Acoustical Analysis of Emotional Speech in Standard Basque for Emotions Recognition

Eva Navas, Inmaculada Hernáez, Amaia Castelruiz, Jon Sánchez, and Iker Luengo

Departamento de Electrónica y Telecomunicaciones, Escuela Técnica Superior de Ingeniería,
University of the Basque Country, Alameda Urquijo s/n, 48013 Bilbao, Spain
{eva,inma,amaia,ion,ikerl}@bips.bi.ehu.es
http://bips.bi.ehu.es

Abstract. This paper presents the acoustical study of an emotional speech database in standard Basque to determine the set of parameters that can be used for the recognition of emotions. The database is divided into two parts, one with neutral texts and another one with texts semantically related with the emotion. The study is performed on both parts, in order to known whether the same criteria may be used to recognize emotions independently of the semantic content of the text. Mean F0, F0 range, maximum positive slope in F0 curve, mean phone duration and RMS energy are analyzed. The parameters selected can distinguish emotions in both corpora, so they are suitable for emotion recognition.

1 Introduction

With the progress of new technologies and the introduction of interactive systems, there has been a sudden increase in the demand for user friendly interfaces. For the correct development of such kind of interfaces, a high quality Text-to-Speech system and a system able to recognize the mood of the user are required. To build this kind of systems, a deeper research of the prosodic characteristics of emotional speech is necessary. To study the prosody of emotional speech in standard Basque, a new database that includes the six emotions considered the basic ones [1][2] (anger, disgust, fear, joy, sadness and surprise) was designed and recorded [3]. This set of basic emotions has been used in different studies related with speech, both for emotion recognition [4] and for emotion generation [5].

The corpus recorded in the database was divided into two parts: one part included emotion independent texts, which are common for all emotions. In this part neutral style has also been considered, to be used as a reference. The other part includes texts semantically related to each emotion, thus, this group is different for all the emotions. The acoustical analysis has been made separately for both parts of the database, to know whether the emotions were expressed in the same way independently of the semantic content of the text and as a result to determine the set of features that can be used for the recognition of emotions.

2 Subjective Evaluation of the Database

To prove the ability of the speaker to accurately simulate the emotions, assessing this way the validity of the database, we prepared a subjective test with the purpose of checking whether listeners could identify the intended emotion above chance level

A. Sanfeliu et al. (Eds.): CIARP 2004, LNCS 3287, pp. 386–393, 2004.

(14%). A forced choice test was designed, where users had to select one of the seven proposed styles (six emotions plus neutral style). Sentences from both the common and specific corpora were selected for the test.

A total of 15 participants took part in the experiments. Fig. 1 shows the total recognition rate obtained for each emotion and the one obtained for the common and specific texts. Signals belonging to the specific corpus are better identified, but it is difficult to determine to what extent this is due to the semantic content of the stimulus, which helps the listener to decide, or to the better expression of the emotion.

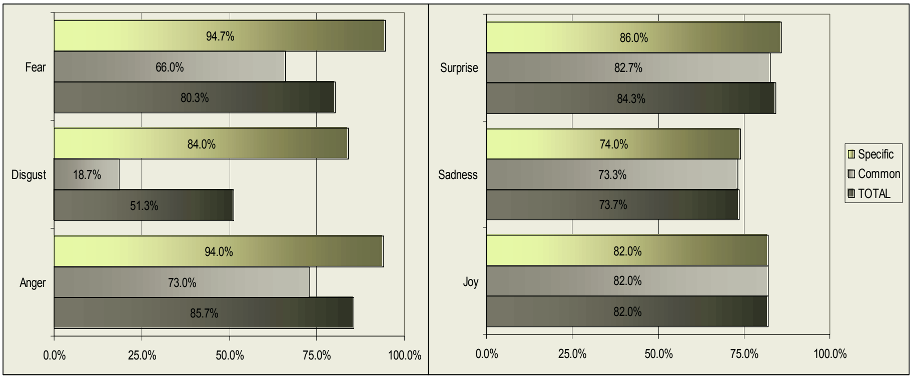

Fig. 1. Result of the subjective test, showing the recognition rate for each emotion separated by text type

Disgust is the emotion with worst recognition results. It has mainly been confused with neutral style. This emotion has also been the most difficult to identify in other works for different languages [6][7][8]. Recognition rates are similar in the common and specific corpus for surprise, sadness and joy. Anger and fear get poorer results for the common corpus, but still well above chance level. Therefore, subjective results for both corpora are good enough to use them to perform the acoustical analysis of the emotions.

3 Acoustic Analysis of the Database

Prosodic features are clearly related with emotion, but the nature of this relation has still to be determined. In this work, several acoustic parameters related with intonation, duration and energy have been automatically measured and analyzed to know how they change to express emotion. These features are mean F0 value (Hz), F0 range (Hz), maximum positive slope in F0 curve (Hz/ms), mean phone duration (ms), mean RMS (dB), mean RMS in low band, between 0 and 2000 Hz (dB) and mean RMS in high band from 2000 to 4000 Hz (dB). These values have been studied in the whole database, but also independently in the common and the specific corpora.

3.1 Comparison of the Distributions of Parameters Between Both Corpora

To know whether the speaker had expressed the emotions in 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 indicate that the parameter that has been more consistently applied by the speaker has been the maximum positive slope of the pitch curve, because none of the differences is significant. Besides, joy is the emotion that has been expressed more similarly in both corpora, because all the parameters studied, except for phone duration have differences not significant between both corpora. Anger and surprise have been expressed in a different way in the common and specific corpora, because most of the parameters have different distributions in both corpora.

When applying ANOVA to the values of the parameters measured in the common corpus, to determine whether differences among distributions were significant for different emotions, most of them were found significant (confidence interval of 95%). Some pairs were considered not significant in some parameters, but they were not the same pairs for every parameter, so a set of parameters that distinguish emotions can always been found in this corpus. The same analysis was applied to the values calculated from the specific corpus, and in this case, more pairs were found no significant, probably due to the fact that the speaker overacted in the common corpus to distinguish emotions that could not be differentiated by the content. In the specific corpus, as semantics indicated which one was the intended emotion, the speaker acted more naturally with less exaggerated emotions.

3.2 Analysis of the Pitch Features

The values of F0 curve were obtained from data provided by the laryngograph with a temporal resolution of 1 ms. Table 1 shows the mean values for the intonation related parameters and their related standard deviations separated by emotion. The first three columns list the values measured in the specific corpus and the rest of the columns have the values of the common corpus. Sadness is the emotion with lower values in all the parameters measured. Surprise is the one with wider pitch range. With regard to F0 maximum positive slope, anger has the larger one in the specific corpus and joy in the common corpus. The emotion with higher mean pitch value is fear.

Fig. 2 displays in the first graph the mean value of the parameters related with intonation for all emotions. Concerning the mean F0, the values corresponding to the specific texts are lower than those corresponding to the common texts for all emotions. This also happens with F0 range, as can be seen in the second graph of Fig. 2, but does not with maximum positive slope of pitch, as the third graph shows.

For all the parameters related with intonation, the emotions have the same relation with neutral style in both corpora: sadness is below neutral level in the three parameters, disgust is below for mean and F0 range and fear and surprise are below neutral level in maximum positive slope of F0.

Table 1. Values of the intonation parameters measured in the specific corpus (*first three columns*) and in the common corpus (*4th to 6th columns*), separated by emotion. Mean value and standard deviation are shown in the form mean ± standard deviation

Emotion	Mean F0	Range F0	MPS F0	Mean F0	Range F0	MPS F0
Anger	256.7±51.9	282.5±79.1	12.3±5.3	370.8±36.5	382.8±73.9	11.6±5.2
Disgust	206.8±33.7	201.4±59.7	9.5±3.9	217.8±28.7	190.3±53.4	8.8±4.0
Fear	322.2±44.2	265.6±104.6	5.5±1.3	379.2±36.3	302.2±112.4	5.6 ±1.5
Joy	306.6±32.1	320.0±80.0	10.9±4.4	314.1±30.8	327.5±87.9	12.3±4.4
Sadness	175.7±21.1	144.0±44.2	2.3±0.7	212.6±18.2	167.1±56.1	2.4±0.7
Surprise	280.0±33.9	371.8±52.3	5.6±1.3	339.0±35.8	398.4±61.8	5.4±1.5

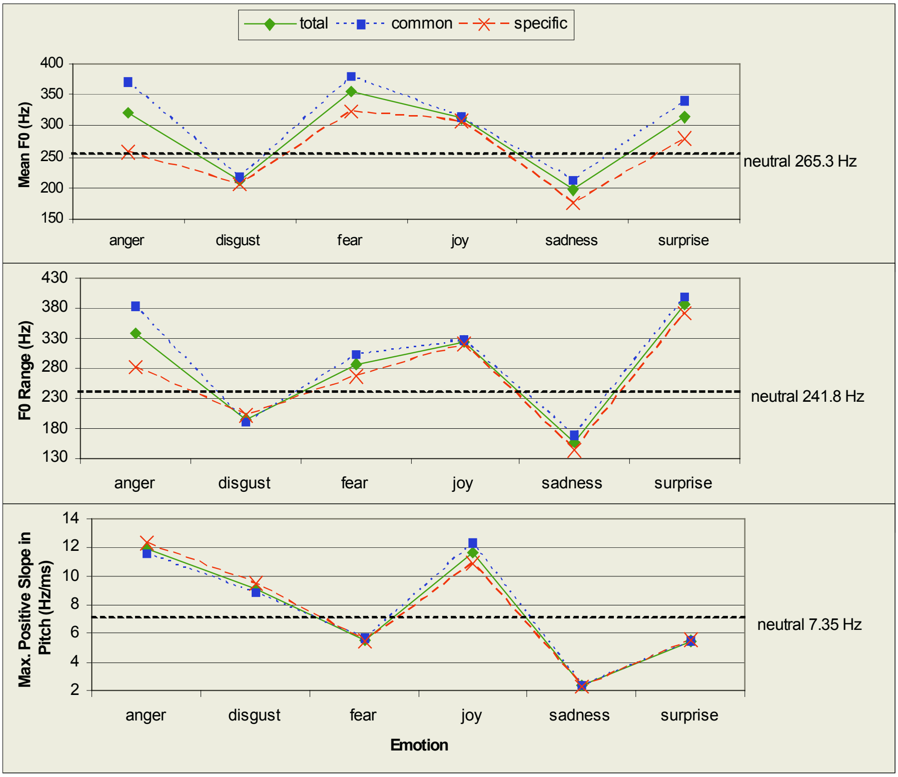

Fig. 2. Comparison of the mean pitch value in the entire database, in the part corresponding to the common texts and the specific texts

3.3 Analysis of the Duration

For the analysis of phone duration the sentences were automatically segmented using an algorithm based in Hidden Markov Models. No manual correction of the time labels was made. Duration of pauses was not considered because the database included only isolated sentences and the number of internal pauses was not sufficient for a statistical analysis.

Table 2 lists mean duration values measured in both corpora: first column has the values calculated from the specific texts and second column the values from the common texts. Sadness is the emotion with lower mean phone duration, therefore, it is the one with faster speaking rate. The emotions with slower speaking rate are fear and surprise.

Table 2. Mean value of phone duration for each emotion in the specific corpus (*1st column*) and in the common corpus (*2nd column*), expressed in ms Standard deviation is also shown in the form mean ± standard deviation

Emotion	Mean duration	Mean duration
Anger	88.5 ± 51.4	89.9 ± 52.0
Disgust	85.0 ± 45.3	84.4 ± 40.0
Fear	97.7 ± 61.1	92.2 ± 44.7
Joy	81.5 ± 46.7	78.4 ± 38.9
Sadness	78.8 ± 37.6	76.0 ± 31.9
Surprise	95.7 ± 61.1	99.5 ± 54.2

Fig. 3 shows a comparison of mean phone duration for all emotions in both corpora. Value measured for the neutral style in the common corpus is also displayed for reference: anger, fear and surprise have slower speaking rate than neutral style and joy and sadness have faster speaking rate. Disgust has a slightly faster speaker rate than neutral style.

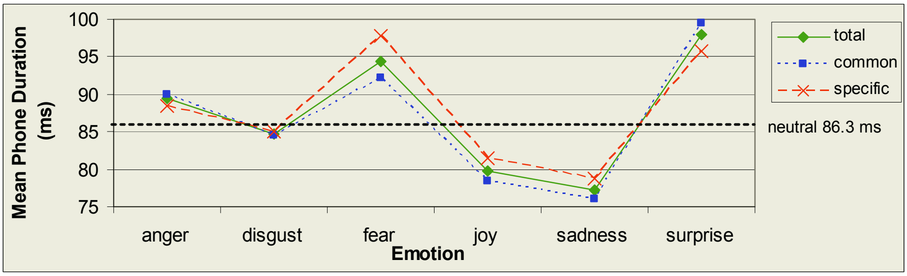

Fig. 3. Comparison of the mean phone duration in the entire database, in the part corresponding to the common texts and the specific texts

3.4 Analysis of the Power Features

For the analysis of the power features, the root mean square (RMS) energy of the signals was calculated. The spectral distribution of the energy has been considered important in other studies about emotional speech for other languages [4], so the RMS energy in the 0-2 KHz band (RMS LB) and in the 2-4 KHz band (RMS HB) have also been measured.

Table 3 lists the values related with energy measured for the specific corpus (columns 1 to 3) and the common corpus (columns 4 to 6). The emotions with more energy are anger and joy and sadness is the one with lower energy.

Table 3. Values related with energy measured in the specific corpus (*first three columns*) and in the common corpus (*4th to 6th columns*). All the parameters are expressed in dB. Mean values and standard deviation are shown in the form mean ± standard deviation

Emotion	RMS	RMS LB	RMS HB	RMS	RMS LB	RMS HB
Anger	20.5 ± 1.5	20.0 ± 1.7	15.6 ± 1.8	19.3 ± 1.5	18.0 ± 1.8	16.6 ± 1.9
Disgust	19.3 ± 2.1	19.0 ± 2.1	13.4 ± 2.9	18.4 ± 1.9	18.1 ± 1.9	13.3 ± 2.9
Fear	19.8 ± 1.3	19.5 ± 1.3	14.0 ± 2.4	16.1 ± 1.6	15.7 ± 1.6	12.2 ± 2.3
Joy	20.2 ± 1.3	19.6 ± 1.3	15.9 ± 1.7	19.8 ± 1.6	19.1 ± 1.8	15.6 ± 2.1
Sadness	16.0 ± 2.1	15.8 ± 2.1	8.2 ± 2.7	16.4 ± 2.2	16.3 ± 2.2	8.5 ± 2.7
Surprise	20.1 ± 1.6	19.7 ± 1.6	15.1 ± 1.9	16.7 ± 1.6	16.1 ± 1.6	13.3 ± 2.3

Fig. 4 shows the differences in RMS energy values in both parts of the database. In this case the values measured in the specific part are in general, higher than those of the common texts. This could be due to the fact that both parts of the database have been recorded in two different days, and no reference level was given to the speaker. RMS energy measured in low and high band have a similar behavior.

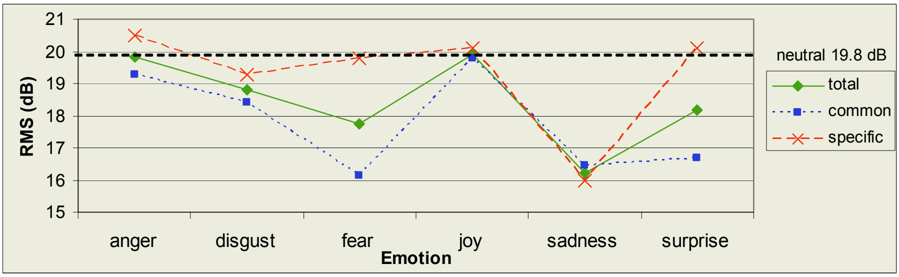

Fig. 4. Comparison of the RMS energy in the entire database, in the part corresponding to the common texts and the specific texts

3.5 Characterization of Emotions

Once the study of the acoustical correlates of emotion is performed, an analysis of their suitability to be used in the recognition of emotions is needed. Emotions are well separated, when representing the values of mean phone duration and RMS measured in both corpora, as Fig.5 shows. Fig. 6 shows the positions of different emotions when considering mean F0 and the maximum positive slope of F0 curve: emotions are also well separated according to these criteria in both corpora.

4 Conclusions

The acoustic correlates of emotion have been analysed in an emotional database for standard Basque. The analysis has been made independently for the part of the database that has common neutral texts and for the part that has specific texts related with emotions.

Subjective tests showed that both parts of the database were suitable for the study of the expression of emotions, because the recognition rates were well above chance

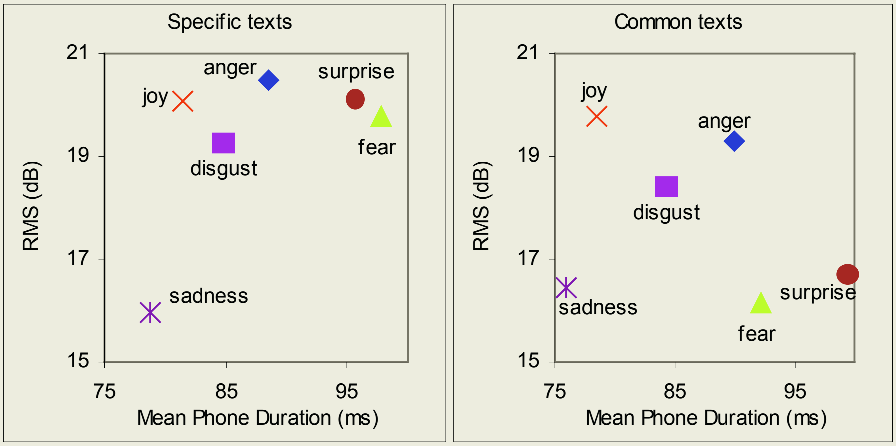

Fig. 5. Position of emotions in function of mean RMS energy and mean sound duration

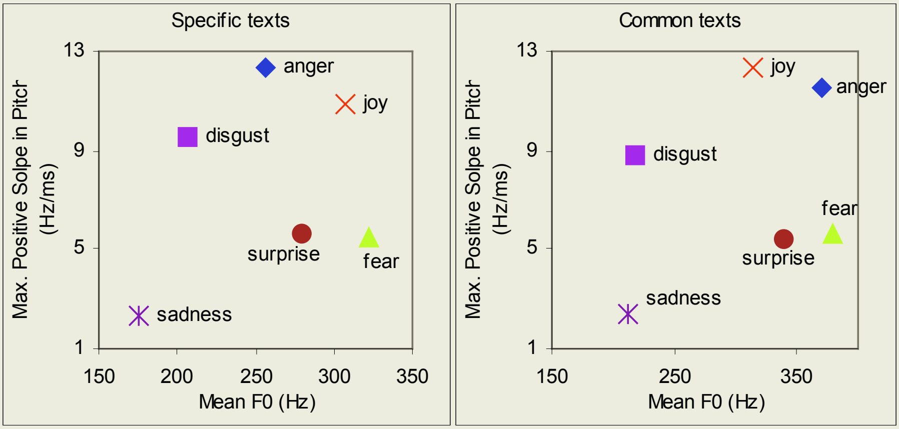

Fig. 6. Position of emotions in function of mean pitch range and maximum positive slope of pitch curve

level. Signals with text related with emotion were recognized with higher recognition rates, but signals with common texts also achieved good results.

Objective analysis of the prosodic features that characterize emotions has showed that mean F0, maximum positive slope in F0 curve, phone duration and RMS energy are suitable for emotion recognition.

Acknowledgements

This work has been partially funded by the Spanish Ministry of Science and Technology (TIC2000-1669-C04-03 and TIC2003-08382-C05-03) and by the University of the Basque Country (UPV-0147.345-E-14895/2002).

References

1. Scherrer, K.R.: Vocal Communication of Emotion: A Review of Research Paradigms. Speech Communication, Vol. 40. Elsevier, Amsterdam (2003) 227-256
2. Cowie, R., Cornelius, R.R.: Describing the Emotional States that Are Expressed in Speech. Speech Communication, Vol. 40(1,2). Elsevier, Amsterdam (2003) 2-32
3. Navas, E., Castelruiz, A., Luengo, I., Sánchez, J., Hernáez, I.: Designing and Recording an Audiovisual Database of Emotional Speech in Basque. Proc. LREC 2004. (2004)
4. Lay Nwe, T., Wei Foo, S., De Silva, L.: Speech Emotion Recognition Using Hidden Markov Models. Speech Communication, Vol. 41(4). Elsevier, Amsterdam (2003) 603-623
5. Boula de Mareüil, P., Célérier, P., Toen, J.: Generation of Emotions by a Morphing Technique in English, French and Spanish. Proc. Speech Prosody. Laboratoire Parole et Langage CNRS, Aix-en Provence (2002) 187-190
6. Iida, A., Campbell, N., Higuchi, F., Yasumura, M.: A Corpus-based Speech Synthesis System with Emotion. Speech Communication, Vol. 40(1,2). Elsevier, Amsterdam (2003) 161-187
7. Burkhardt, F., Sendlmeier, W.F.: Verification of Acoustical Correlates of Emotional Speech using Formant-Synthesis. Proc. ISCA Workshop on Speech and Emotion. ISCA Archive (2000) 151-156
8. Iriondo, I., Guaus, R., Rodríguez, A., Lázaro, P., Montoya, N., Blanco, J.M., Bernardas, D., Oliver, J.M., Tena, D., longhi, L.: Validation of an Acoustical Modelling of Emotional Expression in Spanish using Speech Synthesis Techniques. Proc. ISCA Workshop on Speech and Emotion. ISCA Archive (2000) 161-166