TEXT-TO-PROSODY PARSING IN AN ITALIAN SYNTHESIZER. RECENT IMPROVEMENTS

Barbara Gili Fivela
Scuola Normale Superiore - P.zza dei Cavalieri, 7 - 56126 Pisa, Italy
e-mail:gili@alphalinguistica.sns.it
Silvia Quazza
CSELT - Via Reiss Romoli, 274 -10148 Torino, Italy
e-mail: silvia.quazza@cselt.it

ABSTRACT

This paper describes recent improvements of a Prosodic Analyzer, designed to provide the CSELT Italian text-to-speech system ELOQUENS® with a better handling of prosody. Based on lexical tagging, the Analyzer builds up the prosodic structure of the sentence, inserting proper prosodic markers at word boundaries. The approach belongs to the family of TTS-oriented, 'heuristic' strategies, inferring prosody directly from the building blocks of syntax and exploiting lexical and rhythmical language-dependent information. Latest improvements concern the linguistic knowledge base of the Analyzer, which has been enhanced and optimized. A formal evaluation of the Analyzer's performances is also presented in the paper.

1. INTRODUCTION

Current Text-to-Speech systems can't be expected to read aloud every text with appropriate - although stylized - prosody, as they can't deal with focus, semantic relations, intentions, situations, on which prosody mainly depends. What is commonly accepted is that a compromise can be reached, obtaining richer prosody where texts are controlled, and defining a sort of 'neutral', 'syntax-driven', 'few-breaks' style for unrestricted text synthesis. In fact, the artificial generation of prosody can be split in two different aspects, the acoustic reproduction of natural-sounding prosodic patterns and the automatic assignment of the appropriate patterns to text. An accurate simulation of patterns and a wide repertoire of contours, suitable to different styles and contexts, can make synthetic speech sound much more natural and expressive, even more intelligible, provided that you get the right pattern in the right place. As in many applications you can prepare text in advance or you can generate it (e.g. dialogue systems), most TTS systems allow the insertion of explicit markers in the text, activating the desired patterns. But for any application requiring unrestricted text synthesis (reading for the blind, telephone access to information/ e-mail/news, etc.) prosodic patterns should be assigned to text automatically. This is a demanding task, because text often offers only some broad explicit prosodic markers (punctuation) and very few cues to prosodic nuances. Besides, errors in locating a prosodic break or in choosing the type of pattern may have a heavy negative impact on naturalness and may also affect intelligibility. For this reason, a low-risk strategy should be adopted, giving up the goal of expressive reading and keeping to a more neutral - but still plausible - style. The knowledge on which to rely to fill the gap between text and prosody - given that semantics and knowledge of the world are far beyond the reach of TTS systems - is primarily lexical (or lexico-semantic [1]) and syntactic. On such basis, prosody can be generated through full syntactic parsing, or by statistical rules, or can be inferred directly from the building blocks of syntax, exploiting any language-specific rhythmical or lexical cue easy to extract from a written text ([1,2,3]).

The approach described in this paper follows this last 'heuristic' strategy. Our *Prosodic Analyzer* [4] has been implemented in a laboratory version of the ELOQUENS® Italian text-to-speech system, developed at CSELT [5], with the aim of obtainig a richer prosodic phrasing for 'neutral style' reading. The standard version of ELOQUENS® was already able to associate a limited number of prosodic patterns to punctuation and to locate some other major breaks. To improve on such a simple behaviour while complying with the above mentioned low-risk requirement is not trivial, due to uncomplete grammatical tagging, syntactic ambiguity, non-determinism of syntax-to-prosody relations. Even the evaluation of prosodic performances is a difficult task. We will discuss these points in the following, after describing the algorithmic structure of the Prosodic Analyzer (par.2) and its linguistic knowledge-base (par.3). Finally, we will present an evaluation of its performances, compared with those of human readers, of the standard ELOQUENS® version [6] and of a previous version [4] of the Analyzer (par.4).

2. THE ALGORITHM

The text-to-prosody process covers all the linguistic phases of text-to-speech synthesis, from grapheme processing to computation of phoneme duration, energy and pitch. The Prosodic Analyzer here described performs the central step of the process, where a sequence of lexical words tagged with grammatical categories is organized into an abstract *prosodic structure* whose constituents are associated with specific

	se	ti	manderanno	subito	le	ultime	cartoline	ce	la	farai,	mi	sembra
lexicalword	CLI	CLI			CLI			CLI	CLI		CLI	
clitic group			CLG			CLG						
phonol.phr.				PPH								
interm.phr.							WSC					
intonat.phr.										SSC		
sentence												DEC

Fig.1 *Example of prosodic structure of an Italian sentence (English translation: "*if they send you immediately the last cards you will succeed, I think*"). The right boundary of each phrase is marked by a prosodic label: CLI=clitic, CLG=clitic group, PPH=phonological phrase, WSC=weak suspensive contour, SSC=strong suspensive, DEC=declarative.*

prosodic patterns, later converted into actual prosodic values. The adopted strategy is to analyse words between punctuation marks, identify *syntactic* building blocks on the sole basis of word grammatical tags, then trying to detect minimal *syntactic-prosodic* phrases, delimited by possible candidates for prosodic breaking, with different degrees of strength. Finally, choosing the actual breaks according to rhythmical constraints, trying to balance the length of the resulting prosodic phrases.

2.1. Input and output

Input and output of our Prosodic Analyzer (PA) are defined by the Lexical and Phonetic Modules of the ELOQUENS® system. The *Lexical Module* identifies words and sentences, expands numbers and abbreviations, performs grammatical tagging (relying on a small lexicon and on rules automatically derived from a large database [6]) and provides PA with a single sentence at a time, represented as a *sequence of tagged words* and punctuation marks. The task for PA is to label each word according to its stress degree and to the prosodic type of its right boundary. Depending on the boundary type, the *Phonetic Module* will decide whether to join the word to the following one, as in the case of clitics, or to realize it with a slight final lengthening, or with a specific intonation contour or breath pause marking the end of an intonational phrase. *Prosodic boundaries* can indeed be considered the linear projection of the different layers of the prosodic structure of the sentence. ELOQUENS® distinguishes the layers of clitic group, phonological, intermediate and intonational phrases [7], providing markers for each level and a few alternative contours for intonational phrases (see Fig.1).

2.2. Grouping words into phrases

In order to locate *potential* prosodic breaks, PA performs a minimal syntactic analysis grouping together adjacent words into Phrases (more or less corresponding to phonological phrases), each labeled with its *syntactic type* and the *depth* of its right boundary, i.e. a measure of the syntactic gap with the following Phrase. Depth values range from 1 (mere possibility of a weak prosodic break) to 4 (sure presence of a strong one).

The analysis is performed in three steps, the first based on word-class sequences and the others on short-distance syntactic relations. Steps are computationally efficient, scanning the input stream left-to-right between punctuation marks, with no recursion nor backtracking.

The first step builds up *Word Groups* by joining together words that surely won't be separated by a prosodic break. The analysis is performed by a *transition network* scanning left-to-right the input sequence of grammatically tagged words. The network accepts each successive word tag and decides, depending on the current state, whether or not it can be attached to the previous ones in a single Word Group.

The second step attaches Word Groups into larger *Phrases*, basic constituents of the prosodic structure. Again, the algorithm acts sequentially, deciding for each Word Group whether it should be attached to the current Phrase. Decision is based on local-context *Rule Tables* and on *flags* keeping track of some wider syntactic phenomena. Each generated Phrase is described by a pair <T, D> where T is its syntactic type and D is the depth of its right boundary. When a Phrase is defined, T and D values depend on the current Word Group. For each successive Word Group with label L, the *Rule Table* is searched for the entry <T, L> stating whether or not to attach the Word Group with label L to a Phrase of type T, and possibly providing a new value <T', D'> for the current Phrase. If the Word Group is not to be attached, the current Phrase is terminated and a new Phrase is initiated.

The third step has been recently added in order to deal with coordination. Phrases are scanned left-to-right looking for those labelled with special *coordination tags* marking the Phrases followed by a conjunction. A *Coordination Rule Table* is then looked up, stating the depth to be assigned to such Phrases depending on their type and on the type of the following Phrases.

2.3. Inserting breaks by rhythmical analysis

In order to build up a balanced prosodic structure, a decision should be made as to which of the potential

boundaries should actually be converted into a *prosodic break*, marked by a specific prosodic pattern. Phonological, intermediate and intonational phrases are built up on the basis of the <T, D> labels associated with the minimal syntactic-prosodic Phrases, trying to comply with the following simple rhythmical constraint: prosodic constituents of the same depth should be balanced in length, i.e. they should have, as far as it is allowed by syntax, an equal number of graphemes.
To reach this goal, PA first inserts strong break markers (corresponding to intonational phrases) at the end of Phrases <T, D>, where D = 4. The type of the inserted prosodic marker will depend on T. Then, for each stretch of sentence delimited by the inserted strong markers, PA applies the following procedure:
1. the stretch of sentence is divided into the smallest tracts of equal length L such that L is greater than a given threshold and each tract contains at least one Phrase <T, D> with D=3
2. for each tract: if it contains a single Phrase with D=3, a marker is inserted, whose type, corresponding to an intonational or intermediate phrase, depends on T and on the Phrase's length; if it contains at least two Phrases with D=3, the Phrases are pairwise compared and, depending on their types T1 and T2, it is decided which one should correspond to a marker.
Finally, for each weak boundary, if its distance from the adjacent inserted markers is greater than a given threshold, a prosodic marker is inserted, of type WSC (intermediate phrase, see Fig.1) if the boundary depth is D=2, and of type PPH(phonological phrase) if it is D=1.

3. THE LINGUISTIC KNOWLEDGE

The success of syntax-driven heuristic prosodic phrasing highly depends on the accurateness of its knowledge-base, where lexical, grammatical and rhythmical language-dependent knowledge is formally stated. Any linguistic cue to prosody should be exploited, with special care reserved for function words, which have a key-role both for syntax and for prosodic-specific phrasing. It should be noted that syntactic and prosodic trees are not isomorphic (cfr. *[I eat the cake]*$_{PP}$ *[that you prepared]*$_{PP}$ and *[I eat [the cake [that you prepared]*$_{SP}$*]*$_{SP}$*]*$_{SP}$). Our Prosodic Analyzer takes into account this fact, avoiding attachment of relative clauses to NounPhrases, embedding clause-introductory particles (conjunctions, pronouns) in the clause-intial Phrase, strictly limiting recursion (e.g. allowing a single embedded phrase in Verb Phrases). A careful handling of function words may prevent highly undesirable misplacements of prosodic breaks, as pointed out by diagnostic tests of a former version of PA [4], which suggested a new special treatment of coordination and prepositional attachment.
The grammatical tagging on which PA relies is quite detailed about function words, correctly labels most verbs and by default marks the remaining words as *Noun-or-Adjective*'s (or *Homograph*'s, if they can be both nouns and verbs). The ELOQUENS® Lexical Module is being improved in order to reduce misclassification and further enhance the set of lexical categories. Grammatical tags are the only input provided to the transition network building up Word Groups. The linguistic criteria for Word Group construction aim at keeping together strongly connected words (e.g. a pronoun with the following auxiliary and past participle) and at isolating some key function words which can provide crucial information for the next phrasing steps. For example, while an article will form a Noun Group together with a following sequence of Noun-or-Adjective's, conjunctions will form Word Groups on their own. So, beside the main labels *NounGroup*, *VerbGroup*, the analysis will ouput labels such as *Conj_Coord*, *Conj-Subord*, *Conj&Prep*, or, for relative pronouns or locutions, *RelPron*, *Prep&Rel*. Recent improvements of PA have optimized the classification of function words, keeping only those details actually useful in the next steps (e.g. Word Groups for prepositions have been reduced from 16 to 3, *A* ('to'), *Di* ('of') and a single class for the others).
The detailed labeling of Word Groups allows to build up Prosodic Phrases with a small amount of look-ahead. Inspection of the current Word Group label and possibly of the following one is all what is needed to decide whether to include it in the Phrase under construction, except for a few flags keeping track of position in sentence, subordination, extraposition and parenthesis. The resulting Phrases correspond to a low level in the prosodic tree - phonological phrases - but provide the following rhythmical analysis with information on their attachment into intermediate and intonational phrases, expressed by their label and potential depth. Phrase labels distinguish between *NounP*, *VerbP*, *PrepositionalP* and *Preposition+VerbP* and explicitly state some features concerning the Phrase inner structure, e.g. the presence of a *direct object* inside a Verb Phrase or the kind of *preposition* introducing a Prepositional Phrase, and some higher level structural information, e.g. *embedding* in dependent clauses.
Coordination, where a larger amount of look ahead is necessary to delimit the conjuncts, is now dealt with by a further step (see par.2.2), allowing to scan the already formed Phrases. When a Phrase is marked as *followed by conjunction* also the following Phrases are inspected to decide on the complexity degree of the two conjuncts. On that basis Phrase depth can be altered, imposing a strong prosodic break where complex Phrases are coordinated.

4. PERFORMANCE EVALUATION

Evaluation criteria for prosodic phrasing aren't obvious. Subjective listening would probably be the best way of judging synthetic prosody. But a diagnostic test assessing the specific contribution of phrasing to the overall prosodic performance of a TTS system is difficult to define. To avoid any interference by actual acoustic implementations of the different prosodic patterns, one should directly evaluate the written output of the Prosodic Analyzer, where prosodic breaks are marked by labels. But there is no unique correct phrasing for a given text. Many prosodic breaks are indeed optional and a speaker is free to choose where and how to mark minor boundaries, provided it uses a coherent strategy. The adopted compromise towards an 'objective test' has been for us to compare the output of PA with the *human behaviour* averaged on a few speakers. We defined the following types of possible *errors* by PA:

1. Insertion error. Insertion of a marker where none of the speakers made a prosodic break.

2. Deletion error. Lack of markers where all the speakers made a break.

3. Substitution error. 3.1 Insertion of a different marker where all the speakers inserted an identical marker. 3.2 Insertion of a strong marker where some speakers didn't make a break and the others realized a weak break.

The speakers' behaviour was recorded in a prosodic transcription of their readings, distinguishing four types of breaks, a weak break (intermediate phrase, with no pause) and three strong breaks (two for parenthesis). Such transcriptions were then compared with the automatic phrasing by means of a software realized for that purpose. A first test was performed on a former version of PA [4], resulting in a useful diagnostic. The new improved version of PA was then compared with the old one and with the standard ELOQUENS®version, by means of a new similar test.

Two written Italian texts were chosen, the first (the same used in the former test) from a linguistics treatise and the second from a novel, amounting to 2186 words in total. The texts were read aloud by four naïve but educated Italian speakers. Table 1 summarizes the behaviours of the four naïve speakers NS1, NS2, NS3, NS4, averaged in the "Average" row, and of the three automatic systems PA, old PA (OPA) and standard ELOQUENS® (ELQ). The number of inserted breaks should be compared with the 2186 candidate positions (word boundaries) and with the 238 punctuation marks (plus 5 quotation marks, ignored by the automatic systems), where insertion probability is maximal.

Speaker	*total*	*strong*	*weak*	*no punct.*
NS1	469	326	143	231
NS2	467	324	143	229
NS3	465	334	131	227
NS4	447	315	132	209
Average	462	324.5	140.25	219
PA	454	349	104	216
OPA	445	346	98	207
ELQ	402	402	0	164

Table 1 *Number of prosodic breaks realized by speakers, new (PA) and old (OPA) Prosodic Analyzers, Eloquens (ELQ)*

Table 2 reports number and percentage of errors made by the three systems, classified according to the above described error types. Percentages are computed with respect to the 'relevant positions', depending on the error class (e.g. for class 1: positions where all the speakers made a break; pos. for type 2 include 3.1 pos.).

error class	*pos.*	PA #	PA %	OPA #	OPA %	ELQ #	ELQ %
1 (insert.)	1656	38	2.29	49	2.95	48	2.89
2 (delet.)	399	27	6.76	50	12.53	79	19.79
3.1(sub.)	320	53	16.56	57	17.81	86	26.87
3.2(sub.)	131	9	6.87	11	8.39	26	19.84
Total	2186	127	5.80	167	7.63	239	10.93

Table 2 *Number (#) and percentage (%) of errors in break insertion by PA, OPA and ELQ*

5. CONCLUSIONS

The error rate in prosodic phrasing has decreased from 10.93% (ELQ) and 7.63% (OPA) to 5.8% in the current version. PA turns out to be both more reliable - insertion errors reduced from 49 to 38 - and more effective, with 216 total inserted markers against 207 by OPA and 164 by ELQ, resulting in a much lower percentage of deletion errors. The special treatment of coordination significantly improved performances. We believe that exploiting every piece of information carried by function words can be a good strategy for heuristic prosodic parsing. Lexical tagging should be more reliable and detailed, recognize locutions and, may be, enhance grammatical tags with lexico-semantic information [1]. Automatic phrasing will still be poorer than natural, as a low-risk strategy imposes not to insert breaks where even syntax is ambiguos, as in the case of prepositional attachment. But, hopefully, further enhancement of the linguistic knowledge base could take Text-to-speech closer to the goal of plausible neutral-style prosody.

REFERENCES

[1] A. Lindstrom et al.,"*Generating Prosodic Structure for Restricted and Unrestricted texts*",ICPhS'95, Stockholm, 1995

[2] D. O'Shaughnessy,"*Specifying intonation in a text-to-speech system using only a small dictionary*",ICASSP '87, Dallas, 1987

[3] F. Emerard, et al.,"*Prosodic processing in a text-to-speech synthesis system using a database and learning procedures*", Talking Machines, G. Bailly, C. Benoit (Eds.), North Holland, Amsterdam, 1992

[4] B. Gili Fivela, S. Quazza, "*A Prosodic Parser for an Italian Text-to-Speech System*",Proc. XII SEPLN Congr. (Soc. Espanola para el Proces. del Lenguaje Natural), Sevilla, 1996.

[5] M. Balestri, et al., "*The Cselt System for Italian Text-to-Speech Synthesis*", Proc. EUROSPEECH '93, Berlin, 1993.

[6] S. Quazza, et al., "*Prosodic Control in a Text-to-Speech System for Italian*", ESCA Workshop Prosody, Lund, 1993.

[7] M. Nespor, I. Vogel, *Prosodic Phonology.* Foris Publications, Dordrecht, 1986.