An Objective and Subjective Study of the Role of Semantics and Prosodic Features in Building Corpora for Emotional TTS

Eva Navas, Inmaculada Hernáez, and Iker Luengo

Abstract—**Building a text corpus suitable to be used in corpus-based speech synthesis is a time-consuming process that usually requires some human intervention to select the desired phonetic content and the necessary variety of prosodic contexts. If an emotional text-to-speech (TTS) system is desired, the complexity of the corpus generation process increases. This paper presents a study aiming to validate or reject the use of a semantically neutral text corpus for the recording of both neutral and emotional (acted) speech. The use of this kind of texts would eliminate the need to include semantically emotional texts into the corpus. The study has been performed for Basque language. It has been made by performing subjective and objective comparisons between the prosodic characteristics of recorded emotional speech using both semantically neutral and emotional texts. At the same time, the performed experiments allow for an evaluation of the capability of prosody to carry emotional information in Basque language. Prosody manipulation is the most common processing tool used in concatenative TTS. Experiments of automatic recognition of the emotions considered in this paper (the "Big Six emotions") show that prosody is an important emotional indicator, but cannot be the only manipulated parameter in an emotional TTS system—at least not for all the emotions. Resynthesis experiments transferring prosody from emotional to neutral speech have also been performed. They corroborate the results and support the use of a neutral-semantic-content text in databases for emotional speech synthesis.**

Index Terms—**Evaluation of expressivity, prosody analysis, speech corpus design.**

I. Introduction

NOWADAYS, corpus-based text to speech (TTS) systems are able to produce high-quality speech, with great naturalness. This high quality is mainly due to the large speech database used in building the synthetic signal. Most of those systems produce speech in a neutral style that limits their fields of application and frequently prevents the users from utilizing them. Providing text-to-speech systems with the ability to express emotions is essential for natural interactive interfaces. Various systems have been already implemented in this way, from avatars and modern interactive entertainment toys to automated customer service systems[1] [1].

The text material that is used to build the database is very important when designing corpora for corpus-based emotional speech synthesis. It should be phonetically and prosodically balanced and have a good coverage of the language. When considering emotions, if the semantic content of the texts is related with them, the difficulty of designing the text corpus gets multiplied by the number of emotions to be considered. The collection of emotion dependent texts requires an additional effort, because they are not easy to find. This is why some emotional databases have been designed using only neutral texts to record emotional speech.

For the Basque language, no database suitable for emotional TTS exists. Basque is the only surviving pre-Indo-European language in Western Europe. It is mainly spoken around the border between France and Spain. It is not a Latin language, thus it is totally different from Spanish or French, the other two languages also spoken in the Basque-speaking areas of Europe (see [2] and [3] for a review of the Basque linguistics and origin). Although most of its phonetics is similar to the Spanish phonetics in the Spanish speaking region, Spanish and Basque differ completely concerning lexica, grammar, syntax, and morphology.

Before proceeding to the creation of a suitable emotional database in Basque, the study here presented was performed. It will help us predict whether a text corpus with neutral content will be sufficient to create a good emotional database to be used in corpus-based emotional synthesis. The goal of our work is to study the relationship between the semantic content of the text and the ability of naturally expressing and correctly identifying emotions in speech. This study has been performed both subjectively, checking the ability of human evaluators to identify emotions, and objectively, analyzing a set of acoustic prosodic parameters and building an automatic classifier of emotions. In the objective evaluation, only the statistics of prosodic parameters have been considered. Although it is known that other aspects of speech such as spectral features and glottal source characteristics are important cues in emotion identification, considering all aspects of speech is difficult and requires very exhaustive and extensive work. On the other hand, prosody manipulation is a very important speech processing tool that deserves special attention, e.g., it is the most common operation performed on a diphone-based concatenative TTS system. In such a system, only prosody modeling is used to indicate emotion. It is thus interesting to know to what degree the prosody of emotional speech can be used as the only indicator of the pretended emotion.

To perform our work, a medium-sized database composed of two text subcorpora, one coordinated with the emotion and

Manuscript received May 30, 2005; revised March 2, 2006. This work was supported in part by the Spanish Ministry of Science and Technology under Grant (TIC2003-08382-C05-03). The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Wael Hamza.

The authors are with the Department of Electronics and Telecommunications, University of the Basque Country, Bilbao E48013, Spain (e-mail: eva.navas@ehu.es; inma.hernaez@ehu.es; ikerl@bips.bi.ehu.es).

Digital Object Identifier 10.1109/TASL.2006.876121

[1] [Online] Available: http://www.i-cybie.com.

another one with neutral semantic content has been designed and recorded by a single speaker using emotional speech. For example, a sentence extracted from the corpus coordinated with joy is “Congratulations son, you got the job!” while a sentence from the neutral or not coordinated subcorpus is “Both of them were smartly dressed.” The database design is described in Section II. The database serves as a base to develop our work. Due to its size, it is not useful for corpus-based unit-selection speech synthesis where a more extensive database would be needed. Section III describes the subjective evaluation procedure as well as the evaluation results. The evaluation was performed by two different groups of listeners (Basque speakers and non-Basque speakers). Although subjective evaluation already offers quite definitive conclusions in the direction of the validation of the use of neutral texts, objective analysis of the statistics of the prosodic parameters was performed in two ways. The first objective analysis using ANOVA is presented in Section IV. The second objective analysis using an automatic classifier of emotions is detailed in Section V with a description of the differences in the behavior of the analyzed parameters in both subcorpora. A subjective test using prosody transferring and resynthesis is included in Section VI. Summarizing conclusions are presented in Section VII.

II. Description of the Corpus

A. Selection of the Characteristics of the Corpus

Different types of corpora have been used for the study of emotions in speech. Some databases use corpora of spontaneous speech. Among these databases are the Belfast database [4] for English, consisting of clips from television programs and the JST database [5] for Japanese, English, and Chinese, with natural speech recorded in natural situations. Other databases use elicited speech. This is the case of the database recorded for Swedish in the VERIVOX project [6] and the Hebrew database [7]. Finally, there are databases of acted speech, like the one recorded by Lay Nwe and colleagues for Burmese and Mandarin with amateur speakers for emotion recognition [8], the database used to develop the CHATAKO-AID system in Japanese [9], the RUSLANA database [10] for Russian, the MediaTeam Finnish Speech Corpus [11], and the SES database [12] for Spanish.

All the methods used to obtain emotional speech corpora have advantages and disadvantages [13]. A good review on existing emotional databases can be found in [14].

In our paper, acted speech was selected because it is easier to control and allows an easy comparison among styles. This technique has been accused of recording unnatural examples of emotions not representative of normal speech [15] because simulation leads to prototypical emotions that are more intense than their normal expression. However, as the intended emotion can be recognized, they should be considered satisfactory for speech synthesis studies. Besides, with this type of corpus it is possible to control the content of the recording and, therefore, phonetic variability can be maximized which is an interesting feature when performing studies of sound duration or using the database for corpus-based TTS.

Another important aspect to be considered is the set of emotions to include. Work in this area has usually produced a set of limited pure emotions. Nonetheless, there has been no consensus about the number and identity of these pure emotions. Depending on the approach selected, different sets of important emotions have been considered. Recently the term “Big Six” has been used to group anger, disgust, fear, joy, sadness, and surprise [16], [17]. This set has been used in different studies related with speech, both for emotion recognition [8] and for emotion generation [18]. This is the set selected in our work. An additional neutral style is used as reference.

B. Text Corpus Design

In the design of a corpus for emotional speech, considerations about the semantic content of the texts must be made. There are different theories about the suitability of the database’s texts to be semantically related to the expressed emotion.

On the one hand, the use of texts that are semantically related to the emotion makes it easier for the speaker to express that emotion naturally. However, it makes it difficult to compare the characteristics of different emotions and to phonetically balance the database. The collection of suitable texts to be recorded is also an arduous task. An example of an emotional database containing acted speech with texts related to the emotion is the one used to synthesize emotional speech with the ATR-CHART system [19].

On the other hand, the use of neutral texts (not related to the emotion) eases the comparison among emotions and the phonetic balance of the database, but the work of the speaker to express the emotions naturally is much more difficult. Likewise, the text selection for the corpus becomes easy. Examples of databases that use neutral texts to record emotional speech are the Danish Emotional database [20], the Interface database [21], and the Berlin corpus [22].

Again, each approach has its advantages and disadvantages. However, this is a very important premise to consider when designing corpora for corpus-based synthesis, as the amount of text material differs considerably in both cases and the collection of emotion dependent texts requires an additional effort. In our opinion, it is worth investigating the use of emotion dependent texts to be sure that the higher effort truly results in the expected significant benefits.

To perform this analysis, a small- to medium-sized database was designed, using both approaches.

- One subcorpus consists of emotion independent texts, which are common for all emotions, as well as for the neutral style. This common group of texts was phonetically balanced in order to achieve a phoneme distribution as close as possible to the one that occurs in natural oral Basque language, while at the same time assuring the presence of less frequent units. These texts have neutral semantic content and we refer to them as “Common Subcorpus” from now on.
- A second subcorpus includes texts semantically related to each emotion. Therefore, the texts are different for each of the emotions considered in the database. Neutral style was not considered in this part of the corpus. We refer to these texts as “Specific Subcorpus.”

TABLE I
COMPOSITION OF THE CORPUS

Type of item	Common Subcorpus	Specific Subcorpus
Isolated digits	20	-
Isolated words	20	20
Short sentences	15	15
Medium sentences	30	30
Long sentences	10	10
Total number of items per emotion	95	75
Total number of items	665	450

TABLE II
SIZE OF THE DATABASE

	A	D	F	J	N	Sd	Sp
Length for CS	7:10	6:47	8:15	6:02	7:04	6:27	8:31
Length for SS	5:42	6:09	5:55	4:50	-	5:25	6:39
Word	923	1031	949	951	542	996	953
Syllable	2603	2985	2616	2716	1594	2764	2702
Sound	5309	6110	5395	5595	3280	5712	5499

A = anger, D = disgust, F = fear; J = joy, N = neutral, Sd = sadness, Sp = surprise.
CS = Common Subcorpus, SS = Specific Subcorpus.

It is known that emotion can be reliably identified in very short utterances [20]. Hence, isolated words seem to be suitable for this type of database. However, it is interesting to include also longer sentences to be able to study the location, number, and duration of pauses and the rhythm of the speech. Therefore, both subcorpora were designed to include isolated words and sentences of different grade of complexity and syntactical structure. Interrogative and declarative sentences were used.

Table I shows the number and type of items recorded using neutral common texts (seven styles: six emotions+neutral style) and texts coordinated with emotions (six styles). Short sentences are simple sentences of about five words without internal pauses, medium sentences are sentences that include only one orthographically indicated internal pause (about nine words long) and long sentences include more than one internal pause (about 14 words long). Table II shows the final size of the database measured in time, words, syllables, and sounds recorded per emotion.

C. Database Recording

A professional dubbing actress was recruited for the recordings as she has the ability of expressing the required emotion with enough naturalness.

The recording was made at a professional recording studio during two days. On the first day, emotion-dependent texts (i.e., "Specific Subcorpus") were recorded. On the second day, common sentences (i.e., "Common Subcorpus") were taped. Within a recording session, every emotion was recorded without interruption to avoid the speaker losing concentration. The speaker was allowed to rest between the recordings of texts corresponding to different emotions.

The recording was made using a laryngograph to capture also the glottal pulse signal. Speech and glottal pulse signals were sampled at 32 kHz and quantized using 16 bits per sample. The recorded database is 1 h and 25 min long. Fifty minutes come from the common texts and 35 min from the texts semantically related with emotion.

III. SUBJECTIVE EVALUATION OF THE DATABASE

A subjective evaluation is a valuable tool for the validation of an acted emotional speech database. Besides proving the ability of the speaker to accurately simulate emotions, for our paper, it was also important to establish the differences in the performance for both subcorpora. An evaluation test was designed where participants had to guess the emotional content of sentences from both subcorpora. In a first experiment, all the participants were Basque speakers. Although participants were told to consider only the perceived emotion, regardless of the sentences meaning, it is difficult for the listener not to consider the meaning of the sentences. The results of this first test show significantly better recognition scores for the Specific Subcorpus. We then proceed to a second evaluation experiment where the participants had no knowledge of the Basque language. Due to their ignorance of the language, semantics is not influencing the decision of this second group of evaluators in either subcorpus. This section describes the details of the evaluation test, presents the obtained results, and draws some conclusions.

A. Test Design

A forced choice test was designed where users had to select one of the proposed emotions, including neutral style. Thus, it was a test for discriminating emotions rather than identifying them. The six emotions contained in the database and the neutral style were proposed to the listeners. To check the dependency on the semantic content of the signals, sentences from both the Common Subcorpus and the Specific Subcorpus were selected. For each style, ten sentences with common texts and ten sentences with specific texts were used, generating a total of $(6+7)^*10 = 130$ stimuli. Sentences with different lengths and syntactic complexity were randomly selected for the test. The shortest one had three words (nine syllables) and the longest one 18 words (52 syllables).

After the test was completed, the evaluator was asked to list the emotions that were difficult to identify.

B. Evaluation Process

This test was performed by two different groups of subjects: by Basque speakers and by people who do not understand Basque. Non-Basque speakers were recruited outside Basque speaking area; therefore, they have no familiarity with this language. This way, the semantic content of the Specific Subcorpus could not influence the decision of this second group of evaluators. It might be argued that the ignorance of the language implies at least a partial inability to evaluate the emotion on the speech. Regarding this subject, we should remember that our interest is focused on comparing the performance for Common and Specific subcorpora, and the inability to recognize emotions would be equally present in both of them.

The Basque speaking subjects taking part in the experiment were selected among the students and staff of the Electronics and Telecommunication Department, University of the Basque

TABLE III
CONFUSION MATRIX OF THE SUBJECTIVE TEST FOR BASQUE-SPEAKING SUBJECTS

	A	D	F	J	N	Sd	Sp
A	85.7%	4.0%	0.3%	0.0%	7.3%	0.0%	3.3%
D	9.0%	51.3%	1.0%	0.0%	3.3%	0.3%	0.0%
F	0.0%	7.0%	80.3%	0.0%	0.7%	5.7%	0.7%
J	0.3%	0.3%	0.0%	82%	1.3%	0.0%	4.7%
N	3.3%	23%	0.0%	14%	78.7%	17.7%	7.0%
Sd	0.0%	12.7%	18%	0.0%	0.0%	73.7%	0.0%
Sp	1.7%	1.7%	0.3%	4.0%	8.7%	2.7%	84.3%

Confusion matrix of the subjective test where columns contain true values and rows values selected by listeners.

A = anger, D = disgust, F = fear; J = joy, N = neutral, Sd = sadness, Sp = surprise.

TABLE IV
CONFUSION MATRIX OF THE SUBJECTIVE TEST FOR SUBJECTS THAT DO NOT UNDERSTAND BASQUE

	A	D	F	J	N	Sd	Sp
A	69.0%	6.9%	2.1%	10.0%	11.0%	3.1%	6.0%
D	8.3%	22.9%	3.8%	1.0%	1.9%	1.2%	2.1%
F	3.1%	8.6%	46.2%	1.4%	1.4%	10.7%	4.0%
J	3.8%	2.6%	1.7%	48.1%	4.8%	3.8%	7.6%
N	6.0%	30.7%	2.9%	19.8%	65.7%	29.5%	18.1%
Sd	2.6%	19.5%	35.7%	2.4%	2.9%	47.4%	2.1%
Sp	7.1%	8.8%	7.6%	17.4%	12.4%	4.3%	60.0%

Confusion matrix of the subjective test where columns contain true values and rows values selected by listeners.

A = anger, D = disgust, F = fear; J = joy, N = neutral, Sd = sadness, Sp = surprise.

Country. A total of 15 participants (11 males and four females aged 20 to 36 years) took part in the experiments. All of them were native Basque speakers or fluent in standard Basque. None of them reported speech or hearing problems. Some of them were used to TTS systems, but none of them has a special phonetic training.

The tests were performed in the quasi-silent environment of a research laboratory. Stimuli were presented to listeners over high-quality headphones and reproduced with a standard Sound Blaster soundcard.

The subjects that did not understand Basque were recruited via Internet, primarily from the staff of different Spanish universities. Twenty-one participants (19 males and two females aged 22–41 years) took part in the test. All of them speak Spanish fluently, although not all of them are native speakers. This test is available at http://bips.bi.ehu.es/ahoweb/evaluacion.

The stimuli were presented to subjects by means of electronic forms that grouped ten stimuli for evaluation. Listeners took no training session and got no feedback about their performance. Participants could hear the signals by clicking the adequate buttons and had to select the emotion they identified in the acoustic signal from a list of seven choices. Listeners could hear each stimulus as many times as desired. They had to label all the signals presented in a form before advancing to the next form. Once a form had been completed, they could not return and modify it. The order of the stimuli presented was randomized in all the tests.

The complete test consisted of 13 forms and the duration of the test was about 20 min.

C. Results

Results of the subjective test show that all the emotions are identified above chance level (15%) in both subcorpora and by both groups of evaluators. Table III shows the confusion matrix corresponding to the Basque speakers' group: the values in the diagonal belong to emotions correctly identified and are in the range from 51% to 86%. Table IV shows the confusion matrix corresponding to the subjects that do not understand Basque. In this case, correct identifications vary from 22.9% to 69%.

In both cases, anger gets the best result with 85.7% and 69% of correct identifications respectively, although some Basque speaking listeners commented that the expression of anger in the database was exaggerated. The emotion that has been most poorly identified in both cases is disgust, being mainly mixed with neutral style and sadness. This emotion has also been the most difficult to identify for other languages [19], [23], [24].

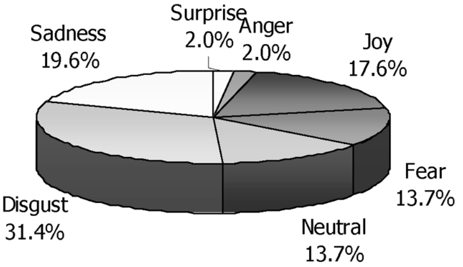


Fig. 1. Perceived difficulty for identifying each emotion.

The main confusions among emotions are similar for both groups of evaluators: fear is identified as sadness in 18% and 35.7% of the cases and sadness as neutral in 17% and 29.5% of the cases, respectively. The subjective impressions of the evaluators agree with these results: when they were asked about the emotions more difficult to discriminate, they chose mostly disgust and sadness, followed by neutral style. Fig. 1 shows the total distribution of answers to the question about the most difficult emotions to identify: disgust and sadness are perceived as the most difficult ones with 31.4% and 19.6% of selections and surprise and anger as the easiest ones with only 2% of selections.

When comparing the results obtained with sentences from the Specific and Common Subcorpora for Basque speakers, most of the emotions achieve a better identification rate when the text is coordinated with emotion. Especially significant is the increase of recognition rate in case of disgust. These results are shown in Fig. 2. In the case of evaluators that did not understand Basque, these differences are not present, as can be seen in Fig. 3. Three emotions achieve a better identification rate for the Common Subcorpus, and the other three are better identified in the Specific Subcorpus. Basque speakers identify the emotions better, even in the case of the Common Subcorpus where the semantic content does not help the listener classifying the emotion. Nevertheless, non-Basque speakers are able to identify all the emotions above chance level. Therefore, an important part of the vocal expression of emotions is not language specific, or at least is common for people with similar cultural

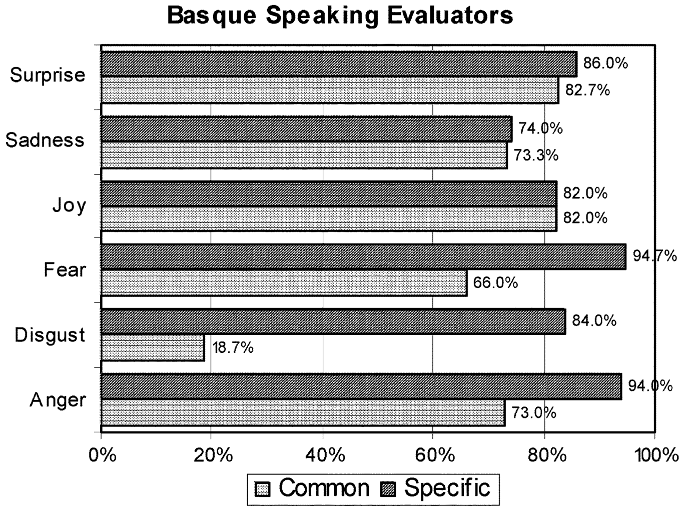


Fig. 2. Results of the subjective evaluation process for Basque speakers.

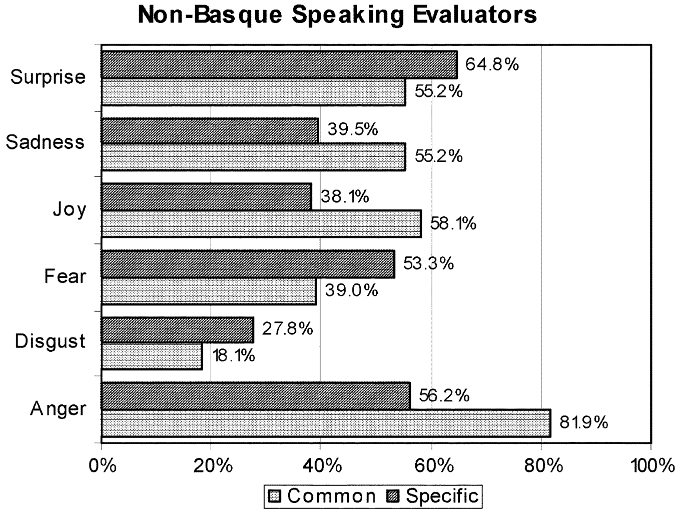


Fig. 3. Results of the subjective evaluation process for non-Basque speakers.

TABLE V
GLOBAL RECOGNITION RATE OF THE SUBJECTIVE TEST

	Common Subcorpus	Specific Subcorpus	Total
Basque speakers	65.9%	85.8%	76.2%
Non Basque speakers	51.3%	46.6%	48.9%
Total	58.6%	66.2%	62.6%

contexts. The complete results of the subjective test are shown in Table V: recognition rate of non Basque speakers is higher for the Common Subcorpus than for the Specific Subcorpus. Total rates are higher in the Specific Subcorpus, due to the good results achieved by Basque speakers who selected the emotion according to the content.

As evaluators had no training session to get used to the signals of the database, results achieved in the first three forms have been compared with the ones from the last three forms. No significant difference has been found, leading to the conclusion that the performance of the evaluators did not increase as they completed the test.

D. Conclusions

Considering the results of the subjective analysis, we can state that the Common Subcorpus plays a similar role as the Specific Subcorpus regarding the ability to express the six emotions considered in this paper. The better recognition rates achieved by the Basque Speakers group of evaluators on the Specific Subcorpus was expected, as it is difficult not to use the semantic information available even if the listener has been instructed not to use it. This result indicates that semantic content is an important cue to identify emotion.

Although disgust is an emotion that is proven to be difficult to recognize also in other works, it could be the actress who is responsible for the poor differentiation of this emotion. A different—more skillful—actor or actress could lead to different results. Multispeaker experiments should be done to eliminate such possibilities. However, for our purposes, the important result is that the scores are similar in both subcorpora, thus validating the use of neutral-text corpora to record emotional speech.

Concerning the test procedure, some evaluators suggested that an option of “Not identified” should have been added to the seven item list for use in cases in which the emotion is not clear. This possibility was considered and discarded in the test design phase to force the listener to make a valid decision. Informal experiments performed previously demonstrated the tendency of the listener to choose the “Not identified” option to speed the completion of the test when faced with a difficult decision. For this reason, together with the desire to avoid reduction of available data to process, we had decided to force the evaluator to take a decision, even if wrong or insecure.

IV. ACOUSTIC ANALYSIS OF THE DATABASE

Prosodic acoustic features such as F0, energy, and duration of sounds are clearly related with vocal emotion. In this paper, long-term averaged parameters related with fundamental frequency and power have been automatically measured and analyzed. This section presents the results of the performed analysis of variance (ANOVA) with the aim of studying the differences in the distribution of these parameters in the Common and the Specific Subcorpora. Sections IV-A–IV-C describe the parameters used. Section IV-D presents a discussion of the differences found and Section IV-E analyzes the capability of each parameter to discriminate among pairs of emotion in each subcorpus.

A. Extraction of the Pitch-Related Features

The values of the pitch curve were obtained from data measured by the laryngograph which provides the following three different synchronized signals as shown in Fig. 4:

— speech signal (Sp);
— glottal pulse signal (Lx): (this signal is captured by the electrodes situated around the neck of the speaker; the local minima indicate the glottal closure);
— quasirectangular signal (Tx): (this signal is created by the laryngograph processor using the information of Sp and Lx signals; it also serves to indicate the closure of the vocal cords).

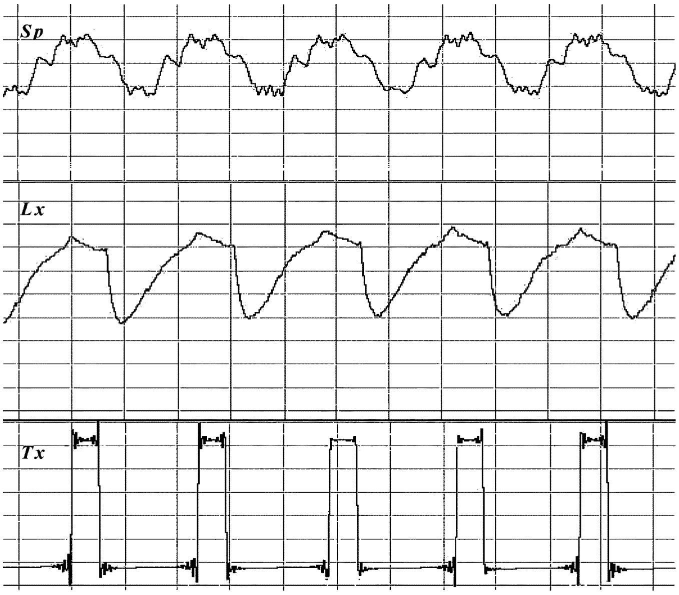


Fig. 4. Signals provided by the laryngograph.

A highly accurate intonation curve was estimated for each of the utterances using the signal recorded by the laryngograph. Pitch values were computed as the inverse of the time elapsed between two consecutive glottal closures.

The voiced/unvoiced (V/UV) information was extracted by detecting the segments where there was glottal closure information and the ones where the glottis was kept open.

The pitch-related parameters analyzed have been as follows.

- Max. F0: Maximum value of pitch curve measured in hertz.
- Mean F0: Mean value of pitch curve measured in hertz.
- F0 Std. Dev.: Standard deviation of pitch curve measured in hertz.
- F0 Range: Difference between maximum and minimum value of pitch curve measured in hertz.
- Max. DF0: Maximum value of first derivative of pitch curve measured in hertz/sample.
- Mean DF0: Mean value of first derivative of pitch curve measured in hertz/sample.
- DF0 Std. Dev.: Standard deviation of first derivative of pitch curve measured in hertz/sample.
- DF0 Range: Difference between maximum and minimum value of first derivative of pitch curve measured in hertz/sample.

B. Extraction of the Power-Related Features

Recordings were analyzed using 25-ms windows and the mean power was calculated every 10 ms for each windowed frame. This gives a power sample every 10 ms. The resulting curves were normalized to the mean value of the neutral style.

Voice activity estimation is necessary in order to reject those frames in which there is no vocal information. This way, noise level during speech silences will not corrupt the calculated features. A voice activity detector (VAD) was implemented based on the computation of the long-term spectral deviation (LTSD) between vocal and noisy frames. The implemented system is based on the one presented in [25] where an adaptive decision threshold is used in order to get the best performance for each noise level.

The power-related parameters analyzed have been as follows.

- Max. POW: Maximum value of power curve measured in decibels.
- Mean POW: Mean value of power curve measured in decibels.
- POW Std. Dev.: Standard deviation of power curve measured in decibels.
- POW Range: Difference between maximum and minimum value of power curve measured in decibels.
- Max. DPOW: Maximum value of first derivative of power curve measured in decibels/sample.
- Mean DPOW: Mean value of first derivative of power curve measured in decibels/sample.
- DPOW Std. Dev.: Standard deviation of first derivative of power curve measured in decibels/sample.
- DPOW Range: Difference between maximum and minimum value of first derivative of power curve measured in decibels/sample.

C. Jitter and Shimmer Estimation

Jitter and shimmer are related to the microvariations of the pitch and power curves, respectively. They can be estimated as the slope change rate of these curves. In this paper, jitter and shimmer were computed as the number of zero crossings of the derivative curves. The result was normalized to the number of frames used for this computation in order to take into account the length of the utterance.

D. Comparison of the Distributions of Parameters Between Common and Specific Subcorpora

To know whether the speaker had expressed the emotions the same way when reading texts related with emotion and texts with neutral content, an ANOVA test has been applied to each parameter with a confidence interval of 99%.

Results of this analysis for the pitch related features are shown in Table VI, where a cell having the value “NO” indicates that differences between the distributions of this parameter in both subcorpora are not significant for the considered emotion. In other words, the speaker has used this parameter the same way in both subcorpora, according to the ANOVA test. Consequently, the parameter that has been most consistently applied by the speaker has been the mean slope of the pitch curve (Mean DF0), considering that the differences are not significant for all the emotions. Maximum slope of F0 (Max. DF0) has also been used the same way in both subcorpora. Besides that, joy is the emotion that has been expressed with the highest similarity in both subcorpora because none of the parameters studied have significant differences between both subcorpora for this emotion. Sadness and anger have the biggest differences in the Common and Specific Subcorpora, as most of the parameters have different distributions in both corpora.

TABLE VI
ANOVA RESULTS FOR PITCH-RELATED PARAMETERS

Feature	A	D	F	J	Sd	Sp
Max. F0		NO	NO	NO		NO
Mean F0		NO		NO		
F0 Std. Dev.		NO		NO		
F0 Range		NO	NO	NO		
Max. DF0	NO		NO	NO	NO	NO
Mean DF0	NO	NO	NO	NO	NO	NO
DF0 Std. Dev.	NO		NO	NO		
DF0 Range	NO		NO	NO		NO
Jitter			NO	NO	NO	NO

A = anger, D = disgust, F = fear; J = joy, N = neutral, Sd = sadness, Sp = surprise.

TABLE VII
ANOVA RESULTS FOR POWER RELATED PARAMETERS

Feature	A	D	F	J	Sd	Sp
Max. POW						
Mean POW						
POW Std. Dev.						
POW Range						
Max. DPOW						
Mean DPOW	NO	NO		NO	NO	NO
DPOW Std. Dev.						
DPOW Range						
Shimmer			NO	NO	NO	NO

A = anger, D = disgust, F = fear; J = joy, N = neutral, Sd = sadness, Sp = surprise.

TABLE VIII
MEAN VALUES OF PITCH-RELATED PARAMETERS

	A	D	F	J	N	Sd	Sp
Max. F0	469.6	295.3	467.5	463.4	379.8	269.3	503.4
Mean F0	320.8	212.9	354.3	310.3	265.0	196.4	312.8
F0 Std. Dev.	80.5	40.8	61.0	76.0	56.2	36.5	102.7
F0 Range	347.5	204.2	288.4	339.3	248.2	158.3	389.8
Max. DF0	70.7	38.6	82.8	63.3	38.6	41.1	94.7
Mean DF0*10^{-2}	-10	-0.59	-8.1	2.6	-1.8	-11	-3.3
DF0 Std. Dev.	4.8	3.1	5.4	4.7	3.1	3.3	7.6
DF0 Range	111.8	77.2	132.4	115.7	77.7	68.4	186.5
Jitter*10^{-2}	2.4	2.8	2.3	2.5	2.4	2.9	2.3

A = anger, D = disgust, F = fear; J = joy, N = neutral, Sd = sadness, Sp = surprise.

TABLE IX
MEAN VALUES OF POWER-RELATED PARAMETERS

	A	D	F	J	N	Sd	Sp
Max. POW	5.5	4.6	3.9	6.3	5.4	2.3	4.2
Mean POW	1.1	0.6	0.5	1.2	1.0	0.2	0.6
POW Std. Dev.	1.8	1.1	1.1	1.8	1.6	0.4	1.3
POW Range	5.5	4.6	3.9	6.3	5.4	2.3	4.2
Max. DPOW	1.5	1.2	1.0	1.7	1.4	0.6	1.1
Mean DPOW*10^{-5}	0.16	-11	0.92	-22	0.34	-2.1	-1.3
DPOW Std. Dev.	0.4	0.3	0.3	0.5	0.4	0.0	0.3
DPOW Range	2.9	2.3	2.0	3.2	2.8	1.2	2.1
Shimmer*10^{-1}	1.4	1.6	1.4	1.5	1.5	1.6	1.4

A = anger. D = disgust. F = fear; J = joy. N = neutral. Sd = sadness. Sp = surprise.

When applying ANOVA to power related features, the results indicate that there are significant differences between both corpora for nearly all the parameters. This is shown in Table VII. Only the mean value of the power slope has been used in the same way in both corpora. This difference might be due to the fact of having divided the recording into two separate sessions and not having given the speaker any reference for the volume she had to use.

E. *Analysis of the Prosodic Features of Emotion*

When applying ANOVA to the values of the parameters measured in the Common Subcorpus to determine whether differences among distributions were significant for the different emotions, most of them (70% of the cases) were significant with a confidence interval of 95%. The same analysis was applied to values measured in the Specific Subcorpus and in this case fewer pairs were found significant (66% of the cases). This is probably due to the fact that the speaker overacted in the Common Subcorpus to distinguish emotions that could not be differentiated by the text content. In the Specific Subcorpus, as semantics indicated what the intended emotion was, she acted more naturally with less exaggerated emotions. These results indicate that either subcorpus can be used alike from an acoustic point of view and considering the measured long-term averaged parameters.

Mean values for all the prosodic parameters were measured in the database. This was done separately for all emotions and the results are shown in Tables VIII and IX. Sadness is the emotion with the lowest values in almost all the parameters measured, except for jitter and shimmer. These two parameters have their maximum values in sadness and their minimum in fear and surprise. Surprise is the one with highest values in almost all the pitch related features, except for: 1) mean F0 that is higher in fear; 2) mean value of F0 slope which is larger in joy; and 3) jitter which is higher in sadness. Joy is the emotion with the highest values in most of power-related features. These values can be used to configure the global F0 and volume values in an emotion-enabled synthesizer.

V. ROLE OF PROSODY IN THE IDENTIFICATION OF EMOTIONS

The goal of the experiments described in this section was to validate the use of the Common Subcorpus by checking the power of a given set of prosodic parameters to discriminate among the emotions in that Subcorpus, using an automatic recognizer. The same tests are also performed with the Specific Subcorpus, so results of both tests can be compared.

A. *Automatic Identification Experiments*

We made an experiment on automatic identification of emotions using Support Vector Machines (SVMs) [26] and the statistics of prosodic features. For our experiment, the set of pitch- and power-related features that was used in the acoustic analysis of the database presented in the previous section was augmented by adding the second difference of pitch and power curves and providing the information also in logarithmic scale. Different

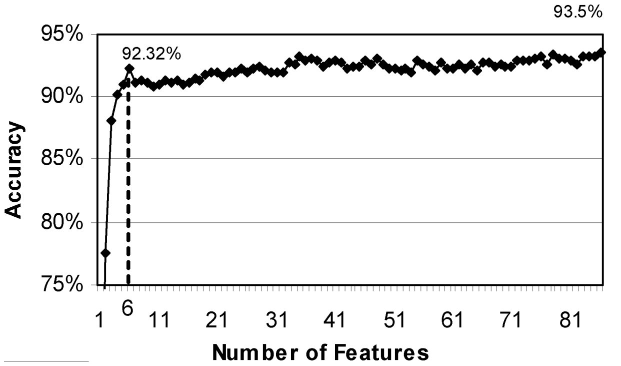


Fig. 5. Recognition accuracy with prosodic features and SVM, for different number of features in the common part of the corpus.

statistical features were computed over these curves, considering also the information provided by the voiced/unvoiced and voice activity detectors. A total of 86 prosodic features were calculated for each utterance.

To compute the recognition accuracy a jack-knife test procedure was used. The utterances corresponding to each emotion were randomized and then divided into five blocks. This randomization ensures that the blocks are balanced as the database contains different types of utterances (from isolated words to long sentences). Then, five different systems were trained and five tests were made with a leave-one-out method. When all five tests were finished, the overall confusion matrix and classification accuracy were calculated. This accuracy was estimated as the number of correctly classified utterances normalized to the total number of utterances in the tests.

The LibSVM v2.6 function library [27] was used for the training and testing of SVM. A radial basis function (RBF) kernel and the one-against-one approach were used for multiclass classification.

In a first experiment, all 86 features that had been extracted were used, achieving an overall accuracy of 93.5% for the Common Subcorpus and 84.3% for the Specific Subcorpus. This difference in the identification score can be due to the overacting of the speaker and it is consistent with the results of the ANOVA.

However, it can be expected that many of the features used are redundant since two versions of the same curves of pitch and power are used to calculate these statistical features, one in linear scale and the other in logarithmic scale. Therefore, a feature selection process was implemented. A forward 3-backward 1-wrapper method [28] was used for the selection of features. During this process, the feature which maximizes the system's accuracy is selected in each step. The accuracy is obtained by training a whole new classifier with a jack-knife test. After three consecutive selections have been made, the least useful, i.e., least influential feature is taken out.

Results of this experiment for the Common Subcorpus are shown in Fig. 5. With as few as six features, 92.32% of accuracy is obtained, only 1.18% less than using all 86 features. Even though using fewer features gives a slightly worse performance than using all 86, the computational cost of extracting all these features and training such a complex system may not be worthwhile. Table X presents the confusion matrix for the case in which only the six most influential features were used. These features were as follows:

— mean pitch;
— mean energy;
— pitch variance;
— skew of logarithmic pitch;
— range of logarithmic pitch;
— range of logarithmic energy;

TABLE X
CONFUSION MATRIX OF THE AUTOMATIC IDENTIFICATION OF EMOTIONS IN THE COMMON SUBCORPUS WITH SVM

	A	D	F	J	N	Sd	Sp
A	94.8%	1.0%	0.0%	2.1%	0.0%	0.0%	0.0%
D	0.0%	82.5%	0.0%	0.0%	3.1%	4.1%	0.0%
F	0.0%	0.0%	96.9%	0.0%	0.0%	0.0%	9.5%
J	2.1%	0.0%	0.0%	90.7%	1.0%	0.0%	0.0%
N	1.0%	6.2%	0.0%	7.2%	94.8%	0.0%	0.0%
Sd	2.1%	10.3%	0.0%	0.0%	1.0%	95.9%	0.0%
Sp	0.0%	0.0%	3.1%	0.0%	0.0%	0.0%	90.5%

Confusion matrix of the automatic identification experiment in the Common Subcorpus using SVM and prosodic features. Columns contain true values and rows values identified by the system.

A = anger, D = disgust, F = fear; J = joy, N = neutral, Sd = sadness, Sp = surprise.

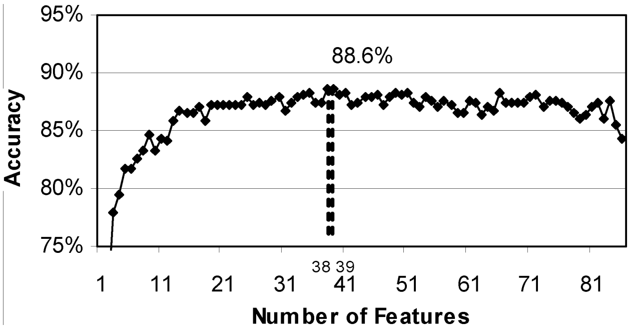


Fig. 6. Recognition accuracy with prosodic features and SVM, for different number of features in the specific part of the corpus.

In the case of the Specific Subcorpus, the identification rate using all the 86 features was 84.3%. Fig. 6 shows the evolution of identification accuracy for the different sets of prosodic features selected. The selection of the optimum set of features gave 88.6% accuracy when using 38 or 39 features. Table XI shows the confusion matrix when the best set of 38 prosodic features is used.

For purposes of reference and comparison, a recognition system using GMM and using only short-term spectral features was also trained with both subcorpora, using HTK [29]. The results of the experiments for a 512 mixture, using MFCC and their first and second derivatives are shown in Tables XII and XIII for Common and Specific Subcorpora respectively. Experiments using prosodic features and GMM have also been performed, offering results that lead to the same conclusions, although they present different recognition rates.

B. Conclusions

Just as has happened in the subjective evaluation test for non-Basque speaking evaluators, the recognition results are better

TABLE XI
CONFUSION MATRIX OF THE AUTOMATIC IDENTIFICATION OF EMOTIONS IN SPECIFIC SUBCORPUS WITH SVM

	A	D	F	J	Sd	Sp
A	86.7%	2.7%	0.0%	5.1%	0.0%	4.0%
D	4.0%	89.3%	2.7%	2.5%	0.0%	4.0%
F	1.3%	4.0%	93.3%	1.3%	0.0%	0.0%
J	4.0%	1.3%	4.0%	82.3%	0.0%	12.0%
Sd	0.0%	1.3%	0.0%	0.0%	100.0%	0.0%
Sp	4.0%	1.3%	0.0%	8.9%	0.0%	80.0%

Confusion matrix of the automatic identification experiment in the Specific Subcorpus using SVM and prosodic features. Columns contain true values and rows values identified by the system.

A = anger, D = disgust, F = fear; J = joy, Sd = sadness, Sp = surprise.

TABLE XII
CONFUSION MATRIX OF THE AUTOMATIC IDENTIFICATION OF EMOTIONS IN COMMON SUBCORPUS WITH GMM

	A	D	F	J	N	Sd	Sp
A	100.0%	0.0%	0.0%	0.0%	1.0%	0.0%	0.0%
D	0.0%	95.9%	0.0%	0.0%	0.0%	0.0%	0.0%
F	0.0%	0.0%	99.0%	0.0%	0.0%	0.0%	0.0%
J	0.0%	0.0%	0.0%	95.9%	0.0%	0.0%	0.0%
N	0.0%	3.1%	0.0%	4.1%	97.9%	0.0%	0.0%
Sd	0.0%	1.0%	0.0%	0.0%	1.0%	100.0%	0.0%
Sp	0.0%	0.0%	1.0%	0.0%	0.0%	0.0%	100.0%

Confusion matrix of the automatic identification experiment in the Common Subcorpus using GMM and MFCC. Columns contain true values and rows values identified by the system.

A = anger, D = disgust, F = fear; J = joy, N = neutral, Sd = sadness, Sp = surprise.

TABLE XIII
CONFUSION MATRIX OF THE AUTOMATIC IDENTIFICATION OF EMOTIONS IN SPECIFIC SUBCORPUS WITH GMM

	A	D	F	J	Sd	Sp
A	88.00%	0.00%	1.33%	4.00%	0.00%	1.33%
D	4.00%	94.67%	0.00%	0.00%	1.33%	2.67%
F	0.00%	0.00%	96.00%	0.00%	0.00%	1.33%
J	1.33%	1.33%	0.00%	86.67%	0.00%	1.33%
Sd	0.00%	0.00%	1.33%	0.00%	98.67%	0.00%
Sp	6.67%	4.00%	1.33%	9.33%	0.00%	93.33%

Confusion matrix of the automatic identification experiment in the Specific Subcorpus using GMM and MFCC. Columns contain true values and rows values identified by the system.

A = anger, D = disgust, F = fear; J = joy, Sd = sadness, Sp = surprise.

TABLE XIV
SUMMARY OF IDENTIFICATION RESULTS

	Best identified		->	->	->	Worst identified	
SPK	Sp	J	N	Sd	A	F	D
SPK	82.7%	82.0%	78.7%	73.3%	73.0%	66.0%	18.7%
SPK	1.0	1.0	0.9	0.9	0.8	0.7	0.0
SVM	F	Sd	A	N	J	Sp	D
SVM	96.9%	95.9%	94.8%	94.8%	90.7%	90.5%	82.5%
SVM	1.0	0.9	0.9	0.9	0.6	0.6	0.0
GMM	Sp	Sd	A	F	N	J	D
GMM	100.0%	100.0%	100.0%	99.0%	97.9%	95.9%	95.9%
GMM	1.0	1.0	1.0	0.8	0.5	0.0	0.0

Summary of identification results: for each experiment first row shows the emotions ordered from the best recognized one to the worst recognized one; second row shows the corresponding identification rates and third row shows these rates normalized to the range [0, 1].

SPK = Basque speakers, SVM = Support Vector Machines with prosodic features, GMM = Gaussian Mixture Models with spectral features; A = anger, D = disgust, F = fear; J = joy, N = neutral, Sd = sadness, Sp = surprise.

for the Common Subcorpus for the two classifiers and parameter types used. This confirms the hypothesis made about the overacting of the actress for texts not coordinated with emotions. The need of more parameters to achieve the maximum percentage of recognition might also be due to the Specific Subcorpus showing less discrimination among emotions. This is coherent with the results for the ANOVA analysis where fewer cases were found significant (70% of the cases in the Common Subcopus versus 66% for the Specific Corpus).

It is also interesting to study the different behavior of every emotion in the automatic systems, and to compare it with the ability of people to identify them. Considering that the automatic systems have not been provided with semantic information, we restrict this analysis to the results obtained with the Common Subcorpus. Also, for the subjective evaluation case, we will consider only the Basque Speakers test, to eliminate the socio-cultural variables that can be involved in the decision of the non-Basque Speakers group. Table XIV shows a summary of the identification percentages for the three cases, ordered from higher to lower recognition rate. Also in this table, the identification rate normalized to the range [0, 1] is shown for the sake of an easier comparison.

The worst identified emotion in all cases has been disgust. It is clear that the analyzed parameters do not show clear cues in this database. This is the emotion that is closest to chance level in the subjective evaluation. Considering the confusion matrix, disgust has been mistaken for neutral style or sadness in all the three experiments. The difficulty of this emotion has also been reported in [19], [23], and [24].

Surprise—the best subjectively identified emotion—seems to have better expression in the short term spectral features than in prosodic ones where the recognition percentage is not very good. Surprise is confused with fear and anger in the automatic systems and with neutral style and anger by people.

It is remarkable that joy, an emotion that has been very well identified by subjects, offers very poor results in the automatic systems. This indicates that this emotion uses other cues shown neither in the statistics of the prosody parameters nor in the statistics of the short-term spectral features.

On the contrary, fear is an emotion that shows very good performance on both automatic systems while people have not shown special ability in its identification. This fact questions the use of long-term statistics by humans, giving rise to the question of the primary cues used by humans in emotion identification.

It is also interesting to compare the results of the automatic recognition using prosody with other experiments performed elsewhere. Results in emotion recognition experiments are hard to compare because different database designs are used: Some use acted speech, whereas others collect real emotions, some are

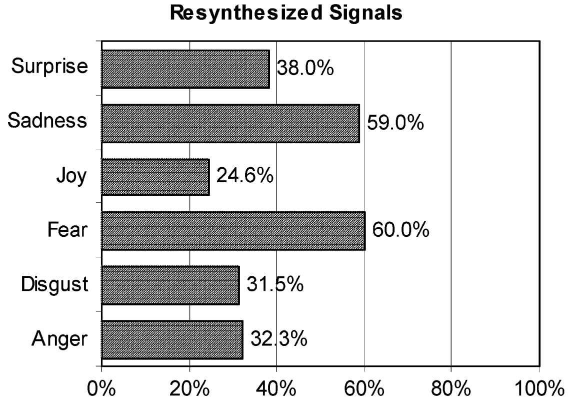


Fig. 7. Results of the subjective evaluation process for resynthesized signals.

multispeaker and others are not, different basic emotions sets are considered, etc.

The work presented in [30] is possibly the closest one to our experiment as acted speech and just one speaker are used, with the same set of emotions. In [30], the Interface database is used to train and test a speaker dependent emotional speech classifier based on a total of 144 prosodic features. Accuracy between 60%–90% is achieved depending on the language. Other works in emotion recognition in speech get an average performance of 78.1% [8], 63.5% in a multispeaker framework [31], 80.7% with only three emotions and neutral style [32], and 82.5% in a multispeaker environment using HMM [33].

VI. Resynthesis Experiment

Automatic identification of the emotions has shown quite good performance even when using only the statistics of the prosodic parameters. This raises the question about the possibilities of using only prosody manipulation to transform a neutral speech sentence into an emotional vocal expression. The purpose of the experiment here described was to validate or reject this possibility by means of a resynthesis experiment.

Pitch curves from sentences pronounced with emotion in Common Subcorpus have been superimposed in the corresponding neutral style sentence using Praat resynthesis tool [34]. Sixty of these resynthesized sentences (ten per emotion) have been used as stimuli in a subjective test taken by 13 Basque speakers. Sound duration and power have not been changed, and F0 curve has been projected at phone level.

Results of this test are shown in Fig. 7. The recognition scores have dropped drastically when compared with the results obtained with natural sentences that are presented in Fig. 2 meaning that prosody alone is not capable of carrying all the emotional content. The improvement on disgust can be attributed to the absence of the "neutral" choice in this test.

The best-recognized emotions are fear and sadness, followed by surprise, with a recognition score very close to the one achieved with natural sentences. These two emotions also have the best recognition score in the automatic system using prosodic parameters. As a consequence, it might be said that fear and sadness are two emotions whose synthesis can be successful by prosody manipulation techniques.

VII. Conclusion

The main goal that led to the work presented here was the validation of the use of a neutral semantic content corpus to build an emotional corpus for TTS for Basque. The compilation of large corpora with emotion-dependent text is laborious. This is especially true for Basque and other minority languages where large corpora are sparse or in the process of construction and tools for semantic analysis do not exist yet [35].

Subjective evaluation and automatic recognition tests have shown that emotions in the neutral content texts were overacted by the speaker. However, the emotions present high identification rates and similar behavior in both Common and Specific subcorpora, and no significant differences were found by the ANOVA test in the analyzed statistics of the prosodic parameters. This indicates that neutral content texts can be used for speech synthesis, although probably they will lead to an slightly exaggerated expressivity. Due to the overacting, these kind of texts should not be used to generate databases for emotion recognition.

The performed analysis has also provided very valuable information about the power of the prosodic parameters to discriminate among emotions. In this paper, only prosodic parameters—in particular F0 and energy—have been used. No analysis has been performed on the other prosodic related parameters such as breaks and the duration of sounds. The study of these parameters requires the accurate segmentation of the speech signals, and has been left for future work. The performed objective analysis has been limited to the statistically averaged values of the parameters. The resynthesis experiment shows that prosody transferring alone cannot provide all the information needed to identify one emotion.

Another interesting result of the study has been the difficulty in the characterization and identification of disgust. The subjective analysis performed shows that semantic information has been essential for the recognition of this emotion. When this semantic information is not present, the recognition rate has been very close to chance level.

Finally, some emotions that have been easily identified by evaluators in neutral content texts get poor results in the tested automatic recognition systems. Besides using other prosodic features, the use of voice quality parameters will improve the characterization of the emotions as has already been suggested in previous works for other languages [36].

Acknowledgment

The authors would like to thank all of the evaluators that took part in the subjective evaluation processes.

References

[1] P. Gebhard, M. Klesen, and T. Rist, "Coloring multi-character conversations through the expression of emotions," *Lecture Notes Artif. Intell.*, vol. 3068, pp. 128–141, Jun. 2004.

[2] J. I. Hualde, J. A. Lakarra, and R. L. Trask, Eds., *Toward a History of the Basque Language*. Amsterdam, The Netherlands: Benjamins, 1996.

[3] R. L. Trask, *The History of Basque*. Evanston, IL: Routledge, 1997.

[4] E. Douglas-Cowie, R. Cowie, and M. Schröder, "A new emotion database: Considerations, sources and scope," in *Proc. ISCA Workshop on Speech Emotion*, Belfast, Northern Ireland, 2000, pp. 39–44.

[5] N. Campbell, "Building a corpus of natural speech—And tools for the processing of expressive speech—the JST CREST ESP project," in *Proc. 7th Eur. Con. Speech Commun. Technol.*, Aalborg, Denmark, 2001, pp. 1525–1528.

[6] I. Karlsson, T. Banziger, J. Dankovicová, T. Johnstone, J. Lindberg, H. Melin, F. Nolan, and K. Scherer, "Speaker verification with elicited speaking-styles in the verivox project," *Speech Commun.*, vol. 31, no. 2, 3, pp. 121–129, Jun. 2000.

[7] N. Amir, S. Ron, and N. Laor, "Analysis of an emotional speech corpus in Hebrew based on objective criteria," in *Proc. ITRW Speech Emotion*, Newcastle, Northern Ireland, 2000, pp. 29–33.

[8] T. Lay Nwe, S. W. Foo, and L. De Silva, "Speech emotion recognition using hidden Markov models," *Speech Commun.*, vol. 41, no. 4, pp. 603–623, Nov. 2003.

[9] A. Iida and N. Campbell, "A database design for a concatenative speech synthesis system for the disabled," in *Proc. 4th ISCA Workshop Speech Synth.*, Edinburgh, U.K., 2001, pp. 189–194.

[10] V. Makarova and V. Petrushin, "RUSLANA: A database of Russian emotional utterances," in *Proc. ICSLP*, Denver, CO, 2002, pp. 2041–2044.

[11] T. Seppänen, J. Toivanen, and E. Väyrynen, "MediaTeam speech corpus: A first large Finnish emotional speech database," in *Proc. 15th Int. Congr. Phonetic Sci.*, Barcelona, Spain, 2003, pp. 2469–2472.

[12] J. M. Montero, J. M. Gutiérrez-Arriola, S. Palazuelos, S. Aguilera, and J. M. Pardo, "Emotional speech synthesis: From speech database to TTS," in *Proc. ICSLP*, vol. 3, Sydney, Australia, 1998, pp. 923–926.

[13] D. Küstner, R. Tato, T. Kemp, and B. Meffert, "Toward real life applications in emotions recognition," *Lecture Notes Artif. Intell.*, vol. 3068, pp. 25–35, Jun. 2004.

[14] E. Douglas-Cowie, N. Campbell, R. Cowie, and P. Roach, "Emotional speech: Toward a new generation of databases," *Speech Commun.*, vol. 40, no. 1, 2, pp. 33–60, Apr. 2003.

[15] N. Campbell, "Getting to the heart of the matter; speech is more than just the expression of text or language (keynote speech III)," in *Proc. LREC*, vol. 3, Lisbon, Portugal, 2004, pp. VII–X.

[16] R. Cowie and R. R. Cornelius, "Describing the emotional states that are expressed in speech," *Speech Commun.*, vol. 40, no. 1, 2, pp. 2–32, Apr. 2003.

[17] K. R. Scherer, "Vocal communication of emotion: A review of research paradigms," *Speech Commun.*, vol. 40, no. 1, 2, pp. 227–256, Apr. 2003.

[18] P. B. de Mareüil, P. Célérier, and J. Toen, "Generation of emotions by a morphing technique in English, French, and Spanish," in *Proc. Speech Prosody*, Aix-en Provence, France, 2002, pp. 187–190.

[19] A. Iida, N. Campbell, F. Higuchi, and M. Yasumura, "A corpus-based speech synthesis system with emotion," *Speech Commun.*, vol. 40, no. 1, 2, pp. 161–187, Apr. 2003.

[20] I. S. Enberg, A. V. Hansen, O. Andersen, and P. Dalsgaard, "Design, recording, and verification of a Danish emotional speech database," in *Proc. 5th Eur. Conf. Speech Commun. Technol.*, Grenoble, France, 1997, pp. 1695–1698.

[21] V. Hozjan, Z. Kacic, A. Moreno, A. Bonafonte, and A. Nogueiras, "Interface databases: Design and collection of a multilingual emotional speech database," in *Proc. 3rd Int. Conf. Language Resources Eval.*, Las Palmas, Spain, 2002, pp. 2019–2023.

[22] A. Paeschke and W. F. Sendlmeier, "Prosodic characteristics of emotional speech; measurements of fundamental frequency movements," in *Proc. ISCA Workshop Speech Emotion*, Belfast, Northern Ireland, 2000, pp. 75–80.

[23] F. Burkhardt and W. F. Sendlmeier, "Verification of acoustical correlates of emotional speech using formant-synthesis," in *Proc. ISCA Workshop Speech Emotion*, Belfast, Northern Ireland, 2000, pp. 151–156.

[24] I. Iriondo, R. Guaus, A. Rodríguez, P. Lázaro, N. Montoya, J. M. Blanco, D. Bernardas, J. M. Oliver, D. Tena, and L. longhi, "Validation of an acoustical modeling of emotional expression in Spanish using speech synthesis techniques," in *Proc. ISCA Workshop Speech Emotion*, Belfast, Northern Ireland, 2000, pp. 161–166.

[25] J. Ramirez, J. Segura, C. Benitez, A. de la Torre, and A. Rubio, "Efficient voice activity detection algorithms using long term speech information," *Speech Commun.*, vol. 42, no. 3, 4, pp. 271–287, Apr. 2004.

[26] N. Cristianini and J. Shawe-Taylor, *An Introduction to Support Vector Machines (and Other Kernel-Based Learning Methods)*. Cambridge, U.K.: Cambridge Univ. Press, 2000.

[27] C. Chang and C. Lin. (2005) LIBSVM: A library for support vector machines. [Online]. Available: http://www.csie.ntu.edu.tw/~cjlin/libsvm

[28] P. Pudil, J. Novovicová, and P. Somol, "Feature selection toolbox software package," *Pattern Recognition Lett.*, vol. 23, pp. 487–792, 2002.

[29] S. Young, J. Odell, D. Ollason, V. Valchev, and P. Woodlans, *The HTK Book*. Cambridge, U.K.: Cambridge Univ. Press, 2000.

[30] V. Hozjan and Z. Kacic, "Improved emotion recognition with large set of statistical features," in *Proc. Eurospeech*, Geneva, Switzerland, 2003, pp. 133–136.

[31] V. A. Petrushin, "Emotion recognition in speech signal: Experimental study, development, and application," in *Proc. ICSLP*, Beijing, China, 2000, pp. 222–225.

[32] T. Seppänen, E. Väyrynen, and J. Toivanen, "Prosody based classification of emotions in spoken Finnish," in *Proc. Eurospeech*, Geneva, Switzerland, 2003, pp. 717–720.

[33] N. Nogueiras, A. Moreno, A. Bonafonte, and J. Mariño, "Speech emotion recognition using hidden markov models," in *Proc. Eurospeech*, Aalborg, Denmark, 2001, pp. 2679–2682.

[34] P. Boersma and D. Weenink. (2005) Praat: Doing phonetics by computer (Version 4.3.16). [Online]. Available: http://www.praat.org/

[35] A. D. de Ilarraza, A. Gurrutxaga, I. Hernáez, N. L. de Gereñu, and K. Sarasola, "HIZKING21: Integrating language engineering resources and tools into systems with linguistic capabilities," in *Proc. Workshop NLP Minority Languages Small Languages*, Batz-sur-Mer, France, 2003.

[36] M. Schroeder, "Can emotions be synthesized without controlling voice quality?," *Phonus 4*, pp. 35–50, 1999.

Eva Navas was born in San Sebastián, Spain, in 1971. She received the university degree and the Ph.D. degree in telecommunication engineering from the Department of Electronics and Telecommunications, University of the Basque Country, Bilbao, Spain, in 1996 and 2003, respectively.

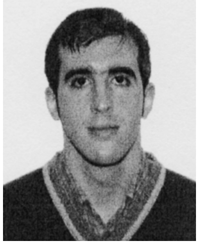
She is a Researcher at the Speech Signal Processing Group (Aholab), Department of Electronics and Telecommunications, University of the Basque Country. She is currently teaching at the Faculty of Industrial and Telecommunication Engineering in Bilbao. She has participated as a Research Engineer in public R&D projects as well as in private contracts. She works mainly in generating synthetic prosody for neutral and emotional speech.

Dr. Navas is member of the International Speech Communication Association (ISCA), the Spanish Thematic Network on Speech Technologies (RTHabla), and the European Center of Excellence on Speech Synthesis (ECESS).

Inmaculada Hernáez received the telecommunications engineering degree from the Universitat Politecnica de Catalunya, Barcelona, Spain, and the Ph.D. degree in telecommunications engineering from the University of the Basque Country, Bilbao, Spain, in 1987 and 1995, respectively.

She is a Full Professor with the Electronics and Telecommunication Department, Faculty of Engineering, University of the Basque Country, Spain, in the area of signal theory and communications. Her research interests are signal processing and all aspects related to speech processing. She is also interested in biometric systems, signature identification, and the use of multimodality.

Dr. Hernáez is a member of the International Speech Communication Association (ISCA), the Spanish thematic network on Speech Technologies (RTHabla), and the European Center of Excellence on Speech Synthesis (ECESS).



Iker Luengo was born in Bilbao, Spain, on October 29, 1979. He received the B.S. degree in telecommunication engineering from the University of the Basque Country, Bilbao, Spain, in 2003, where he is currently working toward the Ph.D. degree in telecommunications.

His research interests include emotional speech, speaker recognition, and voice prosody.

Mr. Luengo is member of the International Speech Communication Association (ISCA) and the Spanish thematic network on Speech Technologies (RTHabla).