A Greek TTS Based on Non Uniform Unit Concatenation and the Utilization of Festival Architecture

Zervas P., Potamitis I., Fakotakis N., Kokkinakis G.

Wire Communications Lab, Department of Electrical & Computer Engineering
University of Patras, 26500, Rion, Patras, Greece
Email: {pzervas,potamitis}@wcl.ee.upatras.gr

Abstract. In this article we describe the first Text To Speech (TTS) system for the Greek language based on Festival architecture. We discuss practical implementation details and we capitalize on the preparation of the diphone database and on the prediction of phoneme duration module implemented with CART tree technique. Two male and one female databases where used for three different speech synthesis engines, namely, residual LPC synthesis, TD-PSOLA and MBROLA technique.

1 Introduction

The waveform speech synthesis techniques can be divided into three categories. The general-purpose concatenative synthesis, the corpus based synthesis and the phrase splicing. The general-purpose concatenative synthesis translates incoming text onto phoneme labels, stress and emphasis tags, and phrase break tags. This information is used to compute a target prosodic pattern (i.e., phoneme durations and pitch contour). Finally, signal processing methods retrieve acoustic units (fragments of speech corresponding to short phoneme sequences such as diphones) from a stored inventory, modify the units so that they match the target prosody, and glue and smooth (*concatenate*) them together to form an output utterance. Corpus based synthesis is similar to general-purpose concatenative synthesis, except that the inventory consists of a large corpus of labeled speech, and that, instead of modifying the stored speech to match the target prosody, the corpus is searched for speech phoneme sequences whose prosodic patterns match the target prosody. Last but not least, at phrase splicing technique the system units are stored prompts, sentence frames, and stored items used in the slots of these frames which are glued together.

General-purpose concatenative synthesis is able to handle any input sentence but generally produces mediocre quality due to the difference of the spectral content in the connection points. On the other hand corpus based synthesis can produce very high quality, but only if its speech corpus contains the right phoneme sequences with the right prosody for a given input sentence. Phrase splicing methods produce natural speech, but can only produce the pre-stored phrases or combinations of sentence frames and slot items. If the slot items are not carefully matched to the sentence frames in terms of prosody, naturalness is degraded. The proposed work is supported by GEMINI (IST-2001-32343) EC project.

2 System Architecture

This paper describes the construction of a Greek TTS based on general-purpose concatenative synthesis architecture. In particular, three different engines have been taken into consideration, the residual LPC synthesizer, the TD-PSOLA and the MBROLA synthesizer.

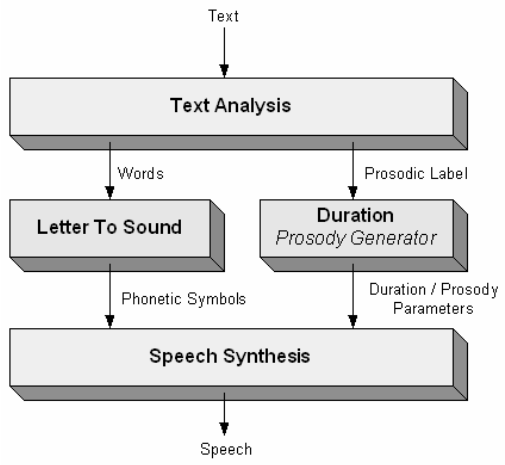


Fig. 1. Text-To-Speech system architecture

Festival is a general multi-lingual speech synthesis system developed at Centre for Technology Research, Edinburgh, Scotland (CSTR) [1, 2]. It consists off a general framework for building speech synthesis systems. It enables the construction of an operational TTS through a number APIs: from shell level, though a Scheme command interpreter, as a C++ library, and an Emacs interface. The architecture of FESTIVAL is diphone-based utilizing the Residual-Exited LPC synthesis technique. In this method, feature parameters for fundamental small units of speech such as syllables, phonemes or one-pitch-period speech, are stored and connected by rules. In our system (Fig. 1), we used a database consisting of diphones.

Mbrola is a speech synthesizer based on the concatenation of diphones coded as Pulse Code Modulation 16 bit linear signals. It takes a list of phonemes as input, together with prosodic information (duration of phonemes and a piecewise linear description of pitch), and produces speech samples using linear coding at 16 bits, at the sampling frequency of the diphone database used. Mbrola is *not* a Text-To-Speech (TTS) synthesizer since it does not accept raw text as input [3].

3 Greek TTS Implementation

Hereafter, we describe the creation of two diphone databases, a male and a female, required from the residual LPC synthesizer provided by the Festival toolbox and for our TD-PSOLA implementation. Diphones are speech segments beginning in the middle of the stable state of a phone and ending in the middle of the stable state of the

following one. Diphones are selected as basic speech segments as they minimize concatenation problems, since they include most of the transitions and co-articulations between phones, while requiring an affordable amount of memory, as their number remains relatively small (as opposed to other synthesis units such as half-syllables or triphones).

3.1 Diphone Database Building

The selection and the recording of the corpus greatly affect the overall quality of the synthesized speech.

A male speaker was asked to read a 900-word phonetically balanced text corpus in a well articulated and at a natural manner. Thus we wanted to ensure that the diphones would be available in a neutral prosodic context.speech database was used for the creation of the male voice concatenation database (Fig. 2). Besides the creation of diphones and some times triphones we created and all the vowels and consonants of our language. As a result our database was consisting of 398 diphones, 24 triphones and 22 phones of the vowels and consonants. The number of the selected units and their partitioning in triphones and diphones has been chosen according to MBROLA requirements.

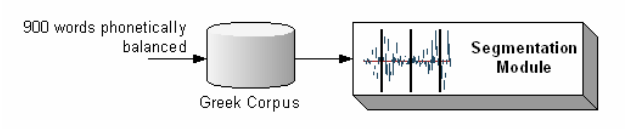
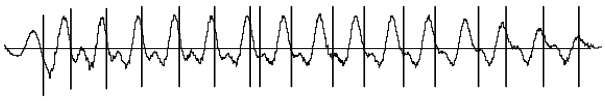


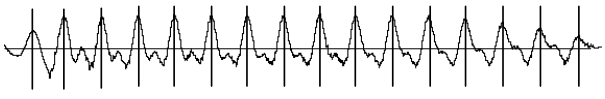
Fig. 2. Greek corpus segmentation procedure

The female voice was created from a 679-word speech database. Contrary to male voice where we used natural carrier words in this case we use ***nonsense carrier words*** to collect all possible diphones and some triphones, following [6].

The words uttered where constructed in a way that the extracted diphone or triphone to be in the perfect condition in order the best merge possible to be achieved. Finally the voice had 565 diphones and 114 triphones covering all the greek language.



(a)



(b)

Fig. 3. a) Automatic placement of pitch-marks. b) Correction of the automatic placement of pitch-marks.

Both voices were recorded in a studio with professional actorsLPC residual synthesis requires the LPC coefficients (perceptual experiments have indicated that 16 coefficients were adequate), residual term of the various speech segments and pitch marks. Epoch-extraction technique was employed to derive the pitch periods of the signal (Fig. 3a). Subsequently, we manually corrected errors in pitch-mark selection (Fig. 3b).

As far as it concerns the voiced parts of the speech, the pitch-marks where placed with a synchronous rate, meaning that we first traced the periods of the signal and then the pitch-marks were placed at the max point of the period. For the voiced parts of the signal they were placed with a constant rate.

As regards the MBROLA synthesizer we have made use of the Gr2 Greek database [4] that has been encoded in TCTS Labs [5].

4. Duration Module

The prediction of the phoneme duration in a specific phonemic and prosodic context is a crucial factor for the performance of TTS systems. For our system we used tree-based modelling and in particular the CART technique. A 500-word speech database was constructed to study the duration model of the Modern Greek language. This database covers all the Greek phonemes and their most frequent contextual combinations. It contains words of various syllabic structures in various locations. The 500 words were spoken in an isolated manner by eight Greek native adult speakers, (four male and four female). The speech database was then labelled manually. The complete database constructed contains a total of about 35.000 entries. Each entry consists of a phoneme label, its duration, its normalized duration, its context and the length of the word it belongs to.

In order to apply tree-based modelling clustering we calculated the mean and standard deviations of duration from the entries. Tree-based modelling is a nonparametric statistical clustering technique which successively divides the regions of feature space in order to minimize the prediction error. The CART technique, a special case of tree-based modelling, can produce decision trees or regression trees with the type of duration. The advantage of the CART technique is the ease of interpreting decision and regression trees. The tree predicts *zscores* (number of standard deviations from the mean) rather than durations directly. After prediction the segmental durations are calculated by the formula:

$$\text{Duration} = \text{mean} + (\text{zscores} * \text{standard deviation})$$

5 F0 Generation Module

For the purpose of the accurate regeneration of the intonation patterns the basic idea was to capture the F0 contour's characteristics by the determination of all its turning points (maxima and minima) in association with discrete textual phenomena along with information about the location of emphasis. For this reason the syllables of the input text were labelled in terms if a set of discrete features (Table 1) and a set of

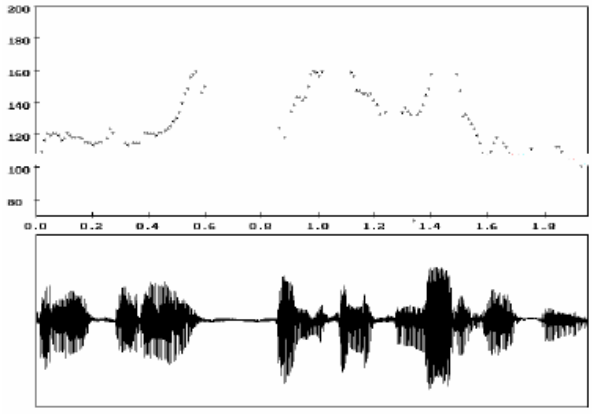
rules which assigns a target F0 level (BASE, MID, TOP or FOCUS) for every syllable was extracted. The kind of textual information used for the syllable's labelling was selected on the basis of its unambiguous extraction directly from the input text except for the information concerning the location of emphasis which is manually provided. The intonation rules have the form:

a,b,c, … = F0 level

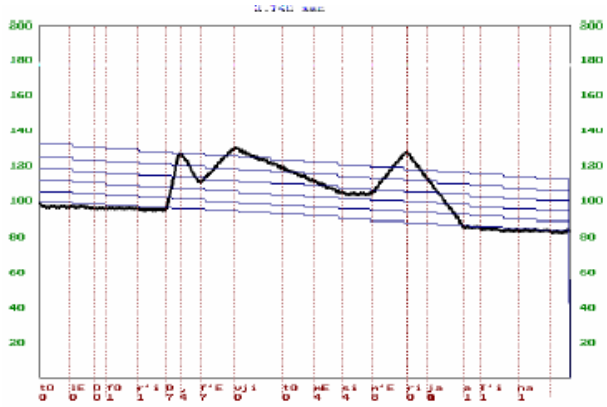
Syllable's features
Stressed/unstressed syllable
Ultimate/penultimate/antepenultimate syllable
Distance in syllables from the previous stressed syllable
Distance in syllables from the next stressed syllable
Distance in syllables from the phrase boundary
Emphatic/non emphatic syllable according to the segmentation of the sentence in the pre-focal, focal and post-focal parts

Table 1. Discrete syllable features used for the association of the turning points with the input text.

The rules do not produce absolute F0 values for every syllable but rather the syllable's corresponding pitch value according to the calculation of the four declined lines with respect to the sentence's duration and the location in time of the emphatic items. For the generation of the appropriate F0 contour the input to the intonation algorithm is the text string enriched with emphatic markings which reflect speaker's intonational focus. First, the declined lines are determined according to the sentence duration and location of the emphasis. Then, the input text is processed and each syllable is assigned a unique vector representing its attributes according to table 1. Finally, every syllable is assigned a F0 level according to the rules and the final contour is constructed by linear interpolation between the successive levels. The resulting pitch contour is a fairly accurate reproduction of the original one as can be seen in figure 4, as far as the patterns used for the analysis are concerned.



(a)



(b)

Fig. 4. Pitch contours of the sentence “My father will come at noon from his work” with emphasis “at noon”: (a) Original contour, (b) F0 contour of prosody module

5 Conclusions

The work we described here was the creation of a Greek diphone-based database for residual LPC synthesizer of Festival architecture and the application of duration derived from CART tree technique. Sample files that demonstrate the high quality of the synthesis results and a Java based web-TTS under construction can be found at http://slt.wcl.ee.upatras.gr/Zervas/index.asp.

Further work focuses on prosody modelling and specifically on the intonation module utilizing the Bayesian networks approach.

References

1. Black A., Taylor P., "The Festival Speech Synthesis System", Technical Report HCRC/TR-83, University of Edinburgh, Scotland, (1997), available at http:// www.cstr.ed.ac.uk/projects/festival.html
2. Black A., Taylor P., “The Festival Speech Synthesis System”, Carnegie Mellon University, Pittsburgh, PA, available at http://www.cstr.ed.ac.uk/projects/festival
3. Dutoit T., “*An Introduction to Text-to-Speech Synthesis*”. Kluwer, (1997)
4. http://www.di.uoa.gr/speech/synthesis/demosthenes
5. http://tcts.fpms.ac.be/synthesis/
6. Isard S., Miller D. Diphone Synthesis Techniques. *Proceedings IEE International Conference on Speech Input/Output* (1986), 77-82
7. Moulines E., Charpentier F., Pitch-synchronous Waveform Processing Techniques for Text-to-speech Synthesis Using Diphones. *Speech Communication*, 9(5/6):453-467 (1990)
8. Galanis D., Darsinos V., Kokkinakis G., Modeling of Intonation Bearing Emphasis for TTS-Synthesis of Greek Dialogues., *ICSLP*, vol.3, USA, (1996)

9. Stylianou Y., Dutoit T., Schroeter J.. Diphones Concatenation Using a Harmonic plus Noise Model of Speech. *Proceedings Eurospeech Conference*, (1997)
10. Sgarbas K., Fakotakis N., Kokkinakis G. "A PC-KIMO Based Bi-Directional Grapheme/Phoneme Converter for Modern Greek", *Literary and Linguistic Computing Journal*, 13(2):65-75, (1998)
11. Yiourgalis N., Kokkinakis G., "A TTS System for the Greek Language Based on Concatenation of Formant Coded Segments". *Speech Communication* (1996)
12. Haan P., Oostdijk M. (eds.) "Prosody in NIROS with FONPARS and ALFEIOS." Proceedings 18, pp.107-119. University of Nijmegen, Department of Language and Speech.
13. Styger T., Keller E. Formant synthesis. In: E. Keller (ed.), *Fundamentals of Speech Synthesis and Speech Recognition: Basic Concepts, State of the Art, and Future Challenges* (1994) 109-128