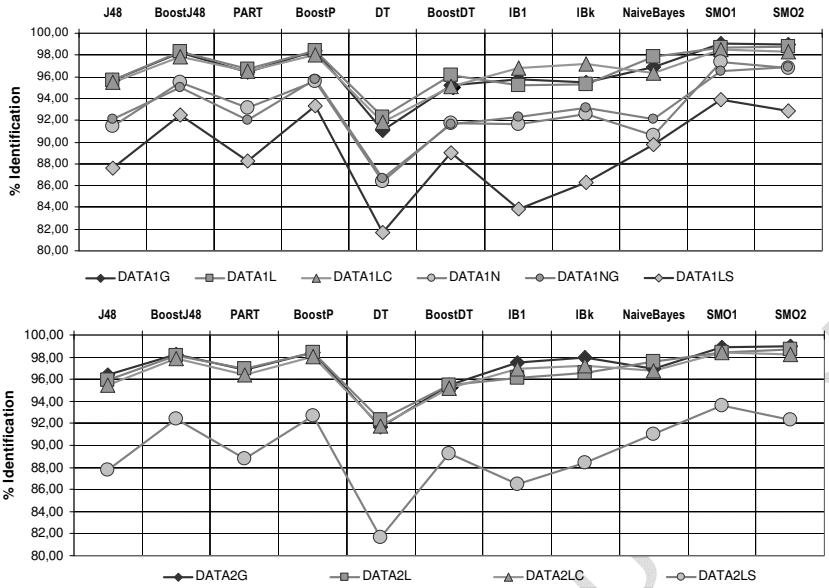


Figure 5: Identification percentage for the ten tested datasets.

pus, and the subjective test reveals that a small percentage of wrong utterances from the expressiveness perspective is not discarded by the automatic system. Therefore, a new corpus refinement method is proposed which considers the following hypotheses:

- Subjective test results are more relevant than those achieved by the automatic system. The aim of the corpus is to be used in speech synthesis, i.e. an application centred on the listener and therefore, a sufficient degree of authenticity from the viewpoint of subjective perception is required.

- Systematizing the pruning of utterances with unclear or erroneous expressive content is required as a manual revision would be very time consuming and would not be able to be re-used for other corpora.

- Achieving an automatic system able to emulate the subjective perception is possible by optimizing the current automatic classification system (improvements in the acoustical analysis, the attribute selection and the combination of algorithms).

The proposed system will automatically classify the corpus utterances into two categories depending on their degree of clarity of expressiveness. The proposal is based on an optimum classifier (algorithm and set of attributes) able to generalize the subjective criteria observed in the listening test. The block diagram presented in Figure 6 summarizes the four main tasks involved in the system:

- **Binary subjective classes**. Each assessed utterance can be binary classified from the expressiveness viewpoint if a consensus among evaluators is achieved. To reach this consensus, we defined a heuristic rule based on the histograms extracted from the subjective test (see Figure 3). Utterances were considered incorrect if they obtained an identification percentage lower than 50% or a *Don't know/Another* percentage greater than 12%, because they can be regarded as outliers in the histograms. From the test set, the 33 sentences out of 480 (6.8%) that corresponded to this rule make up the unclear (UC) class. The rest of the utterances are considered part of the clear (CL) class.

- **Binary automatic classification**. For the automatic classification system, if an utterance is misclassified (different style from its tagging), it is considered UC; on the contrary, the correctly identified utterances are assigned to the CL class.

- **Definition of a comparison measure**. Both the expressiveness level of utterances accord-

ing to the subjective criteria and the automatic classification have to be represented in the same way to enable us to compare them. Given that an utterance is represented by a binary class, the F_1 score (Duda et al., 2001) is considered a suitable measure.

- **Adjustment of the system**. In the training process, the automatic classifier system should be optimized to map the subjective criteria according to the evolution of the comparison measure. The elements of the system that could be tuned are the set of attributes and the classification algorithms.

5.1. Preliminary results

As described in Iriondo et al. (2007a), an initial experiment was conducted considering the dataset Data2-LC (see Section 4.2) since it achieved quite good results (96.58%±1.14) with only 68 attributes out of 464 from the original dataset. This dataset only contains prosodic features that were sufficient to get excellent results in an objective emotion recognition experiment. In that experiment, an attribute selection method was developed in order to maximize the similarity between the subjective test results and the automatic classification. An exhaustive search of the optimal subset of attributes was discarded due to its high computational cost. To solve this maximization process, a greedy approach can be considered (Witten and Frank, 2005). Forward Selection (FW) and Backward Elimination (BW) processes were chosen to carry out this task. The former is an attribute selection method that starts without any attributes and add them one at a time. Beginning with an empty dataset, the attribute that improves the performance of the classifier is chosen to be added for the next iteration. In contrast, the latter method begins with the full dataset and the attribute that improves the comparison measure when it is not considered is removed for the next iteration. In both cases, the process stops when no performance improvement is achieved or there are no more attributes to be added or removed, respectively.

According to the schema presented in Figure 6, the three best rated algorithms (SMO, Naïve Bayes and J.48) from Section 4.2 have been chosen, and they have been combined with both FW and BW attribute selection techniques. Along with instance-based learning (kNN), these three kind of algorithms are the most commonly used applications in automatic emotion recognition from speech (e.g. Ververidis and Kotropoulos (2006); Shami and Verhelst (2007); Morrison et al. (2007)). The classifiers are tested with the 480 utterances that were involved in the subjective test (Section 3) once they have been trained with the 4,158 remaining ones. The wrongly classified sentences are assigned to the UC class, and this assignation is compared with the results obtained from the subjective test, which rejected 33 utterances. The F_1 score is computed over this comparison and it is used to guide the selection of attributes.

Table 7 shows the number of attributes that achieve the maximum value of F_1 score, and also the precision and recall. As can be observed, SMO algorithm obtains $F_1 = 0.50$ with the BW attribute selection technique. J48-FW presents the highest number of coincidences (18) but also a high number of misclassifications (51), which is reflected in the value of $F_1 = 0.43$.

Table 7: Maximum values of F_1 with the related precision and recall for the best configuration of each algorithm (SMO, NB, J48) and strategy (FW, BW). The bold number is the minimum of the range with the same maximum F_1

Alg./Strat.	Attributes	Max. F_1	Precision	Recall
SMO / FW	**18**-35	0.49	0.58 (14/24)	0.42 (14/33)
SMO / BW	**15**-16	0.50	0.56 (15/27)	0.45 (15/33)
NB / FW	**43**-44	0.42	0.39 (15/38)	0.45 (15/33)
NB / BW	**47**-49	0.43	0.52 (12/23)	0.36 (12/33)
J48 / FW	**18**	0.43	0.35 (18/51)	0.55 (18/33)
J48 / BW	**17**-20	0.36	0.45 (10/22)	0.30 (10/33)

5.2. Improvements and final approach

Three guidelines have been considered to improve the results shown in Section 5.1:

- Inclusion of voice quality parameters. A study based exclusively on prosody parameters does not accept discrimination between some styles such as sensual and sadness. The study of these utterances indicates that voice quality parameters (see Section 4.1) can be useful to carry out this task.

- New attribute selection strategy. FW and BW strategies can not undo previous decisions. A combined strategy can solve this drawback and try to avoid getting stuck in poor local maxima.

- Classifiers combination. Given the fact that some classifiers are more precise and others

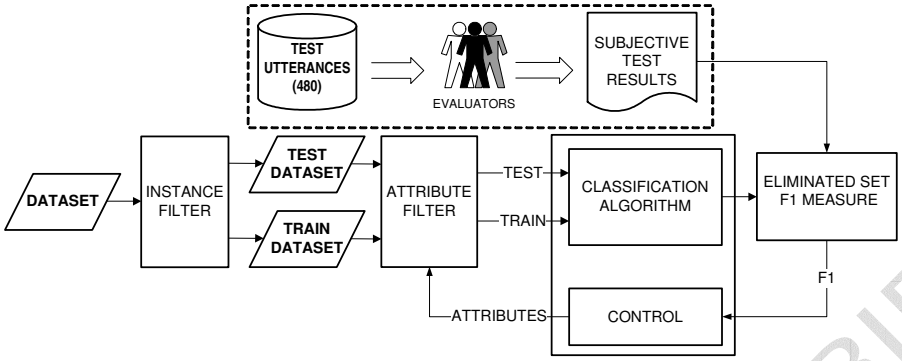


Figure 6: Objective validation adjustment guided by the subjective test results.

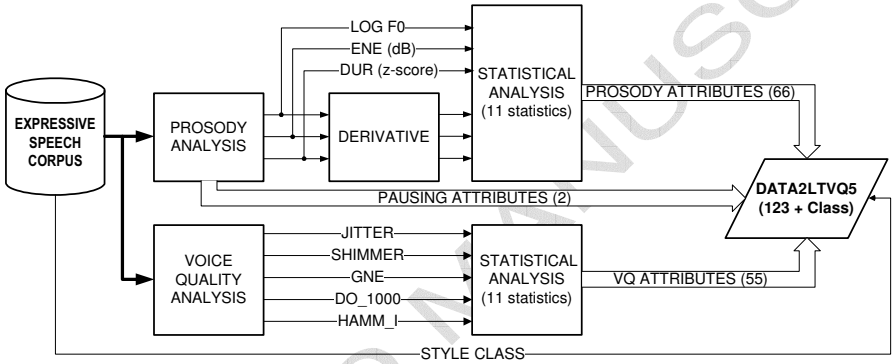


Figure 7: Database generation for the final approach of automatic refinement of the corpus.

show a better recall, a combination of classifiers could improve the final result.

The addition of voice quality parameters according to Section 4.1.2 to the current dataset increases the number of attributes to 123 per utterance (see Figure 7). A pFW-qBW strategy was also considered. The number of p forward and q backward steps can be adjusted (e.g. 3FW-1BW). The third improvement tries to benefit from the different features of each classifier. Experience has shown that combining the predictions from multiple methods often yields more accurate predictions, which can be derived from any one of them (Witten and Frank, 2005). Morrison et al. (2007) uses ensemble methods to improve the emotion recognition from speech. In the stacking strategy, the predictions from different classifiers are used as an input to a meta-learner, which attempts to combine these predictions in an aim to improve the final classification. In our case, the final classification consists of clear or unclear utterances in terms of expressiveness. Therefore, the set of binary outputs from each classifier is the input to this meta-learner. The simplest solution is to implement a voting schema although other simple learners such as tree-based or rule-based classifiers can achieve better results.

The results of the improvements considered by adding voice quality parameters and the implementation of a bidirectional attribute selection are shown in Table 8, where the three algorithms (SMO, J48 and NB) and three attribute selection strategies (FW, 3FW-1BW and 4FW-1BW) applied stand out. Compared with the baseline, both improvements imply a relative increase greater than 20% in terms of F_1 measure. Figure 8 shows the evolution of the maximum of F_1 score according to the best subset of attributes in each iteration of the 3FW-1BW strategy.

For the stacking implementation, a simple voting schema was initially tested with a subset of the seven best classifiers obtained from the different combinations of the three algorithms (SMO, J48 and NB) and the three attribute selection strate-

Table 8: Initial maximum F_1 with FW strategy for SMO, J48 and NB, results for the dataset with VQ and finally, with 3FW-1BW and 4FW-1BW strategies

Algorithm	Without VQ (FW)	With VQ (FW)	With VQ (3FW-1BW)	With VQ (4FW-1BW)
SMO	0.49	0.59	0.61	0.61
J48	0.43	0.52	0.56	0.56
NB	0.42	0.48	0.58	0.54

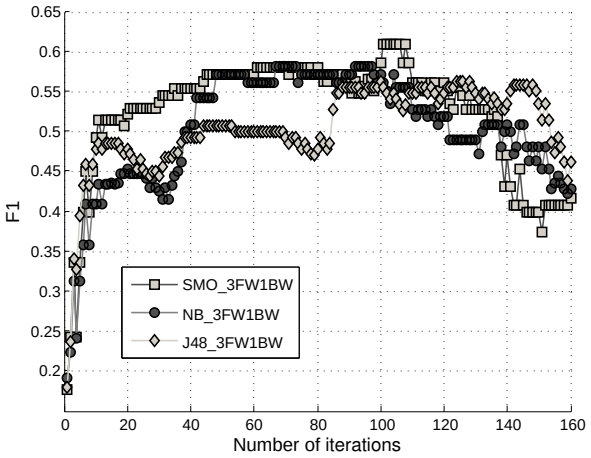


Figure 8: Maximum F_1 values per iteration for 3FW-1BW attribute selection.

gies (FW, 3FW-1BW and 4FW-1BW). Each classifier assigns the UC or CL classes to every utterance and these individual results are then considered in the final decision process by establishing the minimum number of votes required to determine the utterance as unclear. We have observed that the aggressive style suffers greatly when the number of votes is increased. Therefore, this class has been weighted twice as much as the other styles. The highest F_1 measure is 0.7, which is achieved with 4 votes, followed by 3 votes, which yields a closer recall and precision values (Figure 9). The highest single result 0.61 is considerably improved (see Table 8).

Moreover we have trained a rule-based classifier based on PART (Witten and Frank, 2005) that slightly improves this result obtaining $F_1 = 0.73$. The classifiers used by PART are C1=SMO(3FW-1BW), C2=J48(3FW-1BW), C3=J48(4FW-1BW) and C4=NB(3FW-1BW). Algorithm 1 shows the final rules where 0 means correctly classified and 1 misclassified. Correctly / incorrectly classified cases are given in brackets. It is worth pointing out that the aggressive style is processed specifically by the second rule. Finally, both stacking strategies (voting and PART) have been applied to the whole corpus obtaining different subsets of sentences which have to be eliminated. The highest vote increases precision while decreasing it increases the recall.

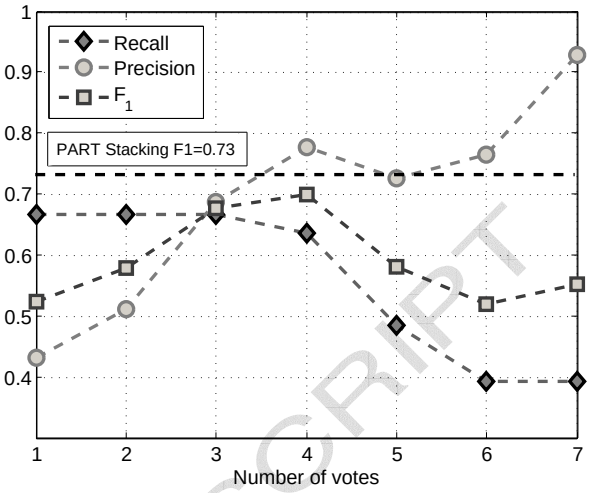


Figure 9: F_1, precision and recall with respect to the minimum number of votes needed to consider that an utterance is unclear.

Algorithm 1 PART decision list for stacking

C1 = 0 and C2 = 0 and C3 = 0 and C4 = 0: CLEAR (408/10)
C1 = 0 and Style = AGR and C2 = 1: UNCLEAR (7/2)
C1 = 0: CLEAR (25)
C3 = 1: UNCLEAR (18/3)
C2 = 0 and C4 = 0: CLEAR (4/1)
C2 = 0: UNCLEAR (2)
: CLEAR

Figure 10 shows the percentage of eliminated utterances detailed per style. Note that, in addition, "Voting-4" and "PART" present similar global results to subjective evaluation as can be observed in Table 3. The best result of F_1 score versus the simple voting schema is due to the increased precision as it selects fewer sentences to be eliminated.

5.3. Assessment of the automatic refinement system

The automatic evaluation process guided by subjective criteria was applied to the whole corpus and extracted a set of utterances classified as unclear from the expressiveness viewpoint. However, this result must be validated in order to guarantee its legitimacy and in order to decide whether to automatically delete unclear utterances or whether to review them once again with subjective evaluators.

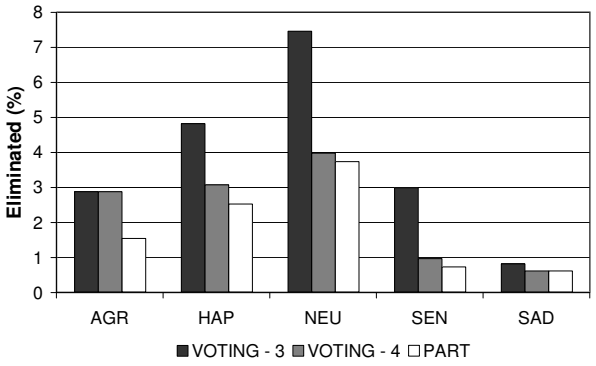


Figure 10: Eliminated utterances per style.

To achieve this goal, a new subjective test has been designed which contemplates the utterances tagged as unclear by the automatic system by taking a set of correct sentences as control points.

This test consists of 75 sentences considered unclear by the stacking version (PART algorithm) of the system which follow the following distribution: aggressive (16 utterances), happy (23), neutral (24), sensual (6) and sad (6). A subset of clear sentences was created by choosing the same number of utterances per style from those previously evaluated by the first subjective test, ensuring that the style they belong to is respectively represented. The total number of sentences in this test is 150, and they are presented randomly to the subjective evaluator. This process follows the same guidelines as those described in the first subjective test (see Section 3). The desired result is to verify that people classify incorrectly, to a greater or lesser extent, those sentences tagged as unclear by the system, whereas they do not fail in those from the other subset. The test was performed by 38 Spanish people and 10 non-Spanish people with heterogeneous profiles.

Results show a greater error rate in the sentences UC tagged than in the CL set. Figure 11 shows the global identification error per style for both UC and CL classes. For the Spanish listeners, the identification error is higher in the UC set than in CL for all the styles except sadness (Figure 11a). For the group of non-Spanish listeners, the results are very similar for all the styles with the exception of the aggressive style, which presents almost the same confusion between UC and CL utterances. The sad style presents a very low identification error ($< 10\%$) for all the cases. This is due to the fact that this style is clearly distinguished because its sound features are very characteristic, as was already observed in the first subjective test (see Figure 1).

This analysis reveals that the system is obtaining the expected results, except for sadness, which would not need a posterior review. Finally, it is interesting to analyze the results by considering each sentence individually. If we apply the same criteria defined in Section 5 to determine if a sentence should be tagged as CL or UC from subjective test results to analyze the evaluations of each individual listener, the F_1 score for the UC class can be computed. The results reveal a correct implementation of the system for sensual, happy and neutral styles and acceptable for aggressive (see Table 9). Sad style is not evaluated because there are no utterances considered to be unclear. Considering these results, the main difference between Spanish and non-Spanish listeners seems to be in the precision, i.e. non-Spanish listeners tend to misclassify more utterances that belong to the CL class than Spanish listeners. It can be concluded that most of the utterances confused by evaluators in this test have been detected by the system (high precision), although some utterances classified as UC by the system are correctly identified by more than 50% of the listeners.

Table 9: Precision, recall and F_1 for the UC class obtained from the comparison between the automatic classification and the results of the final subjective test at utterance level for Spanish and non-Spanish listeners

	Spanish			Non-Spanish		
Style	Precision	Recall	F_1	Precision	Recall	F_1
AGR	1.00	0.25	**0.40**	0.50	0.38	**0.43**
HAP	1.00	0.65	**0.79**	0.80	0.70	**0.74**
NEU	1.00	0.67	**0.80**	0.88	0.58	**0.70**
SEN	1.00	0.67	**0.80**	0.80	0.67	**0.73**
SAD	–	–	–	–	–	–
Total	1.00	0.57	**0.70**	0.76	0.58	**0.65**

6. Discussion

The aim of this paper is to present the method developed to solve the problem which arises when the quality of the utterances produced by speakers does not meet the desired expressiveness. This is a common problem in the development of corpora for expressive speech synthesis. The main idea of the method consists in subjectively evaluating part of the speech corpus and training an automatic classifier to recognize bad utterances in the remainder of the corpus.

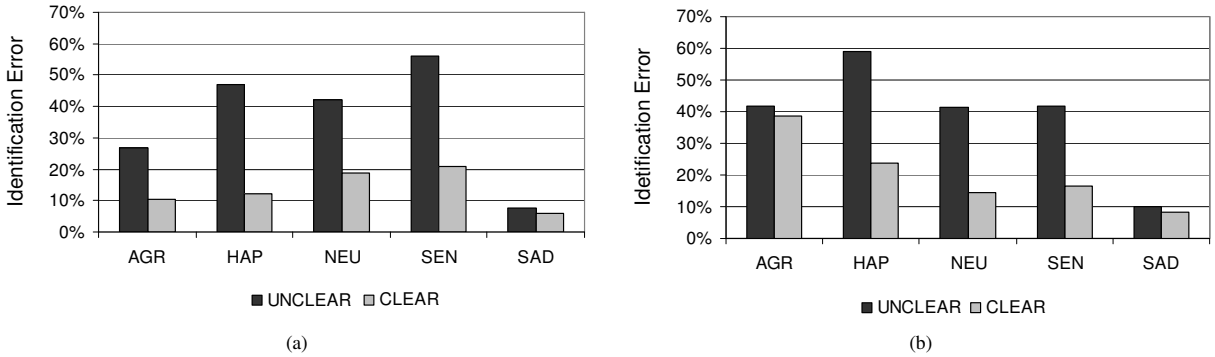


Figure 11: Percentage of error identification per style for the UC and CL classes in the second subjective test, distinguishing between a) Spanish listeners b) Non-Spanish listeners.

An experimental study of this method has been carried out with the Spanish expressive speech corpus, also presented in this paper, recorded from advertising texts in different expressive styles. The aim of this corpus was to achieve different expressive styles with sufficiently varied acoustics but with sufficient coherence intra-style in order to use it in our current research in speech synthesis (prosody and voice quality modeling and synthesis methods). In a broad sense, expressive speech covers, among other things: emotional speech as it is well known, narrative style for story telling (Theune et al., 2006), good/bad news (Pitrelli et al., 2006), soccer announcements (Krstulovic et al., 2007), etc. Therefore, the developed corpus is an additional example of expressive speech data oriented to research instead of a specific application.

The results achieved in the refinement of this corpus are determined by some adjustments of the system. If this method is applied to other data, the main aspects to be considered are the following. Firstly, the heuristic rule defined to reach the consensus among the evaluators should be adapted to the results of the partial subjective test. This adjustment could be manual or included in the automatic process. Secondly, recall and precision have been equally weighted with the F_1 measure. If it is preferable to have a good recall of the unclear class as opposed to achieving good precision, F_2 measure can be used which weights recall twice as much as precision. In this case, the number of utterances classified as unclear would increase and, therefore, it would be necessary to review these utterances in order to decide if they can be deleted from the corpus. If the number of deleted utterances was very high, it would be necessary to repeat their recording with the aim of avoiding a reduction of the intended coverage.

7. Conclusion and future work

There is a high consensus in the speech-synthesis research community regarding the benefits of obtaining emotional speech through acted or stimulated speech regardless of the drawbacks in terms of authenticity. In order to overcome this problem, an assessment of the expressiveness of the recorded utterances is required. At first glance, subjective evaluation appears to be the best method for this purpose.

However, an exhaustive evaluation of the whole database is too costly for large corpora, which are generally required in corpus-based speech synthesis. Nevertheless, to the best of the author's knowledge, a precise and objective evaluation of expressiveness is yet to be developed. Our findings represent a step further on the path to creating a practical and efficient evaluation able to cover this need.

The paper has presented a method to automatically refine a complete expressive speech corpus. First the design, creation and initial objective validation of an expressive corpus have been described. The results of the perceptual tests led to the design of a system that emulates the subjective criteria featured in emotion identification from speech.

The system's training performs an attribute selection that maps the results of a subjective evaluation carried out on a small part of the corpus. The resulting unclear set is validated along with a clear set by means of a second subjective test, which

was performed by Spanish and non-Spanish listeners. According to the results, the system's automatic decision and the subjective perception seem to be closely correlated.

Therefore, it can be concluded that the proposed system is able to erase the unclear utterances in terms of expressiveness (i.e. with emotion/style different to that expected). As a result, the whole initial corpus can be automatically refined using this system. The results are encouraging since the final corpus content shows a higher match with perceptual classification. However new improvements could be introduced in different parts of the proposed method: *i)* the previous subjective test and the rules to determine the expressiveness level of the evaluated utterances; and *ii)* the tuning of the automatic system (the acoustic parameterization, the attribute selection method and the combination of new classification algorithms). The next stage of our investigations will study the convenience of using the proposed method for expressive speech synthesis. The refined corpus could be used in the acoustic modelling of expressive speaking styles by predicting prosodic parameters from text and in the unit database of a corpus-based synthesizer. It would be interesting to compare the results with the ones achieved from the original corpus.

8. Acknowledgments

This work has been partially supported by the European Commission, project SALERO (FP6 IST-4-027122-IP) and the Spanish Government, project SAVE (TEC2006-08043/TCM).

References

Alías, F., Monzo, C., Socoró, J. C., 2006. A pitch marks filtering algorithm based on restricted dynamic programming. In: InterSpeech2006 - International Conference on Spoken Language Processing (ICSLP). Pittsburgh, PA, USA, pp. 1698–1701.

Bozkurt, B., Ozturk, O., Dutoit, T., 2003. Text design for TTS speech corpus building using a modified greedy selection. In: The 8th European Conference on Speech Communication and Technology (EUROSPEECH). Geneva, Switzerland, pp. 277–280.

Campbell, N. W., Isard, S., 1991. Segment durations in a syllable frame. Journal of Phonetics 19, 37–47.

Campbell, N. W., 2000. Databases of emotional speech. In: Proceedings of the ISCA Workshop on Speech and Emotion. Newcastle, Northern Ireland, UK, pp. 34–38.

Campbell, N. W., 2002. Recording techniques for capturing natural everyday speech. In: Proceedings of the 3rd International Conference on Language Resources and Evaluation (LREC'02). Las Palmas de Gran Canaria, Spain.

Campbell, N. W., 2005. Developments in corpus-based speech synthesis: Approaching natural conversational speech. IEICE - Transactions on Information and Systems E88-D (3), 376–383.

Cowie, R., Douglas-Cowie, E., Cox, C., 2005. Beyond emotion archetypes: databases for emotion modelling using neural networks. Neural Networks 18, 371–388.

Cowie, R., Douglas-Cowie, E., Tsapatsoulis, N., Votsis, G., Kollias, S., Fellenz, W., Taylor, J. G., 2001. Emotion recognition in human computer interaction. IEEE Signal Processing 18 (1), 33–80.

Devillers, L., Vidrascu, L., Lamel, L., 2005. Challenges in real-life emotion annotation and machine learning based detection. Neural Networks 18, 407–422.

Douglas-Cowie, E., Campbell, N., Cowie, R., Roach, P., 2003. Emotional speech: towards a new generation of databases. Speech Communication 40, 33–60.

Drioli, C., Tisato, G., Cosi, P., Tesser, F., 2003. Emotions and voice quality: experiments with sinusoidal modeling. In: Voice Quality: Functions, Analysis and Synthesis (VOQUAL'03), ISCA Tutorial and Research Workshop. Geneva, Switzerland, pp. 127–132.

Duda, R. O., Hart, P. E., Stork, D. G., 2001. Pattern Classification, 2nd Edition. Wiley & Sons, Inc., New York.

François, H., Boëffard, O., 2002. The greedy algorithm and its application to the construction of a continuous speech database. In: Proceedings of the 3rd International Conference on Language Resources and Evaluation (LREC'02). Las Palmas de Gran Canaria, Spain.

Goldberg, D. E., 1989. Genetic Algorithms in Search Optimization and Machine Learning. Addison Wesley, Reading, MA.

Hozjan, V., Kacic, Z., Moreno, A., Bonafonte, A., Nogueiras, A., 2002. Interface databases: Design and collection of a multilingual emotional speech database. In: Proceedings of the 3rd International Conference on Language Resources and Evaluation (LREC'02). Las Palmas de Gran Canaria, Spain.

Iriondo, I., Planet, S., Alías, F., Socoró, J. C., Martínez, E., 2007a. Validation of an expressive speech corpus by mapping automatic classification to subjective evaluation. In: Computational and Ambient Intelligence. 9th International Work-Conference on Artificial Neural Networks, IWANN 2007, San Sebastián, Spain, June 20-22, 2007. Proceedings. Vol. 4507 of Lecture Notes in Computer Science. Springer, Heidelberg, pp. 646–653.

Iriondo, I., Planet, S., Socoró, J. C., Alías, F., 2007b. Objective and subjective evaluation of an expressive speech corpus. In: Advances in Nonlinear Speech Processing, International Conference on Non-Linear Speech Processing, NOLISP 2007, Paris, France, May 22-25, 2007. Vol. 4885 of Lecture Notes in Computer Science. Springer, Heidelberg, pp. 86–94.

Iriondo, I., Socoró, J., Alías, F., 2007c. Prosody modelling of Spanish for expressive speech synthesis. In: Proceedings of the 32nd IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP). Vol. 4. Honolulu, HI, USA, pp. 821–824.

Krstulovic, S., Hunecke, A., Schröder, M., 2007. An HMM-based speech synthesis system applied to german and its adaptation to a limited set of expressive football announcements. In: Proceedings of Interspeech-2007. Antwerp, Belgium, pp. 1897–1900.

Montero, J. M., Gutiérrez Arriola, J., Palazuelos, S., Enríquez, E., Aguilera, S., Pardo, J. M., 1998. Emotional

speech synthesis: From speech database to TTS. In: The 5th International Conference on Spoken Language Processing (ICSLP). Sydney, Australia, pp. 923–926.

Montoya, N., 1998. El papel de la voz en la publicidad audiovisual dirigida a los niños. Zer. Revista de estudios de comunicación (4), 161–177, (in Spanish).

Monzo, C., Socoró, J. C., Iriondo, I., Alías, F., 2007. Discriminating expressive speech styles by voice quality parameterization. In: Proceedings of the 16th International Congress of Phonetic Sciences (ICPhS'2007). Saarbrüken, Germany, pp. 2081–2084.

Morrison, D., Wang, R., De Silva, L. C., 2007. Ensemble methods for spoken emotion recognition in call-centres. Speech Communication 49 (2), 98–112.

Murray, I. R., Arnott, J. L., 1993. Toward the simulation of emotion in synthetic speech: A review of the literature of human vocal emotion. Journal of the Acoustic Society of America 93 (2), 1097–1108.

Navas, E., Hernáez, I., Luengo, I., 2006. An objective and subjective study of the role of semantics and prosodic features in building corpora for emotional TTS. IEEE Transactions on Audio, Speech and Language Processing 14 (4), 1117–1127.

Nogueiras, A., Moreno, A., Bonafonte, A., no, J. B. M., 2001. Speech emotion recognition using Hidden Markov Models. In: Proceedings of The 7th European Conference on Speech Communication and Technology (EUROSPEECH). Aalborg, Denmark, pp. 2679–2682.

Oudeyer, P. Y., 2003. The production and recognition of emotions in speech: features and algorithms. Int. Journal of Human Computer Interaction 59 (1-2), 157–183, special issue on Affective Computing.

Pérez, E. H., 2003. Frecuencia de fonemas. eRTH Revista electrónica de Tecnología del Habla (1), (in Spanish).
URL `http://www.rthabla.es`

Pitrelli, J. F., Bakis, R., Eide, E. M., Fernandez, R., Hamza, W., Picheny, M. A., 2006. The IBM expressive text-to-speech synthesis system for American English. IEEE Transactions on Audio, Speech and Language Processing 14 (4), 1099–1108.

Planet, S., Iriondo, I., Martínez, E., Montero, J. A., 2008. TRUE: an online testing platform for multimedia evaluation. In: Proceedings of the Second International Workshop on EMOTION: Corpora for Research on Emotion and Affect at LREC'08. Marrakech, Morocco.

Scherer, K. R., 2003. Vocal communication of emotion: A review of research paradigms. Speech Communication 40, 227–256.

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

Schweitzer, A., Möbius, B., 2003. On the structure of internal prosodic models. In: Proceedings of the 15th International Congress of Phonetic Sciences (ICPhS'2003). Barcelona, Spain, pp. 1301–1304.

Shami, M., Verhelst, W., 2007. An evaluation of the robustness of existing supervised machine learning approaches to the classification of emotions in speech. Speech Communication 49 (3), 201–212.

Theune, M., Meijs, K., Heylen, D., Ordelman, R., 2006. Generating expressive speech for storytelling applications. IEEE Transactions on Audio, Speech and Language Processing 14 (4), 1137–1144.

Ververidis, D., Kotropoulos, C., 2006. Emotional speech recognition: Resources, features, and methods. Speech Communication 48 (9), 1162–1181.

Wells, J., 1993. Sampa: Computer readable phonetic alphabet.
URL `http://www.phon.ucl.ac.uk/home/sampa/`

Witten, I. H., Frank, E., 2005. Data Mining: Practical Machine Learning Tools and Techniques, 2nd Edition. Morgan Kaufmann, San Francisco.