

CONSTRUCTION OF THE ACOUSTIC INVENTORY FOR A GREEK TEXT-TO-SPEECH CONCATENATIVE SYNTHESIS SYSTEM

Costas Christogiannis, Theodora Varvarigou, Agatha Zappa, Yiannis Vamvakoulas

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 development of the Greek Text-To-Speech (TTS) system by NTUA is based on the method of concatenative synthesis and follows the Bell Labs approach to this technique. Concatenative synthesis is one of the simplest methods for speech synthesis and at the same time bypasses most of the problems encountered by articulatory and formant synthesis techniques. The method relies on designing and creating the acoustic inventory of the language by taking real recorded speech, cutting it into segments and concatenating these segments back together during synthesis. The design and implementation of the acoustic database is a key factor for the performance of the synthesizer, since all the possible phone-to-phone transitions must be considered in order to minimize abrupt discontinuities and thus maximize the naturalness of the synthesized utterances.

1. INTRODUCTION

A Text-To-Speech (TTS) synthesizer is a computer-based system able to read any text and convert it into speech that resembles as closely as possible a native speaker of the language reading that text. Thus Text-To-Speech can be defined as the automatic production of speech, through a grapheme-to-phoneme transcription of the sentences to utter.

In general every TTS synthesizer has two basic structural modules: (i) a Natural Language Processing module (NLP), capable of producing a phonetic transcription of the read text, together with the desired intonation and rhythm (often referred to as *prosody*); and (ii) a Digital Signal Processing module (DSP), which transforms the symbolic information it receives into speech. Intuitively, the operations involved in the DSP module are the computer analogue of dynamically controlling the articulatory muscles and the vibratory frequency of the vocal folds, so that the output signal matches the input requirements. In

order to do it properly, the DSP module should take articulatory constraints into account, since it has been known for a long time that phonetic transitions are crucial for the understanding of speech [1]. This can be achieved in two ways :

- Explicitly, in the form of a series of rules which formally describe the influence of phonemes on one another.
- Implicitly, by storing examples of phonetic transitions and co-articulations into a speech segment database, and using them just as they are, as ultimate acoustic units.

Two main classes of TTS systems have emerged from these alternatives, which quickly turned into synthesis philosophies, given the divergences they present in their means and objectives: synthesis-by-rule and synthesis-by-concatenation.

Rule-based synthesizers [2] constitute a cognitive, generative approach of the phonation mechanism and they appear in the form of articulatory and formant synthesizers, which describe speech as the dynamic evolution of parameters [3], mostly related to formant and anti-formant frequencies and bandwidths together with glottal waveforms. Unfortunately, the large number of (coupled) parameters complicates the analysis stage and tends to produce analysis errors. Moreover, formant frequencies and bandwidths are inherently difficult to estimate from speech data.

Concatenative synthesizers, in contrast to rule-based ones, possess a very limited knowledge of the data they handle: most of it is embedded in the segments to be chained up. This feature renders concatenative synthesizers simple but efficient, in terms of the quality of the synthetically produced speech.

In this paper the design and the preparation of the acoustic database to be incorporated in a concatenative TTS system for Modern Greek is described, along with the sequence of all the tasks that had to be completed before the synthesizer could produce its first utterance.

In Section 2 we define the Greek phones and justify the selection of specific diphone segments as the elementary speech units for our inventory. In Section 3 we describe the preparation of the text corpus, which was digitally recorded and stored, from which the complete list of segments was collected. In Section 4 we briefly present our solution to the problem posed by the major coarticulatory effects we observed in Greek, namely the effect of palatal consonants on following vowels and the effect of round vowels on preceding fricatives. Finally our conclusions are summarized in Section 5.

2. PHONE DEFINITION AND SELECTION OF ACOUSTIC UNITS

2.1 Phone Definition

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.

Table 1.
The 33 phones of the Modern Greek

Vowels	i e a o u
Consonants	<i>Stops:</i> p b t d c ʃ k ɣ
	<i>Fricatives:</i> f v θ ð s z ç j x ɣ
	<i>Affricates:</i> tʃ dʒ
	<i>Nasals</i> m n ɲ ŋ m̃ɲ
	<i>Liquids</i> l λ r

Furthermore, 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.

2.2 Design And Selection Of Acoustic Units

As mentioned earlier, 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 [6]. The effort during the construction of the acoustic inventory involves the following tasks:

- Design of the acoustic database, which refers to the appropriate phone-to-phone transitions to be recorded and excised.
- Construction of the text corpus for the speech recording. This task is described in detail in Section 3.
- Recording of the text corpus.
- Performing phone segmentation of the recorded material with the aid of spectrograms and waveforms, through the use of software tools for this purpose.
- Selection of acoustic unit tokens with the goal of simultaneously minimizing the spectral differences between units and the distance of each segment from its “ideal”.

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 (34²). 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.
- (ii) 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 [7]. This fact allows us not to record and store these units as diphones, but to build them up out of singletons. For example the pair /k-t/ consisting of two stop consonants and which is a very frequent cluster in Greek can be synthesized from the pairs /k-*/ and /*-t/. The specific pairs with minimal coarticulation are listed in Table 2.

Table 2.
Consonant pairs with minimal coarticulation.

Phone ₁ (consonant)	Phone ₂ (consonant)
stop	stop
stop	nasal
fricative	fricative
fricative	stop
fricative	nasal
nasal	stop
nasal	fricative
lateral	stop
lateral	fricative

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 3 we summarize the aforementioned collected acoustic units depending on their type.

Table 3.
Type and size of acoustic units for the Greek TTS.

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

3. CONSTRUCTION OF CORPUS TEXT FOR COLLECTION OF ACOUSTIC UNITS

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 [9].

For the TTS system what we need first are phones clearly and fully articulated and (as far as possible) neutral as to context. Any segment that is heavily colored by context is good only when it is used in addition to a basic, colorless set. Embedding diphones in rotating surrounding context through representative place and/or manner of articulations, is recommended for two significant reasons:

- It facilitates the collection of the optimum and most representative utterances for a specific phone.
- It allows us to spot the influence of the context, in terms of coarticulatory effects.

In sum, the strategy of diverse or mixed context 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. Moreover context diversity is far more reliable than logotone context [7], which is based on the idea of repeating a diphone in the same context every time. Units recorded only in similar or identical contexts run the risk of resulting in a sub-optimal situation, during the collection phase, where no representative candidate can be chosen for the database.

After the above assessment we decided to embed each medial diphone in 3 to 5 entirely different contexts, preferably spanning the inventory space. More specifically we decided to embed each diphone in two other phones, one preceding and one following, creating in this way populations of quadri-phones. Then we

looked for existing Greek words that might contain these quadri-phones. Our main concern was to create three or four such contexts for each dyad with high degree of diversity in terms of coloring effects.

As already mentioned, the words for segmentation were placed into sentence frames for recording. These sentences were constructed by hand and their spelling was sufficiently conventional that the speaker uttered them in the way we intended. Also we took care to create short frames, because long sentences may cause unexpected complications such as sentence breaks, changes in intonation or sloppy end ups (e.g. devoicing, creaky voice) by the speaker. Final dyads are embedded in medium long sentences, long enough to get final effects such as pitch lowering, but not so long that the speaker will pause.

It is noted that during the collection of units we never used any part of the frame, since it is highly probable for phones or units in contact with the frame to be "contaminated" by it.

4. COARTICULATION EFFECTS

During the design of the inventory we observed two major sources of coarticulation effects:

- (i) the quality of vowels was affected by the presence of the palatal consonants [c], [j], [ç], [j̥], [ʎ], [m̃j̃] and [ɲ] as preceding phones.
- (ii) affricates ([tʃ] and [dʒ]) and fricatives ([f], [v], [θ], [ð], [s], [z], [ç], [j̥], [x] and [ɣ]) were rounded whenever they were followed by the vowel [u].

In the first case, the effect of palatal consonants was particularly striking in the formant trajectory of [a]. Plotting F1 by F2 of [a], we found one cluster of formant values that represented the realization of the vowel in the majority of contexts, while a separate cluster represented the realizations of [a] when preceded by any one of the aforementioned palatal consonants. In order to take into account this effect and improve the synthesis quality, a smooth transition between the diphones /palatal-a/ and /a-phone/ should be obtained. This could be resolved in two ways: one is to record and collect triphones of the pattern /palatal-a-phone/. An easier solution is to use phone a_{color} from the cluster resulting from the palatal context. An important prerequisite for the latter choice is that in the same cluster all the possible combinations of a_{color} -phone (33 in total) are present. Luckily in the case of the Greek acoustic inventory these combinations did exist in the specific cluster, and thus no further recordings or collection of triphones was necessary. For the moment we have not performed the same process for the other four vowels; however we should note that the coloring effects of the preceding palatal consonants on [o], [i], [e] and [u] is not as strong as for [a].

The coarticulation effects of round vowels on the preceding fricative and affricate consonants were especially strong in the case of [u]. Specifically the presence of this vowel results in lip rounding during the production of the fricative (and the fricative portion of the affricate). This is a widespread phenomenon attested in many languages [7], [8]. In order to address this issue, we collected two different types of each of the Greek fricatives and affricates, from different rounding contexts. Specifically we collected the pure fricative portions of each phone from both an [u]-context and from an [i]-context, which is non-rounding. So

we had two acoustic versions, fricative_{u(round)} and fricative_{i(non-round)} that were involved in patterns of the form /phone-fricative-rounding vowel/. In this case the connection was realized as follows:

/phone-fricative_{i(non-round)}-fricative_{u(round)}-rounding vowel/.

In the first transition the energy is low and the coarticulation has not started yet; however it appears in the second transition. In this way it was possible to eliminate any discrepancies during connection, thus obtaining smooth transitions in the above patterns.

5. SUMMARY

In this paper we presented the framework of the design of the acoustic database for the Greek Text-To-Speech synthesis system. Specifically we defined the phones represented by the letters of the Greek alphabet and their combinations. Furthermore we described the acoustic units that should be included in the database and the technique with which we obtained maximal context diversity for these units. Finally we studied certain coarticulatory effects, the most significant of which were the coloring of vowels (especially low vowel [a]) from preceding palatal consonants and the rounding of fricative consonants when they are followed by the round vowels, [u] and [o].

In conclusion the acoustic inventory for the Greek synthesis system is made out of diphones and is quite compact. Because of the relative phonetic simplicity of Modern Greek, the coarticulation effects we had to face were very few and rather expected, and thus successfully addressed.

6. REFERENCES

- [1] M.J. Liberman, K.W. Curch, "Text analysis and word pronunciation in text-to-speech synthesis", *Advances in Speech Signal Processing*, S. Furuy, M.M. Sondhi eds., Dekker, New York, 1992, pp.791-831.
- [2] J. Holmes, I. Mattingly, J. Shearme, 'Speech synthesis by rule', *Language and Speech*, Vol 7, 1964, pp.127-143
- [3] K.N. Stevens, "Control parameters for synthesis by rule", *Proceedings of the ESCA tutorial day on speech synthesis*, Autrans, 25 Sept 90, pp. 27-37.
- [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] Chilin Shih, Richard Sproat, 'Issues in Text-To-Speech Conversion for Mandarin', *Language Processing*, Vol. 1 No. 1, pp 37-86, 1996.
- [7] Richard Sproat, 'Multilingual Text-To-Speech Synthesis, The Bell Labs Approach', Kluwer Academic Publishers, 1998.
- [8] Karen Livescu, Chilin Shih, Richard Sproat, 'A Romanian Language Text-To-Speech Synthesizer', *Technical Report BL011222-950816-05*, AT&T Bell Laboratories, 1995.
- [9] Joseph Olive, 'A new algorithm for a concatenative speech synthesis system using an augmented acoustic inventory of speech sounds', *Proceedings of the ESCA Workshop on Speech Synthesis*, pp 83-86, 1990.