Designing and Recording an Emotional Speech Database for Corpus Based Synthesis in Basque

Ibon Saratxaga, Eva Navas, Inmaculada Hernáez, Iker Luengo

Aholab - Dept. of Electronics and Telecommunications. Faculty of Engineering. University of the Basque Country
Urkijo zum. z/g 48013 Bilbo
ibon@bips.bi.ehu.es, eva@bips.bi.ehu.es, inma@bips.bi.ehu.es, ikerl@bips.bi.ehu.es

Abstract

This paper describes an emotional speech database recorded for standard Basque. The database has been designed with the twofold purpose of being used for corpus based synthesis, and also of allowing the study of prosodic models for the emotions. The database is thus large, to get good corpus based synthesis quality and contains the same texts recorded in the six basic emotions plus the neutral style. The recordings were carried out by two professional dubbing actors, a man and a woman. The paper explains the whole creation process, beginning with the design stage, following with the corpus creation and the recording phases, and finishing with some learned lessons and hints.

1. Introduction

In the last years, progress in speech synthesis has largely overcome the milestone of intelligibility, driving the research efforts to the area of naturalness and fluency. These features become more and more necessary as the synthesis tasks get larger and more complex: natural sound and good fluency and intonation are mandatory if a long synthesized text shall be understood.

Seeking naturalness, the corpus based (or unit selection based) synthesis methods appeared about the second half of the last decade. These methods use concatenative speech synthesis techniques and try to minimize the signal manipulation. In this way they preserve the original naturalness of the speech, minimizing the number of joints between voice fragments, by using unit selection algorithms which bonus large units (Sagisaka,1998; Sagisaka et al., 1992).

On the other hand, the goal of improved fluency and intonation has led to a lot of research fields. One of the most important of them is the inclusion of emotional features into synthetic speech. Currently our research group is studying the characterization of Basque emotional speech. Moving forward in this research requires a large and specially designed database.

The main purpose of the work discussed in this paper is the design and recording of a speech database which will allow emotional corpus based synthesis and the definition of the prosodic models of emotions for standard Basque.

This paper begins describing the desired characteristics of the database, first detailing the requirements posed by the emotion modelling work, and after, those brought up by the unit selection synthesis objective. Then, the process of the corpus design and its final features are described. Next, the recording phase is explained covering the speaker selection process, the employed equipment and recording sessions' details. Finally, some conclusions are presented.

2. Requirements for emotions modelling

The study of the prosodic models of the emotions requires recording samples of all the basic emotions. We have considered the set of emotions known as "the Big Six" (Cowie & Cornelius, 2003): sadness, happiness, anger, fear, surprise and disgust. Additionally, neutral style has also been considered.

Different types of corpora have been used for the study of emotions in speech. Some groups have employed spontaneous emotional speech, trying to get the greatest authenticity in the emotions. Others have worked with elicited emotions, putting the speaker into situations to rouse a specific emotion. A third option has been to use a speaker with acting skills to simulate emotions. Though this latter technique can exaggerate the emotions, the fact is that they are recognized, so that practical modelling can be derived from them. In this work we selected acted speech corpus because it is the only one that allows a complete control over the recorded text.

Since using different texts for every emotion implies big difficulties to find corpora containing suitable texts, and, furthermore, it hinders from comparing the characteristics of the different emotions, it was decided not to use this kind of texts, recording the same emotion independent text for all the emotions. Besides, previous works showed that a skilled speaker could express emotions naturally even if text content was not related with emotion (Navas et al., 2005).

2.1. Controlling speakers' variability

Previous works have also taught us that it was impossible for the speakers to keep a constant reference level for their rhythm, tone, volume, etc. through a long lasting recording session. The expected recording time for this database spread through several sessions, so the effects of these variations were supposed to be even more important. In order to quantify these deviations and keep on being able of comparing prosodic parameters among emotions, a control text was also designed.

This control text consists of a short continuous text (400 words long), and had to be read with neutral style at the beginning, mid-session and end of every session. In this way, the reference levels in the prosodic parameters for each session will be extracted from this control text, and the data of every emotion will be normalized against these reference levels.

2.2. Supra-sentence level prosody

The last factor relating to the prosodic studies which impacted in the database design requirements, was the

need for studying the prosody of speech above the sentence level. As it will be shown in the next section, the main acted speech text part was going to be composed of isolated sentences, which is very convenient for extracting prosodic models at sentence level (except for pause model), as well as serving to the other purposes of unit selection enablement. However, it is also necessary to study different speech styles such as dialogs, declarative parts, paragraph pausing, etc. which require continuous text.

That is why another piece was added to the database design: a medium-sized continuous text (1,047 words long), which covered different speech styles like dialogues and descriptions. This text was intended to be read in the six emotions plus neutral styles, each of them recorded in one go.

2.3. Database overall structure

Taking into account all the above considerations the overall structure of the database was defined as shown in Table 1.

Section	No. of Recordings	Contents
Main Corpus	One per emotion + Neutral style	Isolated sentences
Continuous Text	One per emotion + Neutral style	Single piece of continuous text with varied speech styles
Control Text	Three times per recording session	Single piece of continuous text

Table 1. Overall structure of the database.

3. Requirements for unit selection techniques

In the previous section we have seen that the prosodic study has posed requirements which affect to what we could call the "external" structure of the database. In this section we will see that the unit selection synthesis objective will set the requirements for the actual text contents of the database: the "internal" structure.

As said before, the unit selection techniques need large databases to provide the selection algorithm with a good choice of candidate units. The main objective of the corpus design for these systems is to ensure that there are candidate units for the biggest number of possibilities, that is, database coverage is broad enough.

The part of the database that will be used for this purpose is the one called Main Corpus, so the following requirements will only affect to the contents of this part.

3.1. Database size

The size of the database has to be carefully fixed in order to assure that, at synthesis time, units as large as possible are found. Appropriate size starts from 1 hour of recordings (Febrer, 2001). This means approximately 40,000 diphonemes, which translated into Basque words (with an average of 6.3 diphonemes per word) yields to some 6,400 words, or 500 phrases.

These figures establish the bottom limit in the database size. The final size will be influenced by the other requirements. Obviously, the bigger the database is, the better the synthesis results could be, but there are performance, resources consumption and even speaker availability constraints that set the upper limit not very far away from the minimum one.

3.2. Phonetic balance

Besides assuring that large units will be found in the database, it is also necessary to assure that all the possible phonemes and certain phoneme combinations of a language are included. If we want to assure that the unit selection synthesis will produce at least the quality of other concatenative methods, we will have to design the database to guarantee that there are at least all the smallest units used by these other methods. A reasonable minimum size for these units is diphoneme.

Once the minimum unit is selected, the purpose of the phonetic balance is to keep the appearance rate of these units in the database corpus, as close as possible to their appearance in the actual language. In this way, usual diphonemes will appear lot of times in the recorded database, in multiple contexts, and rare ones will appear perhaps only once, or even they will have to be explicitly added. In addition, there are some problematic combinations of three or four phonemes, so some units longer than diphonemes, called "poliphonemes" have to be considered. 406 of these poliphonemes have been defined for Basque.

3.3. Lexical balance

In a corpus based synthesis, it is possible that, when units are sought for the synthesis not only diphonemes are found but also bigger pieces, even complete words, so that the number of concatenations in the synthetic speech is minimized. Clearly, the volume of a database which will assure a certain level of coverage at word level for a language will be much higher than the one we are considering. In this sense, the size requirement and the phonetic coverage are more limiting than the possible lexical coverage. Nevertheless, it is desirable that the group of most used words of a language appears in the recorded database, trying also to maximize the number of different words which will be recorded.

On the other hand, when "weird" combinations of phonemes appear, they usually come from foreign words. So, if in the design of the database priority is given to the appearing of all the possible diphoneme combinations, it will very likely imply that the number of foreign words in the corpus increases, taking the place of language's original words.

As we mentioned, lexical criterion is not the determining one for the corpus design, but we should try to maximize the number of distinct words, paying special attention to the most habitual ones. At the same time, we will attempt to keep the number of foreign words under a reasonable level. In any case, final corpus will be lexically analysed and tuned.

3.4. Synthesis domain and type of vocabulary

One final aspect to be taken into account when designing a corpus for a speech database is to define the scope of the application of the Text To Speech (TTS) system, i.e. what the TTS system will more likely say. It seems obvious that if we know beforehand what are the words the TTS system will probably produce, its domain, we could include these words in the database; or at least, we could extract the corpus for the database from larger

corpora of the domain, and the results will probably be very good.

In the case of this work, there is not a planned specific use for the synthesiser; the aim is to create a high quality synthesiser for general use. Unfortunately this means undefined, and therefore very wide, domain.

However, the resulting synthesiser is intended to be able to read newspapers or electronic books, so initial corpora will be taken from these sources.

3.5. Internal requirements of the Main Corpus

To sum up, the requirements brought up by the unit selection synthesis use of the database are detailed in Table 2.

Corpus Size	More than 40,000 diphonemes, or 6,400 words (approx. 500 sentences)
Phonetic coverage	At least one appearance of every detected phoneme in the starting corpora. Total coverage of predefined 406 poliphonemes.
Lexical coverage	Trying to cover the 50% of the most common words of the language.

Table 2. Internal requirements for the main corpus.

4. Creation of the corpus

Once the initial requirements have been stated, the next step in the process is to create the actual corpus to be recorded. Some of these requirements set coverage referred to a pre-existing corpus of the language: and this is actually the first step, getting a great amount of corpora from which the final recording corpus will be extracted.

In this work the initial corpora is a set of texts coming from different sources: the main portion consists of two years of text from a Basque newspaper, other texts come from several novels, and a number of smaller corpora, previously obtained for other works in Aholab, depurated and balanced to get phonetic coverage.

All these corpora together make the Base Corpus, which contains over 580,000 sentences, 7.4 million words (which adds up 243,800 different words[1]), or 46 million phonemes. A more interesting datum is the number of different diphonemes, 897, which is the actual coverage we are required to assure for the recording corpus.

Once the initial corpora were collected and analysed, the next step was to extract the desired corpus from them. The process is carried out using a software tool called CorpusCrt from the UPC[2], which produces a reduced set of sentences keeping the original frequency of the diphonemes as far as it is possible. The huge volume of the Base Corpus required a two step process in order to reduce the initial volume to a more manageable size from which the Main Corpus was extracted.

4.1. Corpus validation and manual tuning

The above mentioned tool produced a reduced set of diphoneme balanced sentences. This set of sentences was checked against the list of poliphonemes (which are composed of more than two phonemes, and thus are not taken into account by CorpusCrt when selecting the sentences). Lacking poliphonemes were looked after in the initial corpus and added if available, some of them had to be included in deliberately created sentences or words.

The resulting corpus and its phonetic transcription were thoroughly revised to correct punctuation and spelling mistakes. Table 3 shows its main features.

Number of sentences	702
Total number of words	6,582
Number of different words	4,308
Total number of phonemes	39,767
Number of distinct phonemes	35
Total number of diphonemes	40,917
Number of distinct diphonemes	897
Poliphoneme coverage	100% (406)
Estimated recording length	80 min

Table 3. Main Corpus statistics.

4.2. Lexical balance

As commented before, the priority of the lexical balance requirements was smaller than the priority of the other requirements. During the tuning stage of the database creation, lexical coverage analysis was done in order to make slight changes to improve it. Final corpus lexical data was compared against the lexical data of the Base Corpus with the following results:

The 4,300 distinct words in the Main Corpus correspond to the 56.3% of the total volume of words of the Base Corpus. Taking any Basque text, Main Corpus should allow taking units at word level for the 56.3% of the words that appear in it.

In a qualitative analysis, the 50% of the words in the Base Corpus consists on 695 different words. From them 570 are included in the main corpus (82%). If we take the 1,000 most common words in Basque, the 73.6% of them are included in the Main Corpus. These figures assure a coverage good enough for the purpose we want.

Finally, the amount of foreign words was studied, which rate was very likely to be increased as these words include the rare diphonemes. A rough study produced a result of 9.6% of foreign words in the corpus, which was considered as acceptable.

4.3. Control and Continuous Text

The last task to finish the corpus definition was to choose the Control and the Continuous Texts. The requirements for Control Text were very loose, and thus a literary descriptive fragment was chosen.

The Continuous Text for prosody modelling had to be larger and should include different speaking styles. A fragment of a monologue was selected including descriptive and dialog style speech.

5. Recording phase

With the whole recording corpus defined, the recording phase started. The very first task to do was the selection of the two speakers we needed, a man and a woman. A limited casting was done in order to achieve two goals: the speaker had to be able of expressing the selected emotions, and also the speakers´ voices should produce good results when synthesising them.

Records from different professional speakers were listened and tested by means of Praat (Boersma & Weenik, 2005), a software tool that allows a very quick resynthesis using PSOLA or LPC based methods. With

[1] This number considers inflected words as different ones; the real number of different meaning words is smaller.

[2] Universidad Politécnica de Catalunya. http://www.talp.upc.es

the help of this tool original voices were manipulated trying to foresee their suitability to be used for synthesis. Finally two professional speakers were recruited: a 40 years old dubbing actor male, and a 37 years old radio speaker and actress female.

5.1. Recording environment and platform

The recording was made at a semi-professional recording studio, during 6 sessions for the female voice and 4 sessions for the male one. Recordings were made emotion by emotion, recording every emotion without interruption, to avoid the speaker loosing his/her concentration.

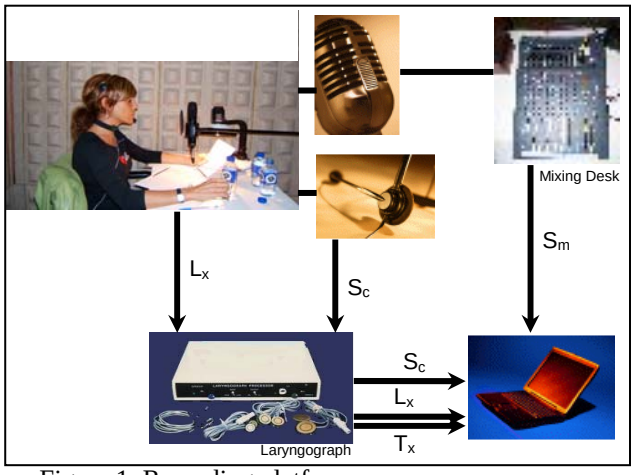

Figure 1. Recording platform.

Recording platform is shown in Figure 1. The recordings were made by means of a portable PC with a professional audio card. Two speech signals were obtained, one coming from a large studio membrane microphone (S_m), and the other from a close-talk electret microphone (S_c). A laryngograph was also used to get the glottal pulse signal. Close-talk signal and a pair of glottal electrodes feed the laryngograph, which produces three outputs: the voice signal from the close-talk microphone (S_c), the glotal pulse signal (L_x), and a quasi-rectangular signal (T_x) derived from the glotal pulse.

The large microphone signal and the three signals coming from the laryngograph were recorded in two stereo signals, one containing S_c and L_x, and the other S_m and T_x. The signals were sampled at 48 kHz, and quantified using 16 bits per sample. The acquisition software was Nanny Record. The equipment used for the recording is shown in Table 4.

Microphones	BeyerDynamic MC740 (Membrane) Emkay VR-3576 (Close-talk)
Mixing Desk	Soundcraft Spirit F1
Laryngograph	Laryngograph PCLX (Laryngograph LTD)
Audio Card	VX Pocket 440 (Digigram)
Software	Nanny Record (UPC) Digigram Wave Mixer

Table 4. Recording platform.

6. Conclusions

The recorded database consists on approximately 1.5 hours per emotion which makes up 10.5 hours of recordings per speaker, more than 20 hours in total. This database represents a new linguistic resource that will allow the study of emotional speech in standard Basque, and also a high quality unit based synthesis. The large extent of the database will also enable future research on other areas, like speech modification, corpus based prosody and so on.

The process has taught us several lessons to be taken into account when recording a large database. First, some issues arise when the database design decisions are confronted with the speakers' ability to perform the texts. We found that unfamiliar foreign words were quite hard for the speaker to utter, and very often the presence of such words affected the overall intonation of the whole sentence. Similar difficulties appeared with long or syntactically complex sentences. These cases can degrade the prosodic models, and thus, the number of sentences of this type has to be kept low. Creating a separate set of sentences with foreign diphonemes not to be used in prosodic modelling can also be a good approach.

Another important lesson is the requirement of a specific casting test to check out the speaker skills for emotion interpretation. It would indeed be good to do an informal emotion recognition blind test with short recordings from the candidates, because acting emotions like disgust or fear with unrelated, sometimes long and complex, sentences requires certain theatrical capacity.

In this sense, another problematic point is the requirement of consistency in the pronunciation and intonation of the performed speech. Consistency is desirable in order to achieve a good modelling of every emotion, but it can go against the naturalness of the emotion. Quite the opposite, for unit selection databases best results are got with uniform and plain intonations. All these often contradictory indications have to be clearly communicated to the speaker and their fulfilment controlled by the recording technician.

7. Acknowledgements

This database was developed with the financial help of the Basque Government within the SAIOTEK program (SPE04UN24) and of the MEC (TIC2003-08382-C0503).

Authors would also like to thank the University of the Basque Country for allowing using its recording studio.

8. References

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.* PhD Thesis. pp 48

Navas, E., Hernaez, I., Luengo, I., Sanchez, J., Saratxaga, I. (2005). *Analysis of the Suitability of Common Corpora for Emotional Speech Modelling 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, pp 483.