Subjective evaluation of an emotional speech database for Basque

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

Aholab – Department of Electronics and Telecommunication. School of Engineering. University of the Basque Country
inaki, ibon, eva, inma, ion, ikerl, igor, imanol @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.

Keywords: emotional speech database, subjective evaluation

1. Introduction

Due to the development of speech synthesis techniques, the intelligibility of most corpus-based TTS systems is at par with human speech. However, the naturalness and fluency of synthetic speech are far from equalling human speech. The adequate expression of emotions is an important factor that is still missing in speech synthesis. Emotions are the means to lessen the monotony of synthetic speech and to improve the communication between humans and the machine.

There have been enough attempts to develop an emotional speech synthesizer over the last years (Ilda and others, 2003; Murray and Arnott, 1996; Bulut and others, 2002). But the last results have fallen short of expectations. In order to express believable emotions, a deep research of the features of emotional speech prosody (tone curve, phoneme duration and energy curve) is necessary; and to be able to do that, it is essential to record an emotional speech database.

Prosodic changes have a great influence on the showed emotion (Vroomen and others, 1993; Montero and others, 1999), but the emotional speech originated in that way is not very natural (Schroeder, 1999), even if real prosody (a copy of the prosody) is used (Heuft and others, 1996). What really needs doing is to extract the spectral features of emotional speech (Rank and Pirker, 2006). For that purpose, instead of modelling acoustic features explicitly (which could be rather arduous), corpus-based techniques that will do the job implicitly could be used.

In corpus-based methods each sentence is created by joining the optimum units of the database. The algorithms used to select the units must minimize the global cost function. That function takes into account the objective cost and the linkage cost, with values between 0 –best case– and 1 –worst case–.

The objective cost measures the similarity between the searched unit (previously marked by the prosody module of the text-to-speech synthesizer) and the units in the database. The linkage cost measures the quality of unit connection (its value is 0 when the units are consecutive in the database). The systems to select units achieve good results when there is a wide choice for each given objective, because there is consequently no need to change the shape of the wave, something that would distort the naturalness of speech.

A large database is needed in order to have enough units to choose from for each emotion.

This paper presents an evaluation of an emotions database. Section 2 describes the design of the corpus and the recording process. Section 3 deals with the evaluation procedure, and the results are presented and discussed in Section 4. Finally, some conclusions are drawn.

2. Design and recording of the corpus

The database described here was created with two main objectives. On the one hand, we want to use it in order to develop a TTS system for Basque based on an emotional corpus; on the other, it needs to be useful for the analysis of the prosody and acoustics of emotional speech.

2.1. Recording of emotional speech

There are many types of emotional speech recordings (using spontaneous emotions, created emotions, simulated emotions and so forth), and each one has its advantages and disadvantages:

- Spontaneous emotions: Obviously, the recording of spontaneous emotions deals with true emotions, but it is morally unsuitable because of its disturbance of privacy. Moreover, the impossibility to control the content makes it nearly impossible to generate a suitable database for a corpus-based synthesis because its phonetic field would be limited.
- Created emotions: The speaker is drawn into a specific situation that brings about a specific emotion. Different speakers react differently to the same situation, and therefore, we cannot be sure of the type of emotion that will be recorded. Another disadvantage is the fact that it is not very ethical to create negative situations in order to record anger and sadness.
- Simulated emotions: This technique involves professional actors reading out a text and trying to give it the suitable emotion. This type of recording could inflate the emotions, and it could be noticed that they are not real; nevertheless, the listener does seem to perfectly recognize them.

For this research work, we have chosen the third method because of its advantages: on the one hand, the content of the database can be detailed, keeping the phonetic balance and acoustic variability of the designed corpus; on the other hand, it makes it easier to analyse and compare the features of each emotion.

We have used for our purpose semantically neutral sentences unrelated to each emotion, and we have used the same collection of texts to record all the emotions. The validity of this alternative is experimentally proved (Navas and others, 2006).

As for the expression content, the Big Six emotions have been taken into account –Anger, Fear, Surprise, Disgust, Happiness and Sadness– (Cowi and Cornelius, 2003) because they are generally the most distinctive. As well as those, we have also used the neutral style for the creation of an emotions free speech with the TTS.

To be able to get an hour's recording for each emotion, 702 sentences were picked, assuring both the phonetic balance and the diphonemic coverage by means of corpus analysis techniques. The corpus is described in detail in an article by Saratxaga and others (2006). Two professional speakers recorded it: a 40-year-old male dubbing actor and a 37-year-old woman, a radio presenter and actress.

3. Evaluation process

In order to know if the emotional content of the recorded database is suitable, a subjective evaluation is needed.

3.1. Design of the research

Listeners took a fixed-choice test to know if they were capable of recognizing the recorded pieces of speech correctly. They were presented with 30 pieces of speech by each actor via a web-based computer network. The pieces of speech were mixed and classified in questionnaires of ten pieces each. The evaluators had to choose between the six emotions. "Unknown" was not given as an option, not even when the emotion was unclear. All were declarative sentences, with the exception of a question. The average length of the sentences was of 8.61 words; the shortest had 4 and the longest 14.

3.2. Evaluation protocol

Each listener took the test individually. They listened to the pieces of speech provided by an ordinary computer sound card and using good quality earphones. The listeners were not given the chance to practise before the test, and they did not get any feedback on their answers during the whole session. They had to give a name to the 10 signals of each questionnaire, and once it was filled, they were not allowed to go back and make any changes.

All in all 20 people took part (14 men and 6 women), with ages ranging between 10 and 53 years. All of them spoke good standard Basque, but only eleven had Basque as their mother tongue. As criteria for the analysis of the evaluation results, we have distinguished the native speakers (group A) from those who had Basque as a second language (group B).

4. Results

The results of the subjective analysis appear in Table 1. The real emotion expressed by the actor or actress is indicated on each line of the matrix; the columns, on the contrary, show the emotions defined by the listeners. The values are worked out in percentages, and each emotion is distinguished with the same letter: A for Anger, F for Fear, S for Surprise, D for Disgust, H for Happiness and X for Sadness.

The precision (P) and recall (R) parameters have also been gathered in the table. The precision measures the rate of correctness of a determined emotion (correct identifications / identified amount of emotional stimuli); the recall rate, in turn, represents the correct amount of occurrences of the emotion (correct identifications / amount of emotional stimuli).

Speakers	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: Mixed matrix of the evaluation process

Clearly, all the emotions are perceived over the chance threshold (17%), even though the corpus consisting of semantically neutral sentences might hinder the identification process. The average identification rate is 76.6%, happiness is the better perceived emotion (94%) and fear, on the contrary, the worst (64%). Sadness is the emotion with the lowest correctness rate because it is chosen when fear, disgust or anger were the stimuli. Disgust shows a low recall rate and high precision, which is understandable considering that it is chosen in few occasions but very accurately so. The opposite case is happiness, which is frequently chosen; it has a high recall rate but low precision.

Tables 2 and 3 show each speaker's mixed matrix. Happiness gets the best results in both cases (96% for the woman and 92% for the man). On the contrary, the worst recognized emotions do not coincide: fear (61%) for Karolina and disgust (59%) for Pello (but with a high precision rate). The differences between both these emotions are wide nevertheless, and they are the worst recognized ones in both databases. The average identification percentages of both emotions are also similar: 75.83% for the woman and a little higher, 76.50%, for the man.

Fear and sadness are the emotions more easily mixed up (fear shows as sadness in 34% of the cases, and the opposite misconception takes place in 20% of the cases). The most frequent misconception for both speakers occurs between the categories of fear and sadness, with mistakes fluctuating between 19% and 34%. The same type of misunderstanding was gathered in the Interfaze database for Spanish (Nogueiras and others, 2001).

Woman	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: Mixed matrix for the woman

Man	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: Mixed matrix for the man

4.1. The listeners' influence on the results

A Student's t-test was carried out in order to see if the listeners' characteristics played a part in the results of the emotion identification. Women reach an identification level of 72.78% (with a confidence level of 95% in the interval from 68.37% to 77.18%) and men of 77.62% (with a confidence level of 95% in the interval from 74.74% to 80.50%). The results seem quite significant (t=1.80, p=0.071 > 0.0) but not in the 95% confidence interval (p=0.05). The differences between the listeners with Basque as their first language (group A) and those with Basque as a second language (group B) are not significant either (t=0.858, p=0.39 > 0.05): 77.12% is the average identification level for group A and 75% for group B. Thus it seems that once the message is understood it is irrelevant whether Basque is their first or second language.

Listeners were not offered training sessions before taking the test. So we studied their level of correctness at the beginning of the evaluation and at the end of it to see any differences. During the first half of the evaluation, the identification level is 72.64% (95% confidence interval from 69.26% to 76.07%), and during the second half, it reaches 79.70% (95% confidence interval from 76.26% to 83.07%). The improvement that the results show during the second half of the evaluation in the 95% confidence interval is statistically significant (t=2.85, p=0.0044 < 0.05). During the second half, a 7% improvement is achieved in the identification level (practically stable for all groups, for the woman and for the man, for group A and for group B). This result is understandable, because the listener is made to choose an answer for the first signals even if not totally sure about them; it is a compulsory choice. As the evaluation continues, the listener learns how the speakers express each emotion, and thus identifies them more easily.

5. Conclusions

The results of the subjective evaluation show that all the recorded emotions are recognized above the chance threshold and for both speakers. Therefore, this database proves a useful resource to analyse and model emotional speech in standard Basque, and to record a system for emotional speech synthesis based on the corpus already started to be produced.

6. Acknowledgements

The evaluation of the database was possible with the assistance of the Basque Government’s ANHITZ programme (ETORTEK06/1 14) and the MEC (TEC2006-13694-C03-02/TCM).

The authors wish to thank all the listeners that took part in the evaluation.

7. References

Bulut, M.; Narayanan, S.; Syrdal, A. (2002). Expressive speech synthesis using a concatenative synthesizer, *ICSLP* 2002: 1265-1268.

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

Heuft, B.; Portele, T.; Rauth, M. (1996). Emotions in Time Domain Synthesis. *ICSLP* 1996: 1974-1977.

Hunt, A.; Black, A. (1996). Unit selection in a concatenative speech synthesis system using a large speech database. *ICASSP* 1996: 373-376.

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

Montero, J. M.; Gutiérrez-Arriola, J.; Colás, J.; Enríquez, E.; Pardo, J. M. (1999). Analysis and Modeling of Emotional Speech in Spanish. *ICPhS* 1999: 957-960.

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

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

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

Rank, E.; Pirker, H. (1998) Generating Emotional Speech with a Concatenative Synthesizer. *ICSLP* 1998: 671-674.

Saratxaga, I.; Navas, E.; Hernaez, I.; Luengo, I.; Sanchez, J. (2006). Korpusean oinarritutako sintesirako euskarazko datu base emoziodun baten diseinu eta grabaketa. *Euskalingua* 9: 173-178.

Schröder, M. (1999). Can emotions be synthesized without controlling voice quality? *Phonus* 4: 37-55.

Vroomen, J.; Collier, R.; Mozziconacci, S. J. L. (1993). Duration and Intonation in Emotional Speech. *Eurospeech* 1993: 577-580.