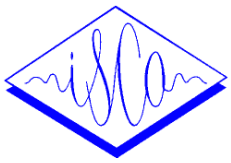

The ESCA Workshop
on Speech Synthesis
Autrans, France
September 25-28, 1990

DESIGN AND GENERATION OF THE ACOUSTIC DATABASE OF A TEXT-TO-SPEECH SYNTHESIZER FOR SPANISH

Alejandro Macarron Larumbe
*Telefonica I+D. Madrid, SPAIN**

ABSTRACT

This paper describes the design and generation of the acoustic database for a Spanish text-to speech synthesizer, developed jointly by AT&T BELL LABORATORIES and TELEFONICA I+D, based on the concatenation of LPC coded units. These units can be referred to as "paraphones" since they can be subphones, diphones, triphones or even polyphones of a greater order. For our work, we first established a list of the allophones required; then constructed a dictionary of paraphones; devised a list of sentences containing all this paraphones; recorded all these sentences and finally segmented from them the units for our database.

1. INTRODUCTION

For the construction of the acoustic database of a text-to-speech converter based on concatenation of parametrically recorded and stored units (J.P. Olive and M.Y. Liberman, [1], [2]), several stages have to be accomplished:

1) A list of all the allophones of the language. We designed the list according to the literature and to the method of synthesis, considerating economy in storage and generation time versus quality of the synthesized speech.

2) A list of all the paraphones for the acoustic inventory. This list is dependent on the allophone set used, and has to be established with similar considerations of the quality of speech versus economy in generation time and memory.

3) A list of sentences containing at least one realization for each of the paraphones, embedded within a context that doesn't cause too heavy coarticulation.

4) Recording of the sentences and segmentation of the units for storage in the appropiate database.

2. THE LIST OF ALLOPHONES

What is the number and identity of the allophones of a given language? The answer depends on how strictly one wants to apply the concept of allophone: any of the

---

* On temporary assignment at AT&T Bell Laboratories, Murray Hill, NJ, USA

different acoustic realizations of a given phoneme. Due to coarticulatory effects of the preceding and following phonemes, any particular phoneme cannot be considered as having the same acoustic properies regardless of its "neighbors". But, to what extent is a given spoken phoneme different from other realizations of the same phoneme depending on the context? The answer to this question must depend on the method used for measuring such difference, and moreover, on what degree of accuracy do we need for our purpose when discriminating the allophones. In speech recognition based in subword units, an allophone can be any phoneme surrounded by any allowed pair of phonemes. Applying strictly such criterion, we might have many thousands of allophones, that for real implementations of speech recognition, or for text-to-speech, is far too costly. If one applies a much laxer criterion for discriminating allophones, as "reasonably different realizations of the same phoneme" one gets, of course, a much more reduced set, that is in turn far more convenient for any application.

There is a clear tradeoff between using many allophones, to generate a better quality of speech at the cost of data gathering time and synthesizer storage, and using a more reduced set with poorer speech quality but great savings in the time needed to generate the database, as well as memory and time to access the elements of the database. This is a crucial aspect in the engineering of any text-to-speech system.

For our work, we started out with the list of Spanish allophones of the AFI/RFE, after A. Quilis [3, 4]. This list contains 45 allophones. It includes many allophones of nasals, nasalized vowels and allophonic variations of /i/ and /u/ in dipthongs. After some empirical studies of some of these allophones, we dropped 19 out of the list, since a smaller number of allophones would result in a shorter list of paraphones, and thus would be easier and faster to generate the database. We dropped these allophones, after considering that:

1) Some allophones were not too different from the standard pronounciation of the phoneme, as was the case of some allophones of /l/ or /n/.

2) Other allophones simply didn't appear differentiated in real speech, as was the case of affricate relizations of /ll/.

3) A third group didn't have a real acoustic personality regardless of the context. Their existance expressed a linguistic concept rather than unique acoustic features. That was the case of /i/ and /u/ within dipthongs.

3. THE LIST OF THE PARAPHONES

The list of paraphones is also a crucial matter for the implementation of the system. This list must enable the synthesis of any possible legal string of characters, for the target language. The list has to be a complete set able to generate any possible context.

A fist approach can be making a list of all allowed diphones (including silence as an allophone). This would be a complete set, but this is not the best solution, since:

1) Some of the diphones may not be necessary. As an example, in our method of synthesis, any transition of a consonant (C) to or from an unvoiced fricative (UF) is synthesized as the concatenation of two units that contain a portion of silence (SI), so we shall have either of these: C-SI + SI-UF, or UF-SI + SI-C. This is done since there is no perceptual quality loss in the speech synthesized in that way, with the advantage of avoiding the need to add to the acoustic database all couples C-UF and UF-C. This means, for Spanish, a saving of 2 x 4 x 20 = 160 units.

2) Some of the possible diphones in Spanish, but most likely in other languages too, don't have an acoustic meaning. For instance, any diphone containing [r] or [rr] is absolutely dependent on context, since [r] and [rr] lack an acoustic target by themselves.

Therefore, not all possible diphones have a place in our list. Some are unnecessary, other don't have acoustic entity.

For this latter case, as well as any in situation with very strong coarticulation and/or very short phonemes, we used triphone units. Thus, our list contains all possible triphones with [r] or [rr] in the middle, so that [r] and [rr] are in any case synthesized with the corresponding triphone.

Fricatives are synthesized by inserting a subphone unit that contains the fricative steady state, between the outgoing and incoming paraphones. Fricatives are assumed to have a stable state of a reasonable duration. A sample of such steady state is segmented from a certain utterance, and is used as a central portion of the given fricative each time it is synthesized.

For the current state of the synthesizer, our list of paraphones (subphones, diphones and triphones) totals 420 paraphones. We are planning to expand this inventory, in a next future, adding triphones and tetraphones, up to about 1000 units.

4. THE LIST OF SENTENCES

For obtaining the paraphones from real speech, under natural conditions, a list of sentences containing at least one realization of all the polyphones has to be designed. Sentences are preferred to isolated words, since it is a more natural context for picking up units to synthesize continous text, as it was our purpose.

The polyphones should appear in contexts without excessive coarticulation; therefore stressed syllables containing the target paraphones are preferred to non-stressed syllables, since the latter may be too coarticulated and/or reduced in duration. The same applies to the words that convey the maximum of lexical information of the sentence, and are therefore likely to be emphasized and over-articulated. It is also a good measure to avoid start and end of phrase for the target syllables, since phrase boundaries (especially the end of phrase) show especial properties with respect to the body of the sentence. And finally, the ever present consideration between economy and

quality dicatates how many target polyphones should contain each sentence. Our conclusion is that 2 is a good compromise. More than 2 per sentence can yield to poor quality units. Just 1 costs at least double time in devising the sentences, and about double time to record and segment them, as compared to 2 polyphones/sentence.

5. RECORDING AND SEGMENTATION OF THE SENTENCES

For the recording of the spoken sentences it is important to choose a suitable speaker. Such speaker must be a person with a clear voice, preferably showing stability in any acoustic domain. A speaker without excessive variations in pitch, stability and cleanliness in the trajectories of the formants, a person with not excessive coarticulation in speech.

For the segmentation of the units, we used a friendly and fast voice editor, WAVES (David Talkin, [5]), with specific tools developed by Jim E Rowley (BL).

After all the paraphones had been segmented, the amplitude was normalized (using a method elaborated by Jan Van Santen, BL), to reduce its variance for each sound, and thus achieve a smoother amplitude pattern of the synthesized speech.

CONCLUSION

We have described the implementation of the acoustic database of a text-to-speech synthesizer for Spanish.

At the time of writing this article, the system is not yet fully implemented as a real time text-to-speech synthesizer, and no systematic tests of intelligibility or naturalness have been carried out, but preliminary impressions are rather satisfactory, with a very high intelligibility and a fair acoustic quality.

REFERENCES

[1] J. P. Olive and M. Y. Liberman. "Text to Speech work at Bell Labs: an overview" JASA (1985).

[2] [1] J. P. Olive and M. Y. Liberman. "A set of concatenative units for speech synthesis". JASA (1979).

[3] Antonio Quilis. "El comentario fonologico y fonetico de textos". ARCO/LIBROS, Madrid.

[4] Antonio Quilis. "Fonetica acustiva de la Lengua Espanola". Editorial Gredos, Madrid. 1988.

[5] "looking at speech". David Talkin, Speech Technology Magazine, April-May 1989