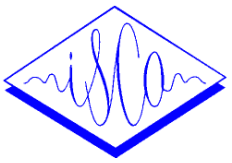
ISCA Archive
http://www.isca-speech.org/archive



6th International Conference on Spoken Language Processing (ICSLP 2000)
Beijing, China
October 16-20, 2000

DESIGN AND IMPLEMENTATION OF A GREEK TEXT-TO-SPEECH SYSTEM BASED ON CONCATENATIVE SYNTHESIS

Costas Christogiannis, Yiannis Stavroulas, Yiannis Vamvakoulas, Theodora Varvarigou, Agatha Zappa
Telecommunications Laboratory
Department of Electrical and Computer Engineering
National Technical University of Athens
9 Iroon Polytechniou, 15773, Athens, GREECE

Chilin Shih
Speech Synthesis Research Department
Bell Laboratories, Lucent Technologies
700 Mountain Avenue, Murray Hill, NJ, USA, 07974

Amalia Arvaniti
Department of Foreign Languages and Literatures
University of Cyprus, P.O. Box 20537, Nicosia 1678, CYPRUS.

ABSTRACT

The goal of this paper is to present the work carried out up to now for the development of the Greek Text-To-Speech (GRTTS) system by NTUA. The system under consideration is based on the method of concatenative synthesis and follows the Bell Labs approach to this technique. In order that the input text to the GRTTS is translated into continuous synthetic speech the following modules have already been studied and implemented: (i) module for the linguistic analysis of the input text; (ii) the acoustic inventory module. On the same time it is under development the duration module of the GRTTS, for the computation of the appropriate temporal structure of synthesized speech. The objectives of the above studies, in combination with the concatenative synthesis technique, which is one of the simplest methods for speech synthesis, are to bypass most of the problems encountered by other synthesis methods such as articulatory and formant synthesis systems. The major objective is to minimize abrupt discontinuities and thus maximize the naturalness of the synthesized utterances.

1. INTRODUCTION

The TTS system for modern Greek (GRTTS) is based on a modular architecture developed by Bell Labs [1]. The overall system can be seen as a pipeline comprising a number of modules, where each module handles a discrete stage of the TTS process:

- The **Transcription Module** consists of three processing steps. The *Lexical Analysis step* receives the raw input text and performs such tasks as classification of the words into grammatical categories, expansion of abbreviations, numerals, dates etc, and syntactic analysis. The *Transcription step* handles the actual transcription into phonetic representation. The *Prosodic Formatting step* is concerned with the prosodic formation of the sentences, the application of sandhi effects and the syllabification of the transcribed text.
- The **Duration Module** computes the duration of the phones, on the basis of a number of factors, such as stress and/or their syllabic position.
- The **Intonation Module** determines the intonational contour of the sentences.
- Finally, the **Synthesizer Module** receives the augmented phonetic transcript and converts it to speech produces the synthesized speech waveform, from the glottal source and other parameters.

In the present paper we describe (i) the development of the transcription first module of the GRTTS that performs the linguistic analysis of an input text in Modern Greek; (ii) the design and the construction of the acoustic database to be incorporated in the synthesizer module of the GRTTS and (iii) the progress on the study of duration module.

In Section 2 we define the phones for Modern Greek. In Section 3 we present how morphological analysis of the input text is performed and we describe the finite state transducers (FSTs)

developed for the morphological analysis and handling of the various categories of Greek words such as abbreviations, dates, numerals and ordinals. Section 4 describes the selection of specific diphone segments as the elementary speech units for our inventory. In Section 5 we present information relative to the duration modeling in order to assign to each phoneme, taking into account various contextual factors. Finally our conclusions are summarized in Section 6.

2. PHONE DEFINITION FOR MODERN GREEK

The Greek alphabet consists of 24 letters. Single letters and combinations of these letters represent 33 phones, five of which are vowels, while the rest 28 are consonants. The IPA symbols of the 33 phones are listed in Table 1.

Vowels	
Consonants	***Stops:*** ***Fricatives:*** ***Affricates:*** ***Nasals*** ***Liquids***

Table 1. The 33 phones of the Modern Greek

In general, there is a direct mapping of letters into phonemes. Pronunciation in Greek is rule-based, in that particular letters and combinations of letters (almost always) correspond to specific phones [6]. There are a few exceptions, however, most of which relate to the syllabification of the word under consideration [2] and result in an additional set of 8 phones. Furthermore, additional rules are needed to account for the fact that some graphemes do not correspond to the same phone in all contexts.

Also, it should be noted that Modern Greek, unlike other languages, such as English, is characterized by simplicity in terms of the following aspects:

(i) Orthography is highly regular, in that graphemes and strings of graphemes represent (almost) always the same phone; this greatly facilitates the procedure of phonetic transcription. In other words, in Greek it is easier to do the conversion from text to phones because many allophonic rules are evident from orthography.

(ii) The language has a five vowels system, the quality and duration of which does vary with stress and context, but not greatly [4], [5].

(iii) Greek has also a small number of infrequently used diphthongs (such as /ai/ and /oi/). Because of their rather marginal status in the linguistic system and for reason of economy, we decided to treat these diphthongs as realization contexts of the five vowels rather than as separate phones.

3. MORPHOLOGICAL ANALYSIS AND TRANSCRIPTION MODULE

The transcription module of the GRTTS is used to perform text analysis (or alternatively linguistic analysis) for the transformation of Greek ASCII text into a phonetic representation output form, readily convertible to speech.

The linguistic analysis and the conversion from text to phonetic transcription is performed in three stages, using *Lextools* [6], a toolkit for producing finite-state linguistic analyzers for various applications. In the first step, the so-called *diphthongs* of Greek (i.e. combinations of letters corresponding to single vowels) are compacted to single symbols. In the second step the letters are mapped to phones. Finally, any assisting symbols, such as hyphens, are deleted.

Using *Lextools*, one can develop analyzers that can handle diverse problems such as grapheme-to-phoneme conversion, lexical (morphological) analysis, abbreviation expansion, numeral expansion, and syntactic analysis. To accomplish this task, pronunciation and syntactic rules of Modern Greek must be taken into account.

The morphological analysis is performed by a transducer that is based on the following modules:

- A *lexicon* containing the stems of all Greek nouns and adjectives, which are used for processing and expanding the abbreviations.
- A *look up table* comprising the paradigms related to the suffixes of nouns and adjectives depending on number, gender and case.
- Four *finite-state transducers* for expansion and transcription of numerals, ordinals, dates and abbreviations respectively.
- Separate *finite-state transducers* for the transcription of single-form words (i.e. adverbs, conjunctions and prepositions).

More specifically as the first step of the morphological analysis, the transducer performs the following conversions to the input text: all upper case letters are converted to lower case; hyphen symbols are translated to space symbols. The resultant string is accepted for further processing if it is a list of words separated by spaces or commas and ends in one of ‘.’, ‘,’, ‘!’, ‘;’ (where ‘;’ is the question mark in Greek).

As a second step, a basic morphological analysis is performed. This is needed so that numerals, ordinals, dates and abbreviations can be accurately expanded using the correct form among several alternatives, since several lexical categories of Modern Greek have various forms.

Specifically, the Greek lexicon includes both words, which have a single form, and words that have several [2]. Adverbs, prepositions and conjunctions have only one form. However, articles, nouns, adjectives, verbs and (most) pronouns have several forms (depending on person, number, tense, aspect and mood for verbs, and on case, number and gender for the other lexical categories). At present, our transducer can handle all single-form words, and definite articles, nouns, adjectives and pronouns, as well as verbs in the present tense.

Abbreviations translate short strings to full text [2]. The generated text may be one or more words. We distinguish between two cases:

(i) Abbreviations the expansion of which has the same form in all contexts;

(ii) Abbreviations which are combinations of single- and multiple-form words, such as nouns (or adjectives). In this case the expansion becomes a more complicated task. Specifically in this case, the transducer:

The transducer that translates *numbers* to text is created with the appropriate utility of LEXTOOLS [1],[6]. The main difference between numerals in Greek and English is that in Greek several numerals have to agree in case, gender and number with the following noun that they determine. At present, only integer numbers are supported.

Similarly to numerals, *ordinals* in Greek are also multi-form words [2]. The expansion of ordinals is facilitated by the fact that when they are written as numbers, they are always followed by the suffix of their corresponding case and number.

The current version of the transcription module also supports syntactic analysis and syllabification. The purpose of the syntactic analysis, as implemented, is to rule out some invalid combinations that are produced by the morphological analysis phase. Syllabification is performed on the written text following the rules described in traditional Greek grammars.

4. DESIGN AND CONSTRUCTION OF THE ACOUSTIC INVENTORY

The Greek synthesizer is a concatenative system, based on a set of prerecorded acoustic inventory elements that represent all the possible phone-to-phone transitions of the language [7].

The proper design of the database is of high importance and requires special care, since we have to include all the units needed for optimum quality during synthesis, and minimize at the same time the size of the inventory. The set of stored speech segments in its totality should cover all legal phone sequences of the language, including inter-word combinations.

Based on the assumption that the Greek system would basically need diphones and not larger concatenative units, we first generated a list of all possible phone combinations. As already stated we have defined 33 phones for our system. Furthermore we take into account silence (represented by the symbol "*") which is used as the initial or final phone in a sentence, or is involved in transitions with silence.

Thus the possible diphone units amount to 1156 (34^2). There are, however, two types of diphones that can be excluded from the inventory of all possible combinations:

(i) Phone-to-phone sequences which never occur in Modern Greek, as a consequence of phonotactic constraints of the language. In the case of Greek language there are no specific rules or conditions for allowed or not allowed sequences of phones but we have to examine each diphone pair individually.

The pairs phone1-to-phone2 where the transition naturally incorporates a section of silence. The existence of transition with silence has only slight coarticulatory effects on each of the two phones [1]. This fact allows us not to record and store these units as diphones, but to build them up out of singletons.

Considering the aforementioned assumptions, we excluded from the inventory 534 diphone pairs with minimal coarticulation and 91 diphones because of phonotactic constraints that disallow them. Thus from the total 1156 dyads, 625 can be constructed during synthesis, while 531 need to be recorded and included in the database of acoustic units. These 531 remaining diphones consist of:

(i) *Medial diphones*, that is sequences of phone pairs occurring within Greek words.

(ii) *Cross-word sequences*. For 365 diphones of the inventory we had to use two words to get a unit because that diphone does not occur within a word. No distinction is made between inter-word and intra-word units, on the assumption that the effect of word boundaries on the phones involved is negligible. In other words, we consider that people can always pronounce two words in the same way as if these were one.

(iii) *Combinations for loan (non-Greek) words*. This case concerns mostly cross-word diphones where the first word is a borrowing. The loan words were used for collecting diphones that are not possible clusters in Greek, but may occur as across-word-boundary combinations (we should mention that Greek allows only /n/ and /s/ word-

finally, except in borrowings, such as ‘parking’). We collected 264 such diphones, which corresponds to 72% of the total (365) cross-word diphone units.

It is noted that by covering the situation of loan words much flexibility is gained for the system, with relatively small cost for the size of the database.

Apart from the medial diphones we also collected:

- 32 starting diphones, that is combinations of silence (*) and phone ;
- 25 ending (or final) diphones, that is combinations of phone and silence;
- 220 triphones that correspond to ending triads of phones, that is combinations in the form phone1-phone2-*.

In Table 2 we summarize the aforementioned collected acoustic units depending on their type.

Type of unit	Population
Medial diphones (including cross-word units and borrowed words)	531
Starting diphones (silence-phone)	32
Ending diphones (phone-silence)	25
Ending (final) triphones (phone-phone-silence)	220
Total Number of Units	**808**

Table 2. Type and size of acoustic units for the Greek TTS.

The segmentation and extraction of the acoustic units requires that they be part of actual words of the language and then that these words be embedded in a sentence environment, so as to maximize the naturalness of the sentences to be recorded. The neighboring phone context is a very important factor. Evidently, the diphone or triphone units needed to synthesize a particular word will most likely not have been originally uttered as part of that word. Thus one of the primary objectives during the construction of such a system is to select units that will minimize possible discrepancies between adjacent diphones in a given synthesized word.

Hence for the recording and extraction of the acoustic units we used the strategy of diverse or mixed context, which is a very efficient technique for increasing the chance that one of the targets from the mixed environment will meet the requirements for quality and naturalness.

5. DURATION MODELING

The duration module of the GRTTS assigns a duration value to each phoneme. This value is estimated by taking into account various contextual factors, using multiplicative models that are fitted from a speech database collected specifically for the duration study. The first stage of the duration modeling is already accomplished and relies on constructing a set of sentences to be read by the speaker, which cover all the desired combinations of the contextual factors. This procedure involves the following steps:

Step 1. Automatically transcribe 20000 short paragraph-sized sentences from a text corpus into their phonetic representation, using the text-analysis tool described in Section 3.

Step 2. Code each segment with relevant factors. In the case of Modern Greek these factors were: segment identity, identity of previous and next segment, presence or absence of stress, syllable type (onset, coda, nucleus), as well as the distance from the beginning and end of the word, phrase and utterance. We then combined the coded factors into factor pairs, each containing the identity of the current segment and another factor.

Step 3. Feed the coded factor pairs to a greedy algorithm [8]. In each round a sentence with the largest number of unseen factor pairs is chosen, until all factor pairs are covered, or the number of sentences has reached a preset limit. In our case a total of 285 sentences, consisting of 38189 segments covering all the factor pairs that are present in the input, were chosen from the 20000 sentence pool.

Step 4. Record the 285 sentences by a male professional speaker.

For the time being we are in the fifth step of this procedure, which is the segmentation stage of the recorded sentences. The final stage is the construction of the data matrix including every phone in the database, its coded contextual factors and its duration. These duration values, along with segment identities and relevant factors will be used at runtime for the computation and derivation of factor coefficients, in the multiplicative duration model, applied during synthesis.

6. SUMMARY

In this paper we have described the various modules we have developed or are under development for the Greek Text-To-Speech synthesis system (GRTTS). Specifically we have presented the module that performs the linguistic analysis of Greek texts, necessary for the appropriate phonetic transcription of the input. Basic morphological analysis of nouns and adjectives and accurately handles abbreviations and

symbols found in written Greek texts. This task is based on the use of lexica, different finite-state-transducers as well as morphological, syntactic and phonological rules. We have also presented the framework of the design of the acoustic database for the GRTTS., in terms of the description of the acoustic units that are included in the database. The acoustic inventory for the Greek synthesis system is made out of diphones and is quite compact. Finally we have presented the completed steps of the ongoing duration-modeling phase.

7. REFERENCES

[1] Richard Sproat, ‘Multilingual Text-To-Speech Synthesis, The Bell Labs Approach’, Kluwer Academic Publishers, 1998.

[2] “PAIDEIA: dictionary of the modern Greek Language - Protypo lejiko ths neas Ellhnikhs”, 1977.

[3] Triantafyllidis, M. “Modern Greek Grammar”, 1998

[4] Amalia Arvaniti, “Acoustic features of Greek rhythmic structure”, *Journal of Phonetics*, Vol. 22, 1994, pp. 239-268.

[5] Amalia Arvaniti, “Secondary stress: evidence from Modern Greek”, *Papers in Laboratory Phonology II*, Chap. 16, pp.398-423, Cambridge University Press.

[6] Sproat, R. “LEXTOOLS: Tools for Finite-State Linguistic Analysis, Technical Memorandum”, Bell Labs, 1995

[7] Chilin Shih, Richard Sproat, ‘Issues in Text-To-Speech Conversion for Mandarin’, *Language Processing*, Vol. 1 No. 1, pp 37-86, 1996.

[8] T.H Cormen, C.E. Leiserson, and R.L. Rivest, ‘In troduction to Algorithms’, The MIT Press, Cambridge, Massachussetts, 1990.