Prosody Generation by Means of a Syntactic Approach and its Application in a Text to Speech System

Enzo Mumolo, *Massimo Teia*
Dipartimento di Elettrotecnica, Elettronica ed Informatica
Universita' di Trieste, Via Valerio 10, 34127 Trieste, Italy
Tel/Fax: +39.40.676.3861/3460
e-mail: mumolo@univ.trieste.it

ABSTRACT

An algorithm for modeling and generating prosody from a written text is described in this paper. Among the several speech processing areas which could benefit of this algorithm, in this paper we have dealt with text to speech synthesis (TTS). An experimental evaluation of the algorithm has been carried out and it has been shown that the naturalness of the produced speech has greatly improved.

1 Introduction

Since the intonation of a spoken message carries information about the meaning of the message, its determination would require techniques which pertain to the artificial intelligence area. However, even such techniques can fail and, actually, this a topic of current research. As a matter of fact, generally the algorithms for prosody modeling and generation are rule based [1]. In [6], for example, a system is described which derives an abstract phonological description of the linguistically relevant modulations of the fundamental frequency and then interprets these descriptions into pre-defined configurations of stylized 'pitch movements', derived with a perceptual approach [7].

In this paper, we propose a syntactical algorithm toward the automatic generation of the intonation of a written text. It is worth noting that, in this paper, we have considered only the fundamental frequency as prosodic parameter. Our main goal was to improve the naturalness of a TTS system without giving too much in complexity with respect to known algorithms. An experimental evaluation of this algorithm with a corpus of testing phrases has shown that a good quality speech with natural intonation is produced for the majority of the tested phrases. The idea behind the algorithm is that, since a real speech pitch profile reflects several events, including intonation and other microvariation, the information semantically relevant can be obtained by averaging profiles corresponding to several phrases with similar propositional structure but different from a phonetic point of view. In this way, the perceptual relevant information are maintained and emphatized while all the others are filtered out. This reasoning is somehow similar to that described in [7], but it is much simpler and suitable to be extended to other languages. In order to generate prosody for a general text, the profiles of a spoken message were roughly divided into a few number of elementary intonational groups. Such elementary groups, which will be called Basic Prosodic Groups (BPGs), are extracted from a corpus of utterances and were eventually used as a dictionary for the reconstruction of the intonation of an arbitrary written text. Moreover, the linguistic processing reduced to the extraction of the BPGs from the written text.

The main problem addressed in this paper concerns the segmentation of the input ASCII text into pieces which correspond to the elementary intonational profiles. An algorithm based on a syntactical analysis, is presented. Once this segmentation is obtained, the final profile can be obtained by concatenation. It is worth noting that the algorithm described in this paper was developed for the italian language.

The paper is organized as follows: in Section 2 the issues concerning the definition of BPGs are considered, and in Section 3 the main topic of this paper, namely the automatic segmentation of a written text into BPG is described. The final processing required to obtain the complete intonational profile and some concluding remarks are reported in Section 4.

2 Identification of the Basic Prosodic Groups

The so called BPGs constitute a model of the prosody. Basically, the goal of the BPGs is to capture the suprasegmental characteristics of the following parts of a spoken utterance: principal vs. secondary, initial vs. non initial and final, both for declarative and interrogative sentences. This division was perceptually motivated. The metodology used in this phase was the following: first, a set of twenty written sentences, equally divided between declarative and interrogative, were first segmented on the basis of their meaning and syntactical structure into the following categories:

- principal initial profile (PI)

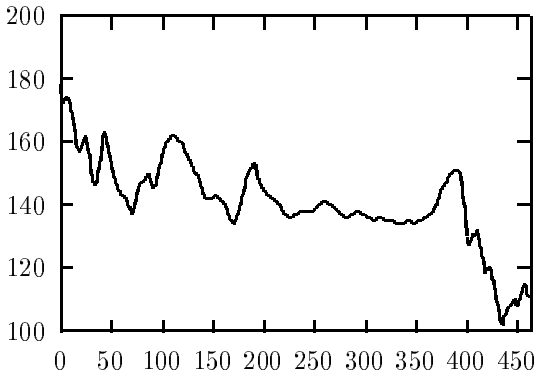

Figure 1: An example of the final BPG for declarative

- principal non initial profile (PNI)
- secondary initial profile (SI)
- secondary non initial profile (SNI)
- final profile (FP)

for the declarative and

- principal initial profile (PIN)
- secondary non initial profile (SNIN)
- final profile (FPIN)

for the interrogative sentences. Then, the phrases were read aloud by a trained speaker accordingly to the above described categorization, sampled and stored in a digital computer. A pitch analysis, followed by a polynomial interpolation in the unvoiced sections, was then performed. The pitch profiles obtained so far were therefore represented by continuous curves. Afterward, the individual pitch frequency profiles corresponding to each BPG were extracted using an interactive tool that allowed for a selective listening of the signal. At this point, a set of profiles for each BPG is available. These profiles are averaged together with a dynamic programming algorithm. As a result of the averaging, a prototype curve for each BPG is obtained; the micro prosodic and idiosyncratic events of each utterance are filtered out by means of averaging, and the linguistic maningful structures which are shared in the utterances, are obtained. The averaged curves, therefore, contain predominantely the peaks which are characteristic to the selected group itself [2]. The profile of a final BPG for declarative sentences is reported in Fig.1, where the f0 values are expressed in Hertz.

3 Automatic segmentation of the input text into BPGs with a syntactical approach

In this Section we describe the identification of the BPGs from an input text. The general linguistic framework for the pre-processing and syntactical analysis phases has been derived from the work of [3].

3.1 Pre-processing of the input text

Consider the sets of symbols defined in the following

Definition 1 *Let us define a set of symbols as* $T_1 = \{gv, gn, gnv, gp, gpv, ga, gav, fr, frv, fs, fsv\}$ *where "g" stands for group, "n" for nominal, "v" stands for verbal, "p" for prepositional, "a" for adjunt, "f" for phrase, "r" for relative and "s" stands for secondary. Define moreover* $T_2 = \{< . >, <, >, <:>\}$ *and* $N_2 = \{< PF0 >, < PF10 >, < PF20 >, < PF30 >, < PF40 >, < PF50 >, < PF60 >, < PF70 >, < PF80 >\}$.

The output of the pre-processing module is a sequence of elements of the set T_1. The first stage toward this goal is to perform a lexical analysis (LA) of the input text. The purpose of the LA is to give the follow-up stages the suitable information on each word of the input text. The lexical analyzer can be built at different levels of complexity but, at least, its output should include the word's grammatical category. Moreover, in case of lexical ambiguities, the LA should furnish all the possible categories. According to [3], we developed an LA composed by a dictionary and an automata. The dictionary contains all the lexicon, using a morphemic approach, while the automata's task is to make a morphological analysis of the input text and to give, as output, the grammatical categories. The dictionary is structured as a graph, which is analyzed with a deph-first strategy. The automata is, as a matter of fact, an Augmented Transition Network (ATN) [5]. The next step of the pre-processing phase concerns the stress assignement. Clearly, stress assignement is fundamental for the correct generation and modeling of the prosody. According to [3], the stress is obtained with a list of rules together with a number of exceptions. The list is searched and as soon as a suitable rule is found, the algorithm ends its task. It is important to note that, besides the main stress assigement, also two additional types kinds of stress, named "secondary" and "reduced", are assigned.

3.2 Syntactical analysis

The goal of the syntactical analysis is to determine high-level soprasegmental characteristics of the sentence, such as the "breath groups" and the prosodic profiles. Moreover, the syntactical information are fundamental to solve ambiguities, or homographs. The input text is divided into "Phonological Words" (PW), defined as a phonetic unit composed by the subset of words which is uttered as it were a single word. The PW contains only one principal stress. The PWs are classified in terms of 'gv' (verbal group), 'gn' (noun group), 'gp' (prepositional group), 'ga' (adjunt group), 'fr' (relative phrase) and 'fs' (secondary phrase).

3.3 Determination of the BPGs in the input text

As introduced before, the meta-text produced so far has to be segmented into BPGs, defined in Section 2. The

segmentation has been realized with a grammar which classify a set of PW as the appropriate BPG according to the structure $[< nominal\ group >] < verbal\ group > [< nominal\ group >] < prepositional\ group >^*$.

We now describe the syntax for the declarative phrases, which has been derived through the analysis of many sentences. Slight modifications are required for the interrogative phrases.

Definition 2 *The grammar used for the segmentation into BPG is a set* $G = \{V_N, V_T, P, S\}$ *where the non-terminal symbols are* $V_N = \{< s_i >, < PF_j >, < p >, < FALL >\}$, *for* $i = 20, 40$ *and* $j = 0, 10 \ldots 80$ *and the set of terminal symbols is* $V_T = T_1 \cup \{PI, PNI, SI, SNI, FP\} \cup \{:, ., , , \{, \}\}$. *The initial symbol is* S. *The set of production rules* P *is given below in BNF.*

1. $S ::= (\{(gav|ga|fs|fsv|gp|gpv|frv|fr) < PF0 >< SI >< s20 >)|(\{(gn|gv|gnv) < PF0 >< PI >< s40 >)$
2. $p ::= , | :$
3. $fr < s20 > ::= fr < PF0 >< PSNI >< s20 >$, same for $fsv, frv, fs, ga, gav, gp, gpv$
4. $gn < s20 > ::= gn < PF0 >< PNI >< s40 >$, same for gv, gnv
5. $< p >< s20 > ::= \{< p > (t, t \in T1) < PF0 >< SNI >< s20 >$
6. $< s40 > ::= \{(t, t \in T1 \text{ or } T2) < PF0 >< SNI >< s40 >$
7. $. < FALL > ::= .\}$
8. $.\}(< PSI >< s20 > | < PI >< s40 > | < SNI >< s20 > | < SNI >< s40 > | < PNI >< s40 >) ::= .\} < PF >$
9. $< p >< FALL > ::= < p >\}$
10. $gv < PF0 > ::= gv < PF10 >$
11. $gpv < PF0 > ::= gpv < PF20 >$, same for fsv, frv, gav
12. $gn < PF0 > ::= gn < PF30 >$, same for fr, fs, ga, gp
13. $< PF10 > ::= (gv < PF10 >)|(gp < PF70 >)|((gn|gnv|gpv|ga|gav|fr|frvfs|fsv) < PF40 >)$
14. $< PF20 > ::= (gv < PF10 >)|(fsv < PF20 >)|((gn|gnv|ga|gav) < PF40 >)|(gp < PF70 >)|((sf|fr|gpv|frv) < PF0 >< SNI >< PF70 >))$
15. $< PF30 > ::= ((ga|gn|gp) < PF30 >)|((gv|gav|gnv) < PF40 >)|((gpv|fr|frvfs|fsv) < PF0 >< SNI >< PF10 >))$
16. $< PF40 > ::= ((gv|gav|gn) < PF40 >)|(gp < PF50 >)|(\{\})$
17. $< PF50 > ::= ((gv|gav) < PF60 >)|(gp < PF50 >)|(\{\})$
18. $< PF60 > ::= ((ga|gp|gav) < PF60 >)|(\{\})$
19. $< PF70 > ::= ((ga|gp|gav) < PF70 >)|(gn < PF80 >)|(\{\})$
20. $< PF80 > ::= ((ga|gav) < PF60 >)|(\{\})$
21. $(n, n \in N2 - \{< PFO >\})|(. < PF0 >) ::= . < FALL >$
22. $(n, n \in N2 - \{< PFO >\})|(< p >< PF0 >) ::= < p >< FALL >$

A parser has been realized by means of a small ATN built according to the described grammar.

Example

Let us consider, as an example, the following italian phrase: '*Il satellite contiene un orologio ed una stazione trasmittente che permettono di sincronizzare i tempi con grande precisione.*'.This phrase is transformed, at the output of the syntactical analysis of Section 3.2, in the following string: '*gn gv gn frv gp gn gp.*' which is left-to-right analyzed with the grammar described above. The results of the analysis are reported below, where the employed production rules and the output strings are indicated.

$P1, P12 \rightarrow \{gn\ PF30\ PI\ s40$
$P15, P16 \rightarrow \{gn\ gv\ gn\ PF40\ PI\ s40$
$P16, P6 \rightarrow \{gn\ gv\ gn\}PI\{frv\ PF0\ SNI\ s40$
$P19, P20 \rightarrow \{gn\ gv\ gn\}PI\{frv\ gp\ gn\}SNI\ s40$
$P6, P8 \rightarrow \{gn\ gv\ gn\}PI\{frv\ gp\ gn\}SNI\{gp.\}PFIN$

which is the final result.

4 Final processing

The result obtained so far contain information concerning the segmentation of the input text in terms of BPG with the associated timing. Additional processing is required to produce the final profile and to align it to the stressed syllables.

4.1 Concatenation of the BPGs

The suitable BPGs are then concatenated and a final prosodic f0 profile is therefore obtained. The discontinuities between different BPG are maintained, as they correspond to an actual phonetic discontinuity.

4.2 Alignement of the prosodic profile to the utterance

The final step is to adapt the prosodic profile to the utterance actually generated. The actual speech synthesis was performed using a text-to-speech system based on segment concatenation [4]. First, the relative maxima of the profile are selected. The selected points are then classified in 5 levels according to the value of the fundamental frequency (lower level correspond to higher f0 values) and such that the distribution of each point is somehow uniform. The classification of the points into levels is needed to avoid a major problem of the adaptation procedure, which may arise if the principal stress of the PW were located toward the end while the biggest peak of the profile is at the beginning. In this case, the profile would be warped in an unreasonable way. Each labeled peak is then associated to a stress point in the phrase using an iterative approach. Namely, the principal stresses are first considered, and the profile is divided into a number of intervals equal to the number of stress points. In each interval, the lower level peaks are chosen and, among them, the one which has the maximum f0 value is finally chosen and assigned to the corresponding stress. This procedure is repeated for the secondary and reduced stress. Once the peaks are assigned to each stress point, the prosodic curve is nonlinearly modified in order to get a one-to-one correspondence between the samples of the signal and the points of the prosodic profile, using the algorithm described below. Let us first define the following symbols:

- NSP = number of stress points
- NSA_i = number of the frame corresponding to the i-th stress
- NP_i = sample number of the i-th peak
- N = lenght of the interval to be distorted (in frames)
- $\tilde{N}$ = length of the final distorted interval (in frames)
- $f0i_n$ = n-th f0 value of the profile to be distorted
- $f0f_n$ = n-th f0 value of the final distorted profile

The pseudocode of the algorithm is

For i=0 to $(NSP + 1)$ Do /* for each stress point */
    $N = (NP_{i+1} - NP_i)$;
    $\tilde{N} = (NSA_{i+1} - NSA_i)$;
    For k = 1 to $\tilde{N}$ Do
        $n = \lfloor (k-1)(N-1)/(\tilde{N}-1) + 1 \rfloor$;
        $j = (k-1)(N-1)/(\tilde{N}-1) + 1 - n$;
        $f0f_k = (1-j) \cdot f0i_{NP_i+n} + j \cdot f0i_{NP_i+n+1}$;
    Od;
Od;

In Fig.2 the f0 profile aligned to the example phrase of Section 3.3 is shown.

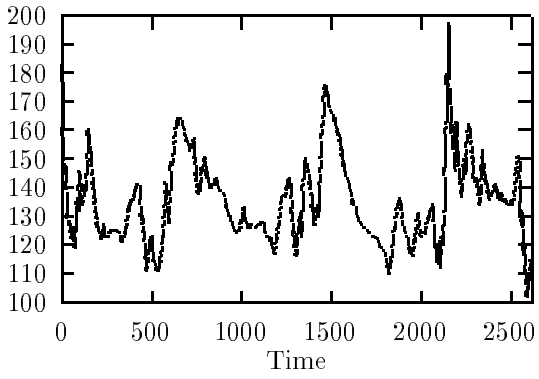

Figure 2: The final aligned prosodic profile for the sentence described in the example

5 Final remarks and Conclusions

The segmentation algorithm has been tested over twenty complex sentences and it gave perfectly correct results with a 50% rate, while in additional 40% it gave acceptable results. Clearly, a refinement of the morphological analyzer would lead to an overall improvement. The algorithm has been integrated into a real time TTS system.

References

[1] J. Allen, M. Hunnicutt, D. Klatt, "From text to speech: the MITalk system", *Cambridge University Press*, 1987

[2] L.Tonelli, E.Mumolo, P.Martini, "Proposta per un Modello dell'Intonazione dell'Italiano", *Atti delle Giornate di Studio del Gruppo di Fonetica Sperimentale*, 11-12 November 1993, Torino, Italy

[3] R.Gretter, G.Mian, R.Rinaldo, M.Salmasi, "Linguistic Processing for an Italian Text-to-Speech System", in Proc. of *Int.Conf. on Speech Tech. VERBA90*, 1990, Rome, Italy

[4] G.Abbattista, E.Mumolo, "High quality real time text to speech system in italian language", in Proc. of *Int.Conf. on Speech Tech. VERBA90*, 1990, Rome, Italy

[5] W.A.Woods, "Cascaded ATN Grammars", *American Jour. of Comp. Linguistics*, Vol.16, N.1, 1980

[6] S.Quazza, P.L.Salza, S.Sandri, A.Spini, "Prosodic control in a text to speech system for italian", in Proc. of *ESCA Workshop on Prosody*, Lund, Sept. 1993

[7] J.t'Hart, R.Collier, A.Cohen, "A perceptual study of intonation: an experimental-phonetic approach to speech melody", *Cambridge University Press*, 1990