Designing and Recording an Audiovisual Database of Emotional Speech in Basque

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

University of the Basque Country
Alda. Urquijo s/n 48013 Bilbao (SPAIN)
eva@bips.bi.ehu.es, amaia@bips.bi.ehu.es, ikerl@bips.bi.ehu.es, ion@bips.bi.ehu.es, inma@bips.bi.ehu.es

Abstract

This paper describes an emotional speech database recorded for standard Basque. This database was recorded in the framework of a project in which the goal was to develop an avatar. therefore, the image corresponding to the expression of the different emotions was also needed. This is why an audiovisual database was developed. The designed database contains six basic emotions as well as the neutral speaking style. It consists in isolated words and sentences read by a professional dubbing actress. At present, this database is being used to study the prosodic models related with each emotion in standard Basque.

Introduction

With the progress of new technologies and the introduction of interactive systems, there has been a sudden increase in the demand of user friendly interfaces, such as avatars. For the correct development of such kind of interfaces, a high quality Text-to-Speech system is required, which may provide a more natural way of communication to the user. This naturalness can be largely improved with the expression of emotions in the synthetic speech. In addition, and focusing on the case of an avatar, it is also fundamental to have a well-trained system for generating synthetic facial expressions. To achieve these goals, a deeper research of the prosodic characteristics of emotional speech is necessary, as well as of facial expressions corresponding to each of the emotions that are intended to be modeled.

In order to carry out this research it is indispensable to have a good audiovisual database of emotional speech. The main purpose of the work discussed in this paper is the design and recording of such a database for the standard Basque.

The paper begins describing the desired characteristics for the database. Then the type of corpus selected and the corpus design made are explained. Next, speaker selection and database recording are described. Finally some conclusions are presented.

Desired characteristics for the database

Among the objectives that were taken into account during the design of this database was to collect samples of all the basic emotions, as well as of the neutral style of speech. With this objective in mind, the six basic emotions that have been considered as "the Big Six" (Cowie & Cornelius, 2003) were selected: sadness, happiness, anger, fear, surprise and disgust.

It is also necessary for the database to be large enough so that the deductions resulting from its study can be considered representative enough.

In addition, an effort was made to include the greatest phonetic variability that was possible, in order to ease the later application of the conclusions to Text-to-Speech and facial expression systems.

Finally, care was taken to collect the glottal pulse signal, synchronized with the speech, to allow a precise analysis of the intonation in each of the emotions.

Choice of corpora

Different types of corpora have been used for the study of emotions in speech:

- Corpora of spontaneous speech: They contain the most authentic emotions, but are very difficult to obtain. There are also moral considerations about privacy when recording spontaneous emotional speech. Therefore, databases of spontaneous speech are not very common. Examples of this type of corpora are the Belfast database (Douglas-Cowie et al., 2000), consisting of clips from television programs and the JST database (Campbell, 2001), with natural speech recorded in natural situations.
- Corpora of acted speech: They consist in texts read by a professional actor. This technique has been accused of recording unnatural emotions, but as the emotion intended can be recognized, they should be considered satisfactory for speech synthesis studies. Most emotional databases use this approach, because acted speech is easier to control than spontaneous speech. Examples of this type of databases are (Lay New et al., 2003), (Hozjan et al., 2002) and (Iida & Campbell, 2001).
- Corpora of elicited speech: To record these databases, the speaker is put into a situation meant to evoke a specific emotion. This method poses ethical problems, because, in order to record all the needed emotions, it is necessary to induce also negative emotions. The database recorded in the VERIVOX project uses this method (Karlsson et al., 2000).

In this work, acted speech was selected, because it is easier to control and allows an easy comparison among styles. Besides, with this type of corpus it is possible to control the content of the recording and therefore phonetic variability can be maximized.

Corpus design

There are different theories about the suitability of the database's texts to be semantically related to the expressed emotion or not. Thus, in the design of a corpus for emotional speech, considerations about the semantic content of the texts must be made.

On the one hand, the use of texts semantically related to the emotion makes easier for the speaker to express that emotion naturally. But it makes difficult to compare the characteristics of different emotions and to phonetically

balance the database. The collection of suitable texts to be recorded is also difficult. An example of emotional database containing acted speech with texts related to the emotion is (Iida et al., 2003).
On the other hand, the use of neutral texts (not related to the emotion) easies the comparison among emotions and the phonetic balance of the database, but the work of the speaker to express these emotions naturally is much more difficult. Examples of databases that use neutral texts to record emotional speech are the Danish Emotional Database (Enberg et al., 1997) and the Berlin corpus (Paeschke & Sendlmeier, 2000).
As each approach has its advantages and disadvantages, so it was decided to divide the selected texts for the database into two different groups.

- One group consists of emotion independent texts, which are common for all emotions, as well as for the neutral style. The common group of texts was phonetically balanced to achieve a phoneme distribution similar to the one that occurs in natural oral language.
- The other group includes texts semantically related to each emotion, and so, this group is different for each of the emotions considered in the database. Neutral style was not considered in this part of the corpus.

Emotion can be reliably identified in very short utterances (Enberg et al., 1997), so 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. So, both groups of texts were designed to include isolated words and sentences of different complexity and syntactical structure.
Table 1 shows the number and type of items recorded using texts not related to the emotion and table 2 shows the same data for the recordings that have texts related to the emotion.

TYPE OF ITEM	NUMBER
Isolated digits	20
Isolated words	20
Short affirmative sentences	10
Short interrogative sentences	5
Medium affirmative sentences	22
Medium interrogative sentences	8
Long sentences	10
Total number of items per emotion	95
Total number of items	665

Table 1: Items not related to the emotion

TYPE OF ITEM	NUMBER
Isolated digits	20
Short affirmative sentences	10
Short interrogative sentences	5
Medium affirmative sentences	22
Medium interrogative sentences	8
Long sentences	10
Total number of items per emotion	75
Total number of items	450

Table 2: Items related to the emotion

Vocal events

Vocal events were also taken into account during the design of the database. Some of these events appear frequently in emotional speech, so it is interesting to record them, as they emphasize and complement the emotional conversation. However, vocal events were recorded separately from the sentences, because they increase the difficulty of extracting intonation parameters, as they introduce information different from normal speech (such as extra pauses).
For the recording of the vocal events, some sentences and explanations were included in the texts, to make easier for the speaker to recognize them and set them in an appropriate context. However, those explanations were not recorded.
The events recorded include both non-lexical phenomena, such as sneezes, laughs and cries; and semi-lexical phenomena, such as voiced pauses and hesitations.

Selection of the speaker

The speaker selection is fundamental when recording an emotional speech database using acted speech, as this speaker must be capable of expressing the selected emotions with enough naturalness. Although it might be convenient to record spontaneous emotional speech, it would make the study of emotions more difficult, since there would be no way of controlling the recordings.
For these reasons, a high experienced dubbing actress was recruited for the recordings, as she has the ability of expressing the required emotion.

Database recording

The recording was made at a professional recording studio, during two days. The first day texts related with the emotion were recorded, and the second day, common sentences and vocal events. Within a recording session, every emotion was recorded without interruption, to avoid the speaker to loose concentration. The speaker was allowed to rest between the recording of texts corresponding to different emotions.

Audio recording

Figure 1 shows the audio set up used for the database acquisition. The speaker read the prompts in a monitor connected to a portable PC where the recordings were made. To get a good sound quality, a professional audio card was used. The signals were acquired with the Nanny Record software, which prompts the text to be read and controls the level of the acquired signal. A laryngograph was inserted in the recording chain to get the glottal pulse signal. The laryngograph provides three different signals, as shown in figure 2:

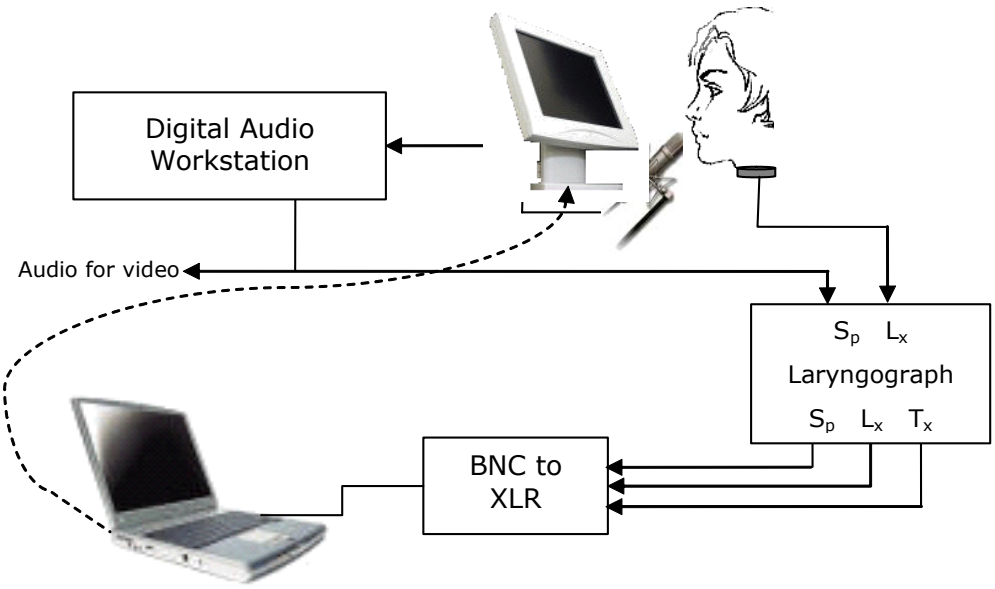

Figure 1: Audio set up for database acquisition

- Speech signal (Sp): This signal is captured with the microphone, and before it is inserted into the laryngograph, it is passed through the digital audio workstation to take it also to the video.
- Glottal pulse signal (Lx): This signal is captured by the electrodes situated around the neck of the speaker. The local minima indicate the moments when the vocal cords close.
- Quasi-rectangular 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 moments of the vocal cords.

The main problem that arose when recording the database was that the fan of the portable PC made too much noise. This noise was captured by the microphone if it was placed too close to the portable PC. To avoid this problem, some coaxial cables were used to move the PC away, but the SNR obtained was no suitable. Thus, finally the PC was covered with an isolating foam structure.

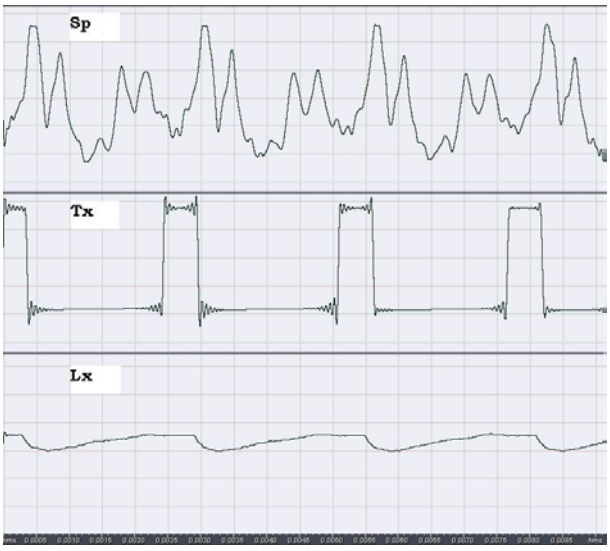

Figure 2: Signals provided by the laryngograph

The three signals provided by the laryngograph were stored in two stereo files. One contained the speech signal in one channel and the glottal pulse signal in the other. The other file contained the speech and the quasi-rectangular signal in each channel. These signals were sampled at 32 kHz, and quantified using 16 bit per sample.

The specific material used for the audio recording is listed in table 3.

Image recording

For the image acquisition two different settings were used. The first day, during the recording of the emotion specific texts, two cameras were used: a Betacam camera for the front image and a VHS camera for the side image. Unfortunately, with this setting the differences in the colorimetry of the images provided by the cameras was too high. Therefore, for the second day, recordings were made only using the Betacam camera, discarding the side image.

Microphone	TLM 103 (Neumann)[1]
Laryngograph	Laryngograph PCLX (Laryngograph LTD)[2]
Audio card	VXPocket 440, 4 channels (Digigram) [3]
PC	Pentium III
Digital Audio Workstation	Pro-Control [4]
Software	Nanny Record (UPC)[5] Digigram Wave Mixer

Table 3: Equipment used for the recording of the audio

1 http://www.neumann.com
2 http://www.laryngograph.com/pr_procs.htm
3 http://www.digigram.com
4 http://www.digidesign.com/products/protoolssystems.cfm
5 http://www.talp.upc.es

To track the movements of the interesting points of the face, several blue marks were marked in the face of the actress.

- Zone marks: They define a zone that moves when expressing emotions. They were used for the eyebrows and lips.
- Point marks: They define several points. They were used for the nose, the chin and the cheeks.

Besides these moving marks it is also necessary to record reference marks, to isolate the movement of the head. For these reference marks green color was used. With this purpose, two points were marked in the nose and five points were placed around the head, using a hairband.

When only one camera was used, more fixed marks were needed, to be able to fix the frontal movements of the head. So two more reference points were added in the forehead of the speaker.

An image of the speaker, with these marks is shown in figure 3.

The specific material used for the image recording is listed in table 4.

Conclusions

The recorded database has 1 hour and 35 minutes length. 50 minutes come from the common texts, 35 minutes from the texts semantically related with emotion and 10 minutes from the events.

This database represents a new linguistic resource that will allow the study of emotional speech in standard Basque. In fact, it is already being used for the study of the phoneme duration and its change from the neutral style, as well as for the characterization of the intonation curves related to each emotion.

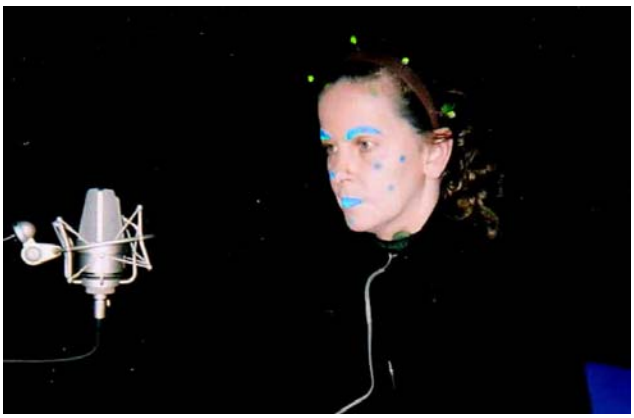

Figure 3: Speaker with marks

Camera 1	BVP 50P Betacam (Sony)
Camera 2	MS-1 S-VHS (Panasonic)
Video Mixer	Digital Production Mixer WJ-MX12 (Panasonic)
TBC	Time base corrector FOR.A FA-310P (Digital)
VTR	Digital Betacam DVW-A500P (Sony)

Table 4: Equipment used for the recording of the image

Acknowledgements

This database was developed within the project ABATEUS (code CN01BA01), with the financial help of Basque Government. It has been also partially financed by the MCyT (TIC C2000-1669-C0403).

Authors would like to thank José Ignacio Ocariz, for his work in the designing of the corpus and recording of the database.

References

Campbell, N (2001). Building a Corpus of Natural Speech – and Tools for the Processing of Expressive Speech – the JST CREST ESP Project. In Proceedings of the 7th European Conference on Speech Communication and Technology. (pp. 1525—1528). Center for Personkommunikation (CPK).

Cowie, R., Cornelius, R.R. (2003). Describing The Emotional States That Are Expressed In Speech. Speech Communication, 40(1,2), 2--32.

Douglas-Cowie, E. Cowie, R., Schröder, M. (2000): A New Emotion Database: Considerations, Sources and Scope. In Proceedings of the ISCA Workshop on Speech and Emotion. (pp. 39—44). ISCA Archive.

Enberg, I.S., Hansen, A.V., Andersen O., Dalsgaard, P. (1997). Design, Recording and Verification of a Danish Emotional Speech Database. In Proceedings of the 5th European Conference on Speech Communication and Technology (pp. 1695--1698).

Hozjan, V., Kacic. Z., Moreno, A., Bonafonte, A., Nogueiras, A. (2002). Interface Databases: Design And Collection Of A Multilingual Emotional Speech Database. In Proceeding of the 3rd Language Resources and Evaluation Conference (pp. 2024--2028).

Iida, A., Campbell, N. (2001). A Database Design For A Concatenative Speech Synthesis System For The Disabled. In Proceedings of the 4th ISCA workshop on Speech Synthesis (pp. 189--194). ISCA Archive.

Iida, A., Campbell, N., Higuchi, F., Yasumura, M. (2003). A Corpus-based Speech Synthesis System with Emotion. Speech Communication, 40(1,2), 161--187.

Karlsson, I., Banziger, T., Dankovicová, J., Johnstone, T., Lindberg, J., Melin, H., Nolan, F., Scherer K. (2000). Speaker Verification With Elicited Speaking-Styles In The Verivox Project. Speech Communication, 31(2,3), 121--129.

Lay New, T., Wei Foo, S., De Silva, L. (2003). Speech Emotion Recognition Using Hidden Markov Models. Speech Communication, 41(4), 603--623.

Paeschke, A., Sendlmeier, W.F. (2000). Prosodic characteristics of Emotional Speech; Measurements of Fundamental Frequency Movements. In Proceedings of the ISCA Workshop on Speech and Emotion. (pp. 75—80). ISCA Archive.