Subjective evaluation of an emotional speech database for Basque

Iñaki Sainz, Ibon Saratxaga, Eva Navas, Inmaculada Hernáez, Jon Sanchez, Iker Luengo, Igor Odriozola

Aholab - Dept. of Electronics and Telecommunications. Faculty of Engineering.
University of the Basque Country. Urkijo zum. z/g 48013 Bilbo
E-mail: inaki, ibon, eva, inma, ion, ikerl, igor @aholab.ehu.es

Abstract

This paper describes the evaluation process of an emotional speech database recorded for standard Basque, in order to determine its adequacy for the analysis of emotional models and its use in speech synthesis. The corpus consists of seven hundred semantically neutral sentences that were recorded for the Big Six emotions and neutral style, by two professional actors. The test results show that every emotion is readily recognized far above chance level for both speakers. Therefore the database is a valid linguistic resource for the research and development purposes it was designed for.

1. Introduction

Due to the progress in speech synthesis techniques during the past years, the intelligibility of most corpus based TTS systems is almost equivalent to that of human speech. However, the naturalness and fluency of the synthetic voice is far from being perceived as human like. An appropriate emotional expression represents one of the key aspects of the naturalness that is still missing in speech synthesis. Emotions can be seen as a way of both decreasing the monotony of synthetic speech and improving the human machine communication.

Several attempts have been made in recent years to develop an expressive text to speech synthesizer [Iida et al., 2003] [Murray & Arnott, 1996] [Bulut et al., 2002]. The final result has not been good enough yet so that it is perceived as natural emotion. In order to convey realistic emotions, a deep research of the prosodic characteristics (pitch curve, phoneme duration and energy curve) of emotional speech is necessary. And for this to be possible, a proper emotional database must be recorded.

While it is true that prosodic modifications have a great influence on the conveyed emotion [Vroomen et al., 1993] [Montero et al., 1999], the expressive speech generated in that way is not natural enough [Schröder, 1999], even when real prosody is being used (prosody copying) [Heuft et al. 1996]. What it is really missing are the spectral characteristics of the expressive speech [Rank & Pirker, 1998]. Instead of modeling these acoustic properties explicitly (which can be difficult), corpus-based techniques can be employed so as to do it implicitly.

In corpus based approaches each utterance is formed by the concatenation of optimal units from the database. The selection algorithm is based on minimizing a global cost function which consists of a target cost and a concatenation cost [Hunt & Black, 1996], both ranging from 0 (optimal scenario possible) to 1 (worst scenario) values.

The target cost measures the similarity between the desired unit (as predicted by the prosody module of the text to speech system) and the candidate units from the database. The concatenation cost is calculated as a measure of the goodness of the join between two units, being 0 if they are consecutive in the database. Unit selection systems succeed when there are enough candidate units for a specific target unit and therefore, it is no necessary to make further modifications to the waveform which may distort the natural quality of the voice. In order to have sufficient candidate units for each desired emotion, a large database is needed.

In this paper the evaluation of an emotional database is presented. Section 2 describes the process of corpus designing and recording. The evaluation protocol is described in Section 3 and the results are presented and discussed in Section 4. Finally some conclusions are drawn.

2. Corpus Design and Recording

The database described here was created with two main purposes. On the one hand we wanted to use it to develop an emotional corpus based TTS system for Basque. On the other hand, it wanted to be useful for prosodic and acoustic analysis of emotional speech.

2.1 Recording of emotional speech

There are different choices for recording emotional speech (spontaneous, elicited or acted), each of which has both advantages and disadvantages:

- Spontaneous emotions: While they inarguably contain the most authentic emotions, they may be difficult to obtain because of moral considerations about people's privacy. Besides, their content can hardly be controlled and therefore makes almost impossible to collect a proper database for corpus based synthesis due to the phonetic coverage constraints.
- Elicited emotions: The speaker is put into a situation to rouse a specific emotion. As each speaker may

react in a different manner even for the same situation, the emotion recorded can not be totally guaranteed. Another disadvantage is the ethical considerations in order to evoke negative situations during the recording of emotions such as sadness and fear.

- Acted emotions: This technique consist of texts read by a professional actor trying to emulate the desired emotion. This way of recording is often accused of generating exaggerated emotions that may not be felt as sincere, but the fact is that after all, they seem to be well recognized by human listeners.

The third option was chosen because of its advantages. On one hand it makes possible to control the content of the database preserving the phonetic balance and acoustical variability of the designed corpus; And on the other, it makes easier to study and compare the characteristics of each emotion. Besides, semantically neutral texts not related with the emotions were used, i.e. one single corpus was used for the recording of all the emotions. The validity of this choice was proved experimentally [Navas et al, 2006].

As far as expressive content is concerned, the Big Six emotions were considered (sadness, happiness, anger, fear, surprise and disgust) [Cowie & Cornelius, 2003] as they represent the most commonly distinguishable ones. Apart from them, neutral style was also considered in order to make possible the generation of non expressive speech with the TTS.

For the recording of approximately one hour of speech for each emotion 702 sentences were selected using corpus analysis techniques, granting both phonetic balance and diphone coverage. The corpus is described in detail in [Saratxaga et al., 2006]. It was recorded by two professional speakers: a 40 years old dubbing actor male, and a 37 years old radio speaker and actress female.

3. Evaluation Process

In order to determine the adecuacy of the emotional content of the recorded database, a subjective evaluation campaign was deployed.

3.1 Test Design

A forced choice test was carried out to discover whether the subjects could correctly recognize the emotion of the recorded speech stimuli or not. 30 stimuli for each actor were presented on a web based computer interface. They were randomized and grouped in forms containing 10 stimuli each. Evaluators could choose among the six possible emotions. There was not an option of 'Not identified' for use in cases in which the emotion is not clear. All the sentences were declarative but one which was interrogative. The average length of the sentences was 8.61 words, varying from 4 for the shorter one to 14 words.

3.2 Evaluation Protocol

Each subject took the test individually. The stimuli were presented over high quality headphones and reproduced with a standard computer sound card. No training session was offered to the listening subjects before the test and they were given no feedback of their answers during the whole session. They had to label the ten signals presented in each form and once a form had been completed they could not return to modify it. However, they were allowed to hear the utterances as many times as they wished before deciding the final answer. The listeners were also free to ask for breaks when needed, but only after a form was completed.

A total of 20 subjects participated in the evaluation (14 males and 6 females) with ages varying form 20 to 53 years. All of them were fluent in standard Basque only 11 being native speakers. Basque is a minority language and the number of speakers has increased notably in recent years due to its promotion in the educational system. We can divide the basque speaking community into two main groups: people for whom Basque is their first language (group A) and those who have learnt it as second language (group B). In Section 4 evaluation results are analysed dividing the subjects with the criteria described above, among some others.

4. Results

The results of the subjective test are shown in Table 1. Each row of the matrix represents the real emotion conveyed by the actor, while the columns show the identified emotions by the listeners. The values are percentages and letters symbolise the emotions: Anger (A), Fear (F), Surprise (S), Disgust (D), Happiness (H) and Sadness (X).

The table also includes precision (P) and recall (R) statistics. Precision measures the portion of the assigned emotions that were correct: Correct identifications / Total stimuli identified with that emotion. On the other hand, recall is measured as the portion of the correct emotions that were assigned: Correct identifications / Total stimuli with that emotion.

Actors	Listeners							
	A	F	S	D	H	X	P	R
A	81.5	2.5	5.5	9	-	1.5	0.78	0.82
F	0.5	64	3	-	1	31.5	0.68	0.64
S	6	2.5	73	1	17.5	-	0.80	0.73
D	15.5	4	3.5	67	2.5	-	0.86	0.67
H	0.5	0.5	5	-	94	-	0.81	0.94
X	-	20	1	0.5	1	77.5	0.66	0.78

Table 1. Confusion Matrix for evaluation process

It is clear that all the emotions are identified above chance level (17%) even when the corpus consisted of

semantically neutral sentences and this could hinder the recognition process. The average recognition level is 76.6%, being Happiness the best recognized emotion (94%) and Fear the worst one (64%). Sadness has the lowest precision (66%) as being selected in stimuli that were actually Fear, Disgust or Anger. The low recall but high precision of Disgust can be explained as being rarely selected during the test but with great accuracy when done so. In the same way, Happiness is a frequent choice and that is why it has such a high recall but a moderate precision.

Tables 2 and 3 illustrate the confusion matrices for each of the actors. Happiness gets the best results in both cases (96% for the female and 92% for the male). However, the worst recognized emotion differs in this case, being Fear (61%) for the actress and Disgust (59%) for the actor (but with a high precision). The difference among both emotions is narrow and they represent the two worst identified ones in both databases. The mean recognition percentage is very similar too: 75.83% for the actress and slightly higher (76.50%) for him.

Actress	Listeners							
	A	F	S	D	H	X	P	R
A	75	-	6	15	-	3	0.76	0.75
F	1	61	4	-	1	34	0.73	0.61
S	10	2	68	-	20	-	0.82	0.68
D	13	-	2	75	1	9	0.83	0.75
H	1	-	3	-	96	-	0.81	0.96
X	-	19	-	-	1	80	0.63	0.80

Table 2. Confusion Matrix for the actress

Actor	Listeners							
	A	F	S	D	H	X	P	R
A	88	4	5	3	-	-	0.81	0.88
F	1	67	2	-	1	29	0.64	0.67
S	2	3	78	2	15	-	0.79	0.78
D	18	8	5	59	4	6	0.91	0.59
H	-	1	7	-	92	-	0.81	0.92
X	-	21	2	1	1	75	0.68	0.75

Table 3. Confusion Matrix for the actor

The emotions most commonly mistaken are Fear and Sadness (Fear is identified as Sadness in 34% of the cases and Sadness confused with Fear in 20% of cases). For both actor and actress the most common confusion fall into Fear and Sadness category with mistakes that range from 19% to 34%. This confusion has also been observed in Interface database for Spanish [Nogueiras et al., 2001].

4.1 Effect of listeners on the results

A Student's t-test was carried out in order to determine whether the different characteristics of the subjects have some kind of correlation with the emotional identification results. Women get a recognition rate of 72.78% (from 68.37% to 77.18% within a 95% confidence interval) while the Men reach 77.62% (from 74.74% to 80.50% within a 95% confidence interval). The results seem significant ($t=1.80 \rightarrow p=0.071 > 0.05$) but not within a 95% confidence interval ($p=0.05$). The difference among those for whom Basque was the second language (Group B) and first language Basque speakers (Group A) are neither significant ($t=0.858 \rightarrow p=0.39 > 0.05$): The group A has an emotion identification mean of 77.12% and B has a rate of 75%. So it seems that once the conveyed message is understood, it does not matter whether basque is your first language or not.

Since the listeners were not offered a training session before the listening test, it was studied whether the listeners scored higher on the second half of the evaluation than on the first one. During the first half, a recognition rate of 72.64% is achieved (95% confidence interval: from 69.26% to 76.07%) while the last part of the evaluation gets a rate of 79.7% (95% confidence interval: from 76.26% to 83.07%). In this case, the improvement mean showed during the second part of the test is statistically important ($t=2.85 \rightarrow p=0.0044 < 0.05$) with a total recognition rate increase of 7% (almost constantly maintained for the various groups: group A, group B, Women and Men). This result is understandable because during the first stimuli and because of the forced choice and the lack of training session, the listeners may select an answer in spite of not being completely sure of it. As the test progresses the subjects learn how the actor conveys some specific emotions and therefore they can also identify the rest more easily.

5. Conclusions

Results of subjective test show that all the recorded emotions are readily recognized for both actors, far above the chance level. Therefore, this database represents a valid linguistic resource that will allow both the study and modeling of emotional speech in standard Basque and the development of a corpus based synthesis system of expressive speech which generation has already started.

6. Acknowledgements

The evaluated database was developed with the financial help of the Basque Government within the ANHITZ program (ETORTEK06/114) and of the MEC (TEC2006-13694-C03-02/TCM).
Authors would like to thank to all the listeners that participated in the subjective evaluation of the database.

7. References

Bulut, Murtaza, Shrikanth S. Narayanan, & Ann K. Syrdal. Expressive speech synthesis using a concatenative synthesizer, In *ICSLP* 2002, pp. 1265--1268

Cowie, R., Cornelius, R.R. Describing the Emotional States that Are Expressed in Speech, In *Speech Communication* 2003, 40(1,2) pp. 5--32

Heuft, B., Portele, T., & Rauth, M., Emotions in Time Domain Synthesis, In *ICSLP* 1996, pp.1974--1977

Hunt, A. and Black, A. Unit selection in a concatenative speech synthesis system using a large speech data base, In *ICASSP* 1996, pp. 373-376.Erlbaum Associates, pp. 252--262

IIda, A., Campbell, N. Higuchi, F., & Yasumura, M. (2003). A Corpus based speech synthesis system with emotion, In *Speech Communication*, 40, pp. 161--187

Montero, J. M., Gutiérrez-Arriola, J., Colás, J., Enríquez, E., & Pardo, J. M., Analysis and Modeling of Emotional Speech in Spanish, In *ICPhS* 1999, pp. 957--960

Murray, I.R. and Arnott, J.L. Synthesising emotions in speech: is it time to get excited?, In *ICSLP* 1996, pp. 1816--1819

Navas, E., Hernáez, I., Luengo, I. An Objective and Subjective Study of the Role of Semantics and Prosodic Features in Building Corpora for Emotional TTS, in *IEEE Transactions on audio, speech and language processing* 2006, vol. 14, n. 4, pp. 1117--1127.

Nogueiras, A., Moreno, A., Bonafonte, A., Mariño, J.B., Speech Emotion Recognition Using Hidden Markov Models, In *Proceedings of Eurospeech* 2001, pp. 2679--2682

Rank, E., & Pirker, H., Generating Emotional Speech with a Concatenative Synthesizer, In *ICSLP* 98, Vol. 3, pp. 671--674

Saratxaga I, Navas E., Hernaez I., Luengo I. Designing and Recording an Emotional Speech Database for Corpus Based Synthesis in Basque, In *Proceedings of the LREC* 2006, pp. 2126--2129

Schröder, M. Can emotions be synthesized without controlling voice quality?, In *Phonus 4, Research Report of the Institute of Phonetics* 1999, Saarland University, pp. 37--55

Vroomen, J., Collier, R., & Mozziconacci, S. J. L., Duration and Intonation in Emotional Speech, In *Eurospeech* 1993, Vol. 1, pp. 577--580