Design and recording of a corpus-based synthesis emotions database in Basque

Ibon Saratxaga, Eva Navas, Inmaculada Hernáez, Jon Sanchez

University of the Basque Country
ibon.saratxaga@ehu.es, eva.navas@ehu.es, inma.hernaez@ehu.es, jonsanchez@ehu.es

Abstract

This paper describes an emotional speech database recorded for standard Basque. The database has been designed with a twofold purpose – to allow it to be used for corpus-based synthesis, and also to allow research into prosodic models for emotions. Thus the database is large for good corpus-based synthesis quality, and contains the same texts recorded in the six basic emotions, plus the neutral style. The recordings were made by two professional dubbing actors, a man and a woman. The paper explains the whole creation process, beginning with the design stage, following with creation of the corpus and the recording phase, and finally some of the lessons learned.

Keywords: speech database, emotions, corpus-based synthesis

1. Introduction

In recent years, following a wealth of progress in voice synthesis, the milestone of intelligibility has been surpassed, and research has focused on naturalness. These characteristics are becoming increasingly necessary as the applications of synthesis expand and become much more complex: natural sound, fluency and intonation are essential factors if a long synthesised text is to be understood.

In the search for naturalness, corpus-based synthesis methods (or methods based on choice of units) appeared towards the end of the last decade. These methods use the linking of units to create synthesis, and attempt to minimise manipulation of the signal. In this way the original natural voice is retained, reducing the linkages between voice units and using algorithms which compensate long units (Sagisaka, 1998; Sagisaka et al., 1992).

Several fields of research have also been examined in an attempt to improve fluency and intonation, particularly adhering the characteristics of emotions to the synthetic voice. Our research group carried out a number of projects on the characterisation of emotions in spoken Basque, and we needed a new database, extensive and purpose-designed, to continue our investigation.

The main objective of the study described in this paper is the design and recording of a database, since this will lead to emotional corpus-based synthesis in standard Basque.

The paper commences with a description of the characteristics of the database, first specifying the requirements for modelling the emotions, subsequently the aims of synthesis based on the choice of units, and then providing an explanation of the design of the corpus and its characteristics. The paper moves on to describe the recording phase, the process of selection of the speakers, characteristics of the equipment used, and details of the recording sessions, and finally a short summary is presented of the conclusions.

2. Requirements of emotion modelling

Investigation of prosodic emotion models requires examples or samples of all the basic emotions. Our research has taken account of those known as "The Big Six" (Cowie & Cornelius, 2003) – sadness, happiness, anger, surprise, fear and disgust – and we have also used the neutral style.

Various different types of corpuses have been used to study emotion within oral speech. Some research units have used spontaneous or natural recordings in order to achieve the greatest degree of veracity in emotions. Others have worked on the effects of the emotion, putting speakers in a specific situation in order to produce a specific emotion. Another possibility is to use a speaker with an actor's capacity to simulate emotions. This can also inflate the emotions – it is considered that recognisable emotions are obtained, and so models can be produced using these techniques. We decided to use represented emotions for this projects, since it gives us full control with respect to the contents of the emotions.

Using different texts for each emotion creates great difficulties in the search for suitable corpuses and, far more important, it presents a considerable obstacle for drawing comparisons between the characteristics of the emotions. The decision was therefore taken not to use special texts, and the same text independent of emotions was used for all the emotions. This decision may be explained by the fact that previous studies have shown that a happy speaker is able to express emotions in a natural manner, if the text does not have any relation to the emotion (Navas et al., 2005).

2.1. Controlling the speaker's changeability

Previous surveys have also shown us it is essential that the pace, tone, volume etc. of the speaker's voice must be maintained with no changes over a long recording session. The recording of this database had to last over several sessions, and we expected that the effect of these changes would be considerable. In order to quantify these deviations, and maintain capacity to draw comparisons between the prosodic parameters of the emotions, a simultaneous control text was also designed.

The control consisted of a short continuous text of approximately 400 words, to be read in a neutral fashion at the beginning of the session, in the middle and at the end. In this way, several reference levels can be extracted for the prosodic parameters in each control text session, and data for each emotion can be standardised using the reference levels.

2.2. Supra-speech level prosody

The last requirement of prosodic research in relation to design of the database was the need to study the characteristics of prosody above the level of speech. As we will see in the next section, the largest section of the recording corpus was to be composed of a single sentence. This is extremely important to produce prosodic models at sentence level, and is also advisable to allow techniques of choice of units. However, different forms of speech must be studied – conversations, statements, pauses between paragraphs etc. To allow us to study all these characteristics, it is essential to have recordings of continuous texts.

In these circumstances, another feature was added to the design of the database – a continuous text of medium length (1,047 words), similar to the style of the speech. This text was to be read in the six basic emotions and in the neutral style and recorded immediately, to allow us to study the pauses between paragraphs.

2.3. General structure of the database

In due consideration of the above requirements, the components of the database were defined as in Table 1 below:

Section	Number of recordings	Contents
Main Corpus	One for each emotion + Neutral	Isolated sentences
Continuous Text	One for each emotion + Neutral	Continuous text with similar style of speech
Control Text	Three times for each session	Continuous text

Table 1: General structure of the database

3. Requirements of the techniques for selection of units

We revised the requirements of the prosodic examinations in the previous section, and these requirements defined the structure or sections of the database. In this section we will observe how the techniques for selection of units determine the contents of the database's text contents, or inner structure.

As we have already mentioned, the techniques for selection of units require large databases to make the selection algorithms suitable for the group of units to be chosen. To allow us to design a corpus for this system, the main objective is to have the largest number of possibilities for each unit: the database coverage must be quite wide.

With this objective, the section of the database to be used is the Main Corpus, and so the next requirements will affect only this section.

3.1. Size of the data base

The size of the database must be established with great care, in order to locate the longest possible units for synthesis. A suitable size starts at one hour (Febrer, 2001). This means almost 40,000 diphonemes, around 6,400 words (assuming an average of 6.3 diphonemes per word in Basque) or about 500 sentences.

These figures set the lower limit for the size of the database. A number of other factors will have an effect on the final size. The larger the database, the better the results of synthesis will be, although there are restrictions in relation to performance and consumption of resources, and also economic restrictions which bring the upper and lower levels closer together.

3.2. Phonetic balance

In addition to ensuring that there are long units in the database, it is necessary to ensure all possible phonemes, and certain combinations between phonemes, are in the database. If we wish to ensure that the quality of synthesis based on selection of units is at least similar to that of other linkage methods, the design of the database must guarantee that the smallest units used by other methods are in the database. The reasonable minimum-size unit here is the diphoneme.

After the minimum unit has been established, the objective of phonetic balance is to maintain as equally as possible the frequency with which these units appear in the database, and the frequency with which they appear in natural speech. This means that the most common phonemes will appear with great frequency in the database recorded, in similar contexts, whilst rarer phonemes will appear only once, or will have to be explicitly added at the design stage. Moreover, there are "difficult" combinations of three or four phonemes, which create problems when they are separated, and to avoid these we define some longer units, known as "polyphonemes". 406 such polyphoneme units have been defined in Basque.

3.3. Lexical balance

In this corpus-based synthesis, when units are chosen it is possible to find larger sections than diphonemes, even complete words, as in this way the number of linkages is reduced. Obviously the size of the database to ensure a coverage rate at word level would be much larger than the database we are considering. Thus the requirement for a smaller size and the requirement of phonetic coverage are more important than lexical coverage. Nevertheless, it is desirable that the most common groups of words in the language appear in the database, making the number of different words as large as possible at the same time.

Moreover, when "rare" phoneme combinations appear, normally they are found in alien words. Thus, if priority is accorded in the design of the database to combinations of all diphonemes, the corpus is sure to increase its volume of alien words, taking the place of normal words in the language.

As mentioned above, the lexical criterion is not the decisive factor in design of the corpus, but we should attempt to maximise the number of different words, with particular care in regard to the inclusion of common words. Likewise, we will also try to keep the percentage of foreign words below a reasonable percentage. In any case, the final corpus produced will be analysed lexically and adjusted.

3.4. Field of synthesis and type of vocabulary

Finally, there is another aspect to be taken into consideration in relation to design of the corpus for a speech database, and this concerns definition of the field or area of application of the synthesiser. In other words, we must know what the synthesiser (Test to Speech or TTS) will say in most cases. As may be observed, if we already know which words the TTS will say, we may add these words to the design of the database, or at least extract the corpus of the database from larger corpuses in that field, presumably producing some good results.

In this case, there is no intention to use the synthesiser in a special field – the aim is to create a good general-use synthesiser. Unfortunately, this means the field of vocabulary is undefined and therefore quite wide.

In any case, we intend to use the synthesiser to read electronic books and newspapers, and so we will take the initial large corpuses from these sources.

3.5. Internal requirements for the Main Corpus

The requirements of the technique for selection of units are summarised in Table 2.

Corpus size	Over 40,000 diphonemes, or 6,400 words (approximately 500 sentences)
Phonetic coverage	At least one hit of all the diphonemes found in the initial corpus. Total coverage of the 406 polyphonemes already defined
Lexical coverage	50% of the most common words in the language in the corpus

Table 2: Design requirements for the Main Corpus

4. Creation of the corpus

After we have set the initial requirements, the next stage in the process is to create the corpus to be recorded. A number of requirements will be defined in accordance with the coverage ratios of several pre-existing corpuses: in other words, this is the first step which must be taken to secure the largest possible text corpus from which the recording corpus will be extracted.

In this project, the initial corpus is a large group of texts produced by a number of sources: the largest section consists of texts from the *Egunkaria* newspaper over two years, other texts are from a number of novels, and other smaller texts have also been used, particularly those used in other AhoLab projects, and so these are extremely pure and phonetically balanced.

All these corpuses make up what is known as the Basic Corpus. It has 580,000 sentences, 7.4 million words, 243,800 of which are different (declined words are taken as different words in this count – the number of different entries for words is smaller), or 46 million phonemes. One of the interesting features of the data is the number of different diphonemes – 897 – and we must ensure coverage of these in the recording corpus.

Once the initial corpus had been created and analysed, the next step was to extract the recording corpus. The process was carried out using a UPC (Universidad Politécnica de Catalunya, http://www.talp.upc.es) computer programme known as CorpusCrt. This tool extracts a small group of sentences from a larger corpus, and these sentences are chosen to maintain a relative frequency of diphonemes in the large original corpus, and in the smaller corpus. The larger size of the Basic Corpus meant that the process had to be carried out in two steps, and in this way it was possible to section off a complementary corpus of a more manageable size from the original large corpus. This eventually produced the Main Corpus (the largest section of the recording corpus).

4.1. Manual adjustments and validation of the corpus

The abovementioned software produced a number of sentences with diphoneme balance. The sentences were compared with the list of diphonemes (since polyphonemes are composed of more than two phonemes, the CorpusCrt programme does not take this into account when choosing sentences). A search was conducted for the missing polyphonemes in the initial corpus, and the new sentences were thus added to the Main Corpus. A number of polyphonemes did not appear in the Basic Corpus, and any special sentences it may have had were added.

The resultant corpus and its phonetic transcription were completely revised to remove any punctuation or transcription errors. Table 3 shows the characteristics of the corpus produced.

Number of sentences	702
Total number of words	6,582
Number of different words	4,308
Total number of phonemes	39,767
Number of different phonemes	35
Total number of diphonemes	40,917
Number of different diphonemes	897
Coverage of polyphonemes	100% (406)
Estimated duration of recording	80 minutes

Table 3: Statistical data for the Main Corpus

4.2. Lexical balance

As mentioned above, less priority was accorded to lexical balance than to other requirements. Analysis of lexical coverage was carried out during the phase of adjustments to the database, with the intention of making minor changes for the sake of improvement. The lexical data in the final recording corpus were compared to data in the Basic Corpus, and the results were as follows:

The 4,300 different words in the Main Corpus accounted for 56.3% of all words in the Basic Corpus. In any Basque text, the Main Corpus would allow the synthesiser to pick up word-level units in 56.3% of words.

In a more qualitative analysis, 695 different words make up 50% of all words in the Basic Corpus. Of these words, 570 are also found in the Main Corpus. If we take the 1,000 most common words in Euskara, 73.6% are found in the Main Corpus. These numbers ensure that coverage is sufficient for our application.

Finally, a survey was made of the number of alien words – their frequency would definitely have increased since these words have rare diphonemes. A rough analysis produced 9.6% of alien words, and this result was considered acceptable.

4.3. Control text and continuous text

The last task for final definition of the corpus was the choice of the control text and the continuous texts. The requirements for the control text were extremely flexible, and so a literary description text was chosen.

The prosodic model continuous text had to be longer and similar in terms of style of vocabulary. A section of monologue was chosen which included a descriptive passage and conversations.

5. Recording phase

The recording phase began when the full recording corpus had been completed. The first task was to select the two speakers, a man and a woman. A small casting was arranged with two objectives: the speaker had to be able to represent emotions, and voice quality had to be suitable for synthesis.

Recordings were produced with professional speakers, and tested with Praat software (Boersma & Weenink, 2005). The Praat tool provides easy and flexible resynthesis, using the PSOLA and LPC methods. In this way the original recordings were manipulated in order to predict the suitability of their synthesis. Finally, two professionals were hired for the project: a 40-year old male dubbing actor, and a 37-year old woman, a radio presenter and actress.

5.1. Recording platform and environment

Recordings were carried out in a semi-professional recording studio. Six sessions were required for the woman's voice, and a further four sessions for the man's voice. The recordings were made from emotion to emotion, and recording did not stop between each emotion so that the speaker would not lose concentration.

The recording platform is shown in Figure 1. Recordings were made using a portable PC with a professional audio card. Two voice signals were taken – one was captured using a large studio microphone (S_m), and another from an elektret proximity microphone (S_c). A laryngograph was also used to pick up glottal pulse. The proximity signal and a pair of glottal electrodes go into the laryngograph, and this produces three outgoing signals: a voice signal picked up on the proximity microphone (S_c), the glottal pulse signal (L_x), and a quasi-rectangular signal (T_x) produced by the glottal pulse.

The three signals produced by the studio microphone and the laryngograph were recorded in dual-signal stereo – one in S_c and L_x, and another in S_m and T_x. The signals were sampled at 48 kHz frequency, and quantified using 16-bit sample resolution. The signal pick-up equipment used was Nanny Record. The characteristics of the unit used are shown in Table 4.

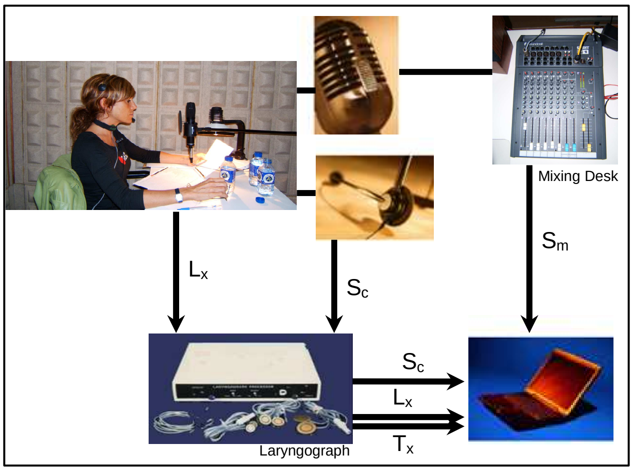

Figure 1: Recording platform

Microphones	BeyerDynamic MC740 (studio mike) Emkay VR-3576 (proximity mike)
Mixing desk	Soundcraft Spirit F1
Laryngograph	PCLX laryngograph (LTD laryngograph)
Audio card	VX Pocket 440 (Digigram)
Software	Nanny Record (UPC) Digigram Wave Mixer

Table 4: Recording equipment

6. Conclusions

The data base contains approximately 1.5 hours of recording for each emotion. A total of 10.5 hours are added for each speaker, making more than 20 hours overall. This database represents a new linguistic resource, which will prove useful in terms of research into the emotional vocabulary of standard Basque and high-quality synthesis based on choice of units. The large dimensions of the database will be useful for other research projects in a number of areas – for example, voice transformation, corpus-based prosody etc.

The process has taught us a number of lessons to be taken into account when recording a large database. Firstly, we have found that problems arise when decisions taken in relation to design of the database come up against the speaker's ability to imitate. The speaker found it hard to pronounce uncommon alien words, and on many occasions the presence of such words affected the intonation of the entire sentence. Similar problems were encountered when long or syntactically complicated sentences appeared. These situations can adulterate the prosodic models, and so an attempt should be made to reduce the number of such sentences. Creating another special group of sentences, which would not be used in the prosodic model, could be a suitable solution to this problem.

Another valuable lesson learned was the need for a specific casting test to ascertain the ability of the speaker to represent the emotions. It might be a good idea to arrange an informal blind test on the knowledge of emotions, since it would seem that imitating emotions such as fear or disgust while reading a long complex sentence out of context requires considerable dexterity.

In this sense, another point in relation to potential problems is the need to maintain coherence in terms of pronunciation and intonation. This kind of emotion coherence is desirable, because the example thus attained will be high quality, although it can work against the naturalness of the emotion. With opposing requirements, in the databases chosen by units, the most successful results are obtained with smooth uniform intonation. These opposing obligations must be made extremely clear to the speaker, and the recording technician must pay close attention to ensure whether these instructions are carried out.

7. Acknowledgements

In our production of this database we received assistance from the Basque Government's SAIOTEK (SPE04UN24) programme, and also from MEC (TIC2003-08382-C0503).

The authors wish to thank the University of the Basque Country for allowing them to use its recording studio.

8. Reference material

Boersma, P., Weenink, D. (2005). *Praat: doing phonetics by computer (Version 4.3.16)* [Computer program]. http://www.praat.org/

Cowie, R., Cornelius, R.R. (2003). *Describing the Emotional States that are Expressed in Speech.* Speech Communication, 40 (1, 2), 2 – 32.

Febrer, A. (2001). *Síntesi de la parla per concatenació basada en la selecció* (Speech synthesis by selection-based concatenation). PhD Thesis. p. 48.

Navas, E., Hernaez, I., Luengo, I., Sanchez, J., Saratxaga, I. (2005). *Analysis of the Suitability of Common Corpora for Emotional Speech Modeling in Standard Basque.* LNCS 3658, pp. 265 – 272.

Sagisaka, Y. (1998). *Speech synthesis by rules using an optimal selection of non-uniform synthesis system.* International Conference on Acoustics, Speech and Signal Processing, pp. 679 – 682.

Sagisaka, Y., Kaiki, N., Iwahashi, N. and Mimura, K. (1992). *ATR – vTALK speech synthesis system.* International Conference on Spoken Language Processing, p. 483.