AUTOMATIC PROSODIC MODELING FOR SPEAKER AND TASK ADAPTATION 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

One of the most important demands for future TTS systems is their ability to improve naturalness when embedded in a particular task or application that requires a particular speaking style for a particular speaker. In this paper, we present a new prosodic modeling procedure for improving naturalness by adapting a TTS system to a new speaker and a new speaking style. The proposed procedure is an extension of our automatic data-driven methodology presented in [1], to model both fundamental frequency and segmental duration. Automatic linguistic and acoustic analysis are performed on both a task dependent text corpus and the recorded material from the selected speaker.

1. INTRODUCTION [1]

Nowadays Text-to-Speech (TTS) technology has reached a point where an important number of applications are possible, specially in the field of Interactive Voice Response (IVR) systems. However, although the global quality of most TTS systems is acceptable, there are still important challenges demanding more and more research interest.

The rapid building of new voices and the adaptation to specific tasks are open research fields involving different related techniques in TTS systems. In our particular case, the focus will be on the generation of proper prosodic information, which is one of the most important information related to the final naturalness of the synthetic speech.

In this paper, we present an automatic prosodic modeling procedure based on an acoustic and linguistic analysis on a corpus designed to cover the functionality of a particular task and recorded by a particular speaker whose speaking style the system tries to imitate. The acoustic analysis includes two major steps:

a) Sound segmentation using an HMM-based automatic speech recognition system and a fine procedure for spectral transition detection.

b) High-precision fundamental frequency analysis for pitch contour estimation.

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

This methodology starting from a monospeaker recorded corpus provides, when compared to a "manual" one [1], a good prediction of frequency and duration patterns. Therefore in this work we extend it to reach the mimic necessary to provide a TTS system with both task and speaker adaptation capabilities. Two main points are addressed:

a) How to extend the general automatic prosodic modeling technique towards adaptation facilities.

b) Experimental results by using the proposed methodology over two different tasks demanding different speakers and speaking styles.

In section 2, we describe the basic functionality of the proposed automatic prosodic modeling strategy. The different steps needed to perform the speaker and task adaptation are discussed in section 3. Experimental results are given in section 4.

2. AUTOMATIC PROSODIC MODELING

Generating proper prosodic information is one of the most important issues for synthesizing speech. Although many approaches of prosody generation have been proposed in the past for text-to-speech (TTS), it still remains a problem to model the variability and fluency of natural speech. For a number of years, the build-up and systematic use of a prosodic corpus (see for example [2] for French) have been recognized as the key to generate more natural synthetic speech. The problem is how to extract the prosodic knowledge from this database.

A possible solution to prosodic modeling is the use of "manual" procedures, as we proposed in [3] for example. Prosodic patterns were obtained by averaging some manually labeled data from a single speaker. 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. As we said above, the use of a prosodic database is the current tendency to obtain a prosodic model. This is what we call a data-driven approach. The prosodic model describes the relationship between some linguistic features extracted from a text corpus and some prosodic features extracted from

a related speech corpus. Then, a TTS system relies on a prosodic model to generate the prosodic parameters.

The general scheme we propose for producing a data-driven prosodic model is shown in Figure 1. The input to the system is a monospeaker recorded prosodic corpus and its textual representation. Our system analyzes every sentence of the corpus.

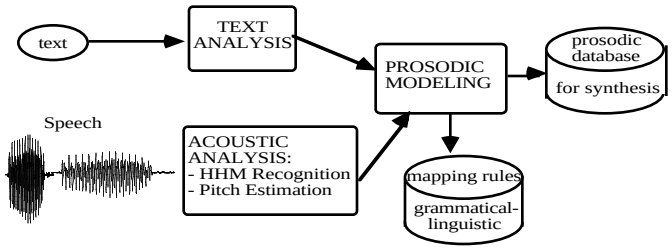

Figure 1: Data-driven methodology for prosodic modeling.

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). So syllables, which will be the basic units used in our prosodic modeling, are embedded into accent and breath groups. Syllables will be linguistically and prosodically distinguished.

The different steps of the proposed methodology are described in the next paragraphs of this Section.

2.1. Text Analysis

For every sentence of the corpus rule-based text analysis gives as results:

- grammatical categories of every word.
- accent groups in the sentence.
- phonetic transcription.
- vowel nucleus of each syllable.

At this point we have three features related to each vowel nucleus: name of the nucleus (the name of the vowel), place of the syllable in the accent group and distance of the syllable to the lexical accent.

2.2. Acoustic Analysis

In order to get some more linguistic features and some prosodic features we take the recorded speech related to the sentence we are analyzing and we perform an analysis of the speech signal. This analysis includes acoustic processing of the speech to obtain sound segmentation and pitch contour estimation.

Segmentation is made through an HMM-based automatic speech recognition system. This segmentation is refined with information from a spectral detection algorithm, including:

- adjustment of the segmentation between two vowels of similar energy;
- removal of the short silence usually detected before occlusive consonants;
- adjustment of the segmentation of silence segments.

The fundamental frequency estimation is based on the high-resolution algorithm described in [4]. This algorithm is modified to restrict the fundamental frequency range, and thresholds are adapted to the reference speaker.

As result of the acoustic analysis we obtain the following information:

- segmentation of sounds
- pauses in speaker's utterance.
- prosodic contour of each vowel nucleus. This contour is represented by 5 parameters (see Figure 2) 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).
- duration of the consonants.

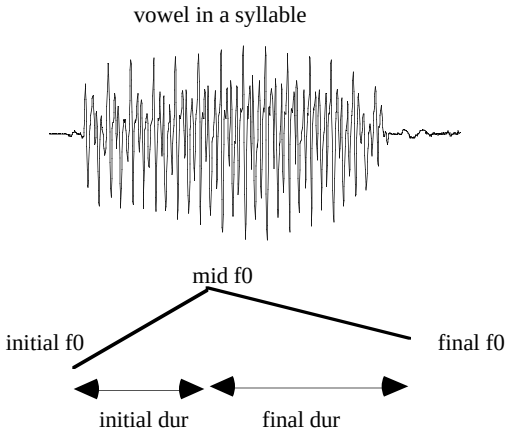

Figure 2: Prosodic contour in a syllable defined by 5 parameters to form the prosodic syllabic pattern (PSP).

2.3. Prosodic Modeling

In this step results from text and acoustic analysis are melted to obtain the prosodic model. We use the information about pauses position to compute breath groups. Next, we categorize accent groups according to their position into the computed breath groups: initial, inner or final position. Type of breath group is defined by the prosodic contour of the last vowel nucleus of the current breath group. We can already complete the set of linguistic features related to syllables adding two more features: type of accent group and type of breath group.

We also consider at this point the prosodic features of each syllable. These features are:

- the duration of the pause after the syllable (for pause modeling), if it is the last syllable preceding a silence.
- the rhyme lengthening and its difference to the syllable onset lengthening (both calculated as in [5] for consonant modeling)

- two features for vowel modeling: the vector representing the PSP of the current syllable (PSP1) and the pattern of the following syllable in the speech corpus (PSP2).

These five prosodic features plus the five linguistic features mentioned before contain relevant information for each syllable. A register containing this information is created for each syllable in the prosodic data-base shown in Figure 1. As process described for features extraction is performed on every sentence of the corpus, there will be as many registers in the database as syllables there are in the corpus.

In order to reduce the storage space in the prosodic database we observed the following strategy: five dimensional vectors representing PSPs were quantified (we have obtained good results in our applications with only 64 centroids), so we store just one number (centroid identification number) instead of storing the five components of the vector.

There is still one more aspect in this prosodic modeling stage. Let's imagine the system working on synthesis. For each input sentence we will access to the prosodic database to obtain information for every syllable. We have some information (text analysis information) from the linguistic analysis, but another important parameter to access database is still needed: the type of breath group of each syllable. In our system this parameter is derived from a set of rules inferred automatically from the training corpus: these rules make a mapping between grammatical categories and linguistic features derived from the acoustical analysis. The process to generate the rules is as follows:

First, 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, the following rule is generated:

{C30,C27,C42,C28,C32,C9} →

→ {	7	7	7	15	15	15	breath gr.
	yes ,	no ,	yes ,	no ,	yes ,	no }	accented
	no	no	yes	no	no	yes	pause

This mapping is made for every sentence in the corpus. 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. Then a process of selection of rules is performed to obtain the final set of rules. This process involves rule-length adjusting, and checking of contradicting rules. If two rules have the same antecedent (sequence of grammatical categories) and a different consequent (linguistic features), we preserve the one that appears more frequently.

Several criteria can be followed in this process. In the following section, we explain it for a task-adapted system and its difference for a general purpose system.

Therefore, prosody generation process to synthesize a text is as follows:

1. Grammatical parsing to obtain the categories of each word in a sentence.
2. Application of grammatical-linguistic mapping rules to derive the linguistic features .
3. Information retrieving from the prosodic database using linguistic features obtained in previous step (2) for synthetic prosody generation

It is important to note that the prosodic information extracted from the prosodic database is referred to each syllable. Concatenation of each prosodic pattern is necessary to create a natural prosodic contour. For this purpose every selected pattern must have some degree of relation with the previous one avoiding large variation of pitch values. This is achieved taking into account not only the pattern of the current syllable (PSP1) but also the pattern of the following syllable (PSP2). Therefore in synthesis, having two consecutive syllables A and B , we consider that their PSP1´s will produce a good concatenation, if the PSP2 of A is similar to PSP1 of B. More information about pattern selection can be found in [6].

3. SPEAKER & TASK ADAPTATION

According to the general prosodic modeling strategy described in the previous section, the prosodic database and the mapping rules, automatically generated, keep prosodic information from the speaker and syntactic structures found in the training corpus.

Our rule-generation procedure is quite flexible: maximum and minimum length of the rule can be fixed. This is important because we have 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 is obtained.

Another important advantage of our system is that grammatical parsing can be adapted to the task with minimum effort. 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 we have included specific grammatical categories for these words or syntactic structures that provide a more accurate prosodic modeling and improve the naturalness of the

synthetic speech. Note that once these categories have been included the whole procedure runs automatically.

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.

It is important to point out that, of course, building new voices in synthesizers by concatenation of speech units also requires the new speaker to generate an acoustic database (i.e. to extract diphones). Automatic generation of these speech units as in [5], although beyond the scope of this work, can be considered as a complementary research field towards the design of future TTS systems.

4. EXPERIMENTAL RESULTS

The proposed methodology has been tested in two different applications. In both cases we use a diphone-based speech synthesizer. We have only used the speech concatenation units of one speaker to isolate the improvements derived as result of the prosodic modeling, from those due to different sets of concatenation acoustic units.

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 two experiments for each application. One experiment consisted of prosodic modeling using grammatical categories designed for a general purpose TTS, while in the second experiment special grammatical categories were used. 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...).

A preliminary evaluation has been made over a population of 12 people . A set of 20 sentences were synthesized for both applications using three different prosodic models (these speech material can be accessed through our www address: www. gaps.ssr.upm.es/tts) :

1. Model A: general purpose prosodic model.
2. Model B: speaker and task prosodic adapted model with general grammatical categories.
3. Model C: speaker and task prosodic adapted model with specific grammatical categories.

The three synthetic realizations of each single sentence were randomly presented to the listeners, who had to sort them according to their subjective preferences. Table I represents the percentage each model was selected as the one providing the highest naturalness

Table I: Subjective evaluation results

	TRAINS	TELEPHONE
MODEL A	9.6%	7.5%
MODEL B	31.6%	38.3%
MODEL C	58.8%	54.2%

It can be seen 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. Differences between applications are probably due to the more complex structures found in the train information system, what can only be managed through the use of specific categories.(model B works better in the telephone operated case because structures are not so complex and can be properly managed with standard categories).

REFERENCES

1. 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.
2. F. Emerard et. al. "Prosodic processing in a TTS synthesis system using a database and learning procedures" in *Talking Machines: Theories, Models and Applications* Editors G. Bailly and C. Benoit. Elsevier 1992.
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. Y. Medan, et al "Super Resolution Pitch Determination of Speech Signals" in *IEEE Trans. on Signal Processing*, pp. 40-48. Volumen 39, Number 1. Jan. 1991.
5. C.W. Wightman and M. Ostendorf "Automatic Labelling of Prosodic Patterns" in *IEEE Trans. on Speech and Audio Processing* , pp. 469-481. Volumen 2, Number 4. Oct. 1994.
6. 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.
7. O. Böeffard, et al "Automatic segmentation and quality evaluation of speech unit inventories for concatenation-based multilingual PSOLA text-to-speech systems", *Proc.EUROSPEECH 93*,Berlin, pp.1449-1452.