Database Processing for Spanish Text-to-Speech Synthesis

J. Gómez-Mena, M. Cardo, J.L. Madrid, C. Prades

E.T.S.I.T.-U.P.M.
Ciudad Universitaria, 28040-Madrid, Spain
mena@gtd.ssr.upm.es

Abstract. In this paper we will present the method we have used to collect a large database using live recording of a radio transmission and of an announcer. Our selection method uses phonetic and prosodic context information and frequency domain measurements in a search algorithm based on minimising two cost functions: the concatenation cost between adjacent speech units and the distance between the selected unit and a previously calculated target segment. For each basic unit, on average, we have collected around 100 similar units in different phonetic and prosodic contexts. To segment the basic units we have used our highly accurate speech recognition system described in [1]. This collection of units has been used in our TTS system giving marked improvement in naturalness. With this new system we hope to help people with disabilities that need to listen or speak using artificial speech.

1 Introduction

Speech synthesis today is a mature technology in terms of intelligibility, but in naturalness it has yet some severe shortcomings to overcome. Being our major goal the use of the text to speech system by people with speaking disabilities, its lack of naturalness has hindered its practical applications, because of the perception of “machine voice” by the listeners.

The synthesis techniques that currently present the best quality are data-based techniques, in which the synthesised voice is made by means of concatenating different pre-recorded voice segments, called “synthesis units”. Some transformation is applied to these concatenated units; we use the TD-PSOLA (Time Domain Pitch Synchronous OverLap-Add) method [2].

Our approach consists in improving the quality and size of the voice segment database. In previous versions of our TTS system, only one token of each synthesis unit was stored. This implies that most of the time, the time domain and spectral characteristics of the synthesised unit will differ from the features of the stored voice segment. This segment would have to go through heavy signal processing, thus suffering from distortion and the loss of some subtle but important phonetic and prosodic information.

The main difficult with this approach is the very high cost of the collection and adaptation of the voice segments. The huge size of the new database made unfeasible the manual methodology we had used before, so we had to develop automatic procedures to obtain and segment the synthesis units.

2 Database collection

To improve the overall quality of synthesised speech, reducing distortion upon voice segments, we decided to increase the number of tokens of each voice segment contained in the database. Each token would have been uttered in a different phonetic and prosodic context, so the modifications made by the synthesis program in order to adapt the segment's features to the synthesised prosody would be much smaller, leading to a less noticeable distortion.
We planned to include in the database about 100 tokens of each synthesis unit. Because we use 780 different units, the database size grew to almost 80,000 voice segments that made unfeasible the manual selection, segmentation and inclusion of each segment in the database.

2.1 Automatic data adquisition

Instead of working with a speaker, we chose to follow another approach: the recording of a live FM radio program. We could achieve a fairly good signal quality using a high-end receiver plus a data acquisition card that digitised the analog broadband audio signal; the quality was almost so good as we could have achieved in our sound laboratory. The sound samples were stored in a hard disk, allowing further processing.
This method let us capture a lot of information, pronounced in a state-of-the-art studio by a professional speaker. The huge range of prosodic situations recorded was important for our purposes of recording voice in different contexts, and probably better than what could be obtained hand-crafting such an enormous number of phrases for a speaker.
After recording, the voice segments must be segmented and the pitch must be marked, in order to be used in the TD-PSOLA synthesis method. This would have been a cumbersome task but for the automatic segmentation application we developed base on a speech recognition system.

2.2 Automatic segmentation

Our research group has developed a Flexible Vocabulary Speech Recognition System based on semi-continuous HMM (Hidden Markov Models) [1]. These are statistical state models used for pattern recognition that provide a natural and highly reliable way of recognising speech, which is characterised as a parametric random process.

This recognition system must be trained with a known utterance; it segments the input speech, generating the model network. We have used the training feature to automatically segment the recorded speaker's speech. Figure 1 shows the process:

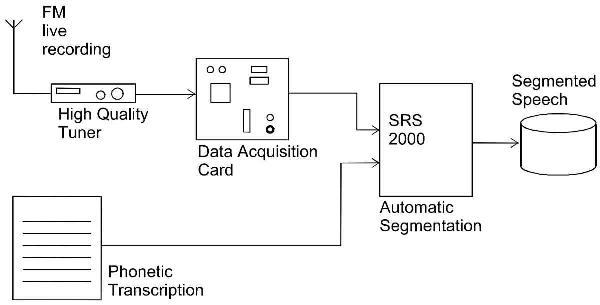

Fig. 1. Database collection

Because our system is still under development, some manual corrections of the segmentation results had to be made on no more than 5% of the recorded utterances.

3 Database construction

Before including a segment in the speech database, it must endure some transformations: power normalisation, a segmentation analysis in which segments that do not comply with some minimal requirements can be rejected, and the calculation of some parameters which will be useful at synthesis time.
Each token of every synthesis unit will be stored with some auxiliary information, to be able to identify and to differentiate the context in which it was recorded, and to allow the computation of a *distance* to other synthesis units [3].
The following features are computed and stored for each token:
Phonetic context: information about the allophones preceding and following each segment.
Prosodic context: fundamental frequency, duration, logarithmic power and the first cepstral coefficients. The cepstrum is a measurement of the spectral envelope; it frees us from the always difficult computation of formants. They are computed on two windows, at the beginning and at the end of each segment.
When including a new token in the database, these features are measured and then compared to the other tokens belonging to the same synthesis units; thus allowing us to select for inclusion only different enough tokens, covering the widest possible decision space [4]. A segment will be added to the database only if it broadens the choice of different contexts at synthesis time.

4 Speech synthesis

A new method for the selection of units at synthesis time was implemented. It is based on a function minimisation algorithm; this function is the sum of two cost functions:

The *target cost*, an estimation of the difference between a predicted target unit t_i, and each of the tokens of this unit stored in the database.

The *concatenation cost,* an estimation of the quality of the join between two consecutive units u_{i-1} and u_i.

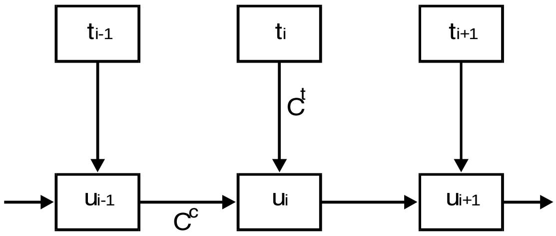

Fig. 2. Costs for unit selection

The target cost is computed as a weighted distance, using the following features: phonetic context, duration, power and pitch. When computing the concatenation cost, our goal is to minimise the acoustic distortion, so to calculate the distance between the units at both sides of the join, we chose to use the pitch difference and the cepstral distance (the Euclidean distance of the cepstral coefficients, which has proven to be a good measurement of acoustic distortion).

Special care had to be taken in the selection and optimisation of a good set of weighing coefficients for both target and concatenation costs; an automatic tuning method comparing the mean Euclidean cepstral distance between vectors from selected units and target segments, known it the special test case of synthesising natural speech utterances from the speaker of the source database. This method is computationally expensive but produces much better speech than hand tuned weights.

Both the database and the selection procedure were implemented with the purpose of transferring the most of the computational load to the process of building the database itself, thus freeing the synthesis program, which can be run in any low-end personal computer.

5 Results

The method that we have described here undertook some preliminary testing, showing a promising improvement in the naturalness of the synthesised sentences.

Due to time constraints, we have still not tested the described system completely. We are planning to write a thorough testing procedure in which a large number of

listeners will be exposed to voice synthesised from different speakers resembling different genders, ages and social background.
The conclusion that we draw from the few experiments made is that having in the database so many speech samples extracted from a broad spectrum of phonetic and prosodic contexts enables us to find a near optimal candidate at synthesis time, minimising the extent of the signal-processing transformations and thus reducing the acoustic distortion in the synthesised speech.

6 Discussion and future work

Any improvement on the quality and naturalness of our system will increase its applications in the realm of aids for the speaking-impaired people, because synthetic speech will be more readily accepted by any listener. This can increase social acceptance for people with disabilities.
The architecture of our system was intended to allow easy modification or substitution of any software module. Thus, improvements in the HMM-based speech recogniser or in the synthesis algorithm can be quickly ported to our system, improving its overall quality or performance.
Future work will aim to improve the prosody generation module, because in this work we have still used not very complex models. It should be time to include the latest theories of energy, duration and pitch modelling, with which we could achieve dramatic improvements, specially in long sentences.

References

1. SRS-2000: A Speech Recognition System, Internal Report, Escuela Técnica Superior de Ingenieros de Telecomunicación, Universidad Politécnica de Madrid, 1999.
2. Eric Moulines and Francis Charpentier. : "Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones". Speech Communication, 1990, Vol. 9, pp. 453-467.
3. Alan W. Black and Nick Campbell: "Optimising selection of units from speech databases for concatenative synthesis". Proceedings of EUROSPEECH'95, Madrid, 1995, Vol. 1, pp. 581-584.
4. W. J. Wang, N. Campbell et al. : "Tree-based unit selection for English speech synthesis". Proceedings of the International Conference on Acoustics, Speech and Signal Processing, Minneapolis, 1993, Vol. 2, pp. 191-194.