EUROSPEECH '97
5th European Conference on Speech
Communication and Technology
Rhodes, Greece, September 22-25, 1997

AUTOMATIC CORPUS-BASED TRAINING OF RULES FOR PROSODIC GENERATION IN TEXT-TO-SPEECH

Eduardo López-Gonzalo, Jose M. Rodríguez-García, Luis Hernández-Gómez and Juan M. Villar

E.T.S.I. de Telecomunicación. Univ. Politécnica Madrid
Dep. Señales, Sistemas y Radiocomunicaciones.
Ciudad Universitaria. 28040-Madrid (Spain).
Tel:34.1.5495700. Fax:34.1.3367350. e-mail: eduardo@gaps.ssr.upm.es

ABSTRACT

In this paper, we discuss a methodology for automatic prosodic modeling in Text-to-Speech (TTS) systems. The proposed methodology can be seen as a data-driven strategy to train prosodic rules from the automatic analysis of a specific text and its related speech material. Therefore, our corpus-based training procedure is based on an automatic linguistic analysis of the text and on an acoustic analysis of the speech using automatic speech recognition techniques. Together with the automatic derivation of prosodic rules, our method can be easily extended to obtain specific grammar categories suitable for accurate prosodic modeling of specific tasks. Evaluation results over two different applications and speaker styles, reveal that the proposed automatic prosodic generation procedure is able to provide a noticeable increase in naturalness when adapting TTS system to a new speaker and a new speaking style.

1. INTRODUCTION

Text-to-Speech technology has reached a point where general purpose systems designed for reading texts are being incorporated in several applications. Although the general quality is acceptable, there are applications where a more specific prosody and signal generation are needed. The rapid building of new voices and the adaptation to specific tasks are open research fields involving different related techniques in TTS systems. Specially in the field of man-machine communication, naturalness of state-of-the-art TTS systems need to be improved to properly handle machine responses including task specific and speaker specific speaking styles. However, traditionally, prosodic modeling in TTS systems rely on a set of manually derived rules for prosody generation. The process of deriving these rules is time consuming and is also difficult to generalize to a new voice or a new speaking style.

In [1], we presented an extension of our automatic data-driven methodology (see [2]),for adapting a TTS system to a new speaker and a new speaking style. So far, in this contribution, we will focus on the automatic extraction procedure of prosodic rules based on an acoustic and linguistic analysis of a task-specific corpus. Two main points are addressed:

a) How to extend the general automatic prosodic modeling technique towards specific domain TTS systems.

b) Experimental results by using the proposed methodology over two different tasks.

The organization of the paper is as follows. In Section 2, we briefly describe the automatic prosodic methodology for prosodic modeling. The automatic generation of prosodic rules is presented in Section 3. Section 4 discusses the possible improvement in prosodic modeling when task-specific grammar categories are automatically extracted and included in our methodology. Experimental results and conclusions are given in Section 5.

2. AUTOMATIC PROSODIC MODELING

A possible solution to prosodic modeling is the use of “manual” procedures, as we proposed in [3] for example. This “manual methodology” is based on subjective criteria and it is a tedious time-consuming work.

Automatic prosodic modeling is therefore the key point for adapting a TTS system to a specific task or speaker. The general scheme we propose for producing a data-driven prosodic model (see [1,2,3]) is shown in Figure 1. The input to the system is a monospeaker recorded prosodic corpus and its textual representation. The output of the system is a database of prosodic patterns for prosodic generation and a set of prosodic rules that achieve a mapping between grammatical categories and some linguistic features related to the prosodic generation process of our synthesizer.

____

1 This work was supported by a CICYT contract under the project TIC96-0956-C04-03

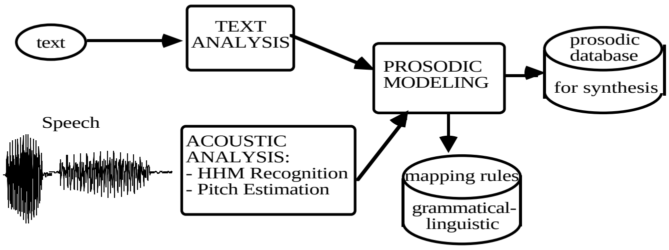

Figure 1: Data-driven methodology for prosodic modeling.

Our system analyzes every sentence of the corpus. We consider in our modeling that sentences are formed by syllables, accent groups (groups of syllables with one lexical accent) and breath groups (groups of accent groups between pauses).

For every sentence of the corpus rule-based **text analysis** gives some grammatical information as categories of words, etc. (see [1,2,3] for a more comprehensive description).

In order to get some more linguistic features and some prosodic features we perform an **acoustic analysis** of the speech signal. This analysis includes acoustic processing of the speech to obtain sound segmentation (including pauses) and pitch contour estimation. This contour is represented by 5 parameters as defined in [3]: two duration values and three pitch values. This set of five parameters can be seen as vectors of 5 components that we refer to as Prosodic Syllabic Pattern (PSP).

These prosodic features and linguistic features mentioned above are stored in the prosodic data-base shown in Figure 1. Although the size of the data-base is proportional to the size of the speech corpus, some methods have been developed to reduce its size (see [4] for more details). The problem we are going to focus now strives in deriving some rules for relating the prosodic and linguistic features. These rules are named mapping rules and are described in the following section.

3. AUTOMATIC EXTRACTION OF MAPPING RULES

Mapping rules between linguistic and acoustic features are essential on our data-driven strategy. These rules make a mapping between grammatical categories obtained from the text analysis and linguistic features derived from the acoustical analysis. These mapping rules are very important for a correct use of the information stored in the prosodic database starting from the text. They need to accurately represent the training material but they also need to be able to generalize to similar structures not present in the training corpus.

The approach we will follow to define the set of mapping rules is based on the derivation of pauses position, breath groups and accent groups [1] And the generation of these rules can be divided in three major parts, that we have called: pre-generation, generation and selection of rules. These three steps are described in the following paragraphs together with an example:.

3.1 Pre-generation of rules

Firstly, given a phrase from the training corpus, the sequence of grammatical categories is obtained by the text analysis module

	(Wait	a	moment	please		.)
Phrase words:	*Espere*	*un*	*momento*	*por*	*favor*	.
Categories:	C30	C27	C42	C28	C32	C9

Secondly, we match the linguistic features from the acoustic analysis module

Phrase:	*Espere*	*un*	*momento*	*por*	*favor*	.
breath gr. type:	7	7	7	15	15	15
pause after word:	no	no	yes	no	no	yes

Finally we obtain information about accent groups

Phrase:	*Espere*	*un*	*momento*	*por*	*favor*	.
accent group:	initial |		final	|	final	

Now, a set of pre-rules are generated. In our example the following pre-rule is generated:

$$\{C30,C27,C42,C28,C32,C9\} \rightarrow$$

$$\rightarrow \left\{\begin{matrix} 7 & 7 & 7 & 15 & 15 & 15 \\ \text{yes}, & \text{no}, & \text{yes}, & \text{no}, & \text{yes}, & \text{no} \\ \text{no} & \text{no} & \text{yes} & \text{no} & \text{no} & \text{yes} \end{matrix}\right\} \begin{matrix} \text{breath gr.} \\ \text{accented} \\ \text{pause} \end{matrix}$$

This mapping is made for every sentence in the corpus. The type of breath group is automatically derived from the acoustic analysis (see [4]).

The whole set of pre-rules need now to be processed to deal with two major problems: a) different lengths of rules; and b) the possibility to derive contradictory rules. Therefore we need to discuss now the different approaches we have developed to accomplish rule-length adjusting, and checking of contradicting rules. We refer these two tasks as generation and selection of rules.

A unique rule for each sentence in the corpus is not very useful due to there is a low probability of using it in another different sentence. The rule only could be used for a sentence with exactly the same sequence of categories. Furthermore, there should be a maximum limit for the length of a rule. Therefore, the

generation of several rules for each sentence is necessary in a practical system.

Another factor that must be taken into account is that two rules could have the same antecedent (sequence of grammatical categories) and two different consequents (linguistic features), therefore a process for rule **selection** is necessary to fix a finite set of rules providing a good compromise between complexity and expected quality of the system. This process assures different antecedents for all rules. It should be noted here that two rules of different length are considered to have always different antecedents.

3.2 Generation of rules

Our rule-generation procedure is quite flexible, we have tested two different approaches.

In the first one, rules as large as possible are generated ending in a pause. Following this procedure, we observed that the great amount of different structures that a general purpose TTS has to deal with, cannot be properly managed with large rules extracted from the corpus, due to the low power of generalization of these rules. However, large rules perfectly suit in particular tasks where usually there are fixed parts such as: *El teléfono marcado es...* (the dial number is...)). These structures are easily identified in the corpus and stored in the prosodic model, so a good mimic of the natural prosody was obtained in this case.
A second approach was tested for those situations where more general rules are needed. In these cases, in order to derive more general rules, we start as we have said before and then generating all subset of rules embedded in each large rule. The rules obtained are consequently shorter and more general and the generated prosody is better for structures that do not appear in the training corpus. As an example, the generated rules with the pre-rule described in section 3.1 were:

{C30,C27,C42,C28,C32,C9} → { ... }
{C30,C27,C42} → { ... }
{C27,C42} → { ... }
{C27,C42,C28,C32,C9} → { ... }
{C42,C28,C32,C9} → { ... }
{C28,C32,C9} → { ... }
{C32,C9} → { ... }

The second method is identical to the first one but the restriction of the rule ending in a pause does not apply, so all combinations of embedded rules inside the pre-rule are generated. Clearly, we generate more rules following this method. Experimental results comparing the two methods are given in section 4.

3.3 Selection of rules

After the generation of rules, it may be the case that there are two rules with the same antecedents and different consequents, so a method for the final selection of rules is necessary.

Two methods have also been proposed for rule selection here. In the first one we keep all rules that are consistent. In the case of several inconsistent rules, we look for the common part of the consequent, then we generate from the rules the sub-rule with the common part. If there is not such a common part, then the rules are not selected. The cause for this phenomenon is that the speaker reading the speech corpus can vary his intonation when saying a particular structure that appears more than once in the corpus. As an example, if we have these two rules:

{C20,C30,C29,C30,C30} → {8,8,2,14,14} breath gr.
{C20,C30,C29,C30,C30} → {4,4,2,14,14} breath gr.

Then, the rule finally generated would be:

{C29,C30,C30} → {2,14,14} (breath gr.)

In the second method, in case of an inconsistent set of rules with the same antecedent, we look for a majority of rules with the same consequent, if the majority exceeds a fixed percentage (35%), then a rule with this consequent is generated. If there is not such majority, then we look for consistent consequents refered only to the position of the pauses; if there is agreement in the position of the pauses then a rule is generated keeping these positions and deriving the type of breath-group by majority. It should be noted that less rules are eliminated following this method, and then more rules are finally generated.

4. FURTHER IMPROVEMENTS THROUGH SPECIFIC GRAMMATICAL CATEGORIES

The methodology just described is general, and it can also be used for non-specific corpus. But in specific applications where specific corpus are recorded further improvements can be obtained. As it is obvious, in a particular task, some syntactic structures and vocabulary words are more relevant, from a prosodic point of view, than others. Therefore specific grammatical categories for these words or syntactic structures should be included.

Another important advantage of our system is that grammatical parsing can be adapted to the task with minimum effort. We have developed a simple method for including these specific categories based only on giving specific categories to the more frequent words in the corpus automatically. As we describe in the next section, the inclusion of this categories provides a more accurate prosodic modeling and improves the naturalness of the synthetic speech because more rules are generated. Note that with specific categories, we gain in specificity but we may loose in generality.

At this point it is straight to design a procedure for adapting the prosodic information of the TTS system to a particular speaker and application task. When an application designer wants to adapt a TTS system to a particular Interactive Voice Response application he only has to perform two easy tasks: a) to prepare a text corpus including a sample of typical sentences the application will generate and b) to record these sentences from the selected speaker. From this information (text and speech), the general prosodic modeling process can be applied to generate the desired prosodic information. As result of this methodology we will provide the TTS system with the necessary mimic to reproduce the speaking style characteristic of both the application task and target speaker.

5. EXPERIMENTAL RESULTS

The proposed methodology has been preliminary tested in two different applications. The first application deals with message generation for a telephonic IVR service providing information about railway stations: departure time, time-tables, fares...The second application is an IVR service for an automatic telephone operated system which includes a dialogue manager module. In this case there are control, confirmation and information messages.

For these two applications particular corpora were designed and recorded by two different speakers. In the first case we used a corpus with 180 sentences while 97 sentences were enough considering the smallest set of typical sentences. We carried out several experiments for each application. The experiments tried to compare the different methods of generation and selection of rules previously described, and combining these methods with the use of both grammatical categories designed for a general purpose TTS and task-adapted grammatical categories. Task-adapted categories included in the trains information system focused on: dates, hours, places and specific words (train, from, to, departure, arrival...). For the telephone operated system the categories were related to telephone numbers, directory names and some specific words (telephone, collect, call, busy...).

In [1], we did a test for task adaptation, where we showed that the automatic methodology applied to a specific corpus improved greatly the naturalness of the synthetic speech. Further listening tests have been made for comparing the different methods for mapping rules training. Two small corpus were designed, the first one was a sub-set of the training corpus, and the other one was a related corpus with different structures that the training corpus although with some specific words in common. Several comparison tests were performed comparing methods for generation and selection of mapping rules.

One general result we observed was that the generation by the second method (all possible combinations) generates more rules, specially if specific grammatical categories are introduced. For instance, in the train corpus the number of rules generated without specific categories was 906 with the first method of generation; this number increases to 3354 with the second method of generation; and adding 15 new specific categories, the number of rules increased to 1393 rules for the first method of generation and 4881 for the second method.

In general, a better improvement in subjective quality is observed when the second method of generation is used, specially this is noted without specific categories. It seems logical that the more rules there are the better generalization is obtained. So when the sub-set of the training corpus is used for test both methods generate prosody very close to the original one. But when the other corpus for test was used better results were obtained with the second method of generation. With regard to the selection method, it was noted that there was a small advantage using the second method. This result was observed using for both selection methods the first method for rule generation.

As a general conclusion, it can be seen from the experiments that the automatic methodology presented, when adapted to a particular speaker and task, provides an important improvement over a general purpose TTS system. It can also be noted that the use of specific grammatical categories results in a noticeable increasing naturalness. More information and some synthetic speech examples can be accessed through our www address: www. gaps.ssr.upm.es/tts.

6. REFERENCES

[1] E. López-Gonzalo, J.M. Rodríguez García, L.A. Hernández-Gómez and J.M. Villar "Automatic Prosodic Modeling for Speaker and Task Adaptation in Text to Speech" in *Proc. ICASSP* . Munich (GERMANY). Apr. 1997.

[2] E. López-Gonzalo and L.A. Hernández-Gómez "Automatic Data-Driven Prosodic for Text to Speech" in *Proc. EUROSPEECH* pp. I-585 I-588. Madrid (SPAIN). Sep. 1995.

[3] E. López-Gonzalo and L.A. Hernández-Gómez "Data-driven Joint F_0 and Duration Modeling in Text to Speech Conversion for Spanish" in *Proc. ICASSP*, pp. I-589 I-592. Adelaide (AUSTRALIA). Mar. 1994.

[4] E. López-Gonzalo and J.M. Rodríguez-García "Statistical Methods in Data-Driven Modeling of Spanish Prosody for Text to Speech" in *Proc. ICLSP* pp. 1373-1376. Philadelphia (USA). Oct. 1996.