



SYNTACTIC PRE-PROCESSING FOR HIGH QUALITY TEXT-TO-SPEECH

S. Quazza, G. Varese, E. Vivalda

Ing. C. Olivetti & C. S.p.A.
D.O.R. Speech & Language Laboratory
C.so Svizzera 185, 10149 Torino, Italy

ABSTRACT

To the end of improving the naturalness of the intonation generated by a text-to-speech system for Italian, a syntactico-prosodic pre-processor has been devised, inserting prosodic markers at syntactic boundaries in the input text. The relations between syntax and prosody, to be exploited by the system, are being investigated on a speech data base obtained by digital recording of a carefully designed corpus of sentences read by a professional speaker. First results concern the duration and position of breathing pauses, correlated to sentence syntactic structure. A prototype of the syntactico-prosodic parser has been implemented and tested on written texts.

1. INTRODUCTION

High-quality text-to-speech systems are able to read aloud any written text with a satisfying degree of intelligibility. They fail to sound "human" mostly because of their dull or unnatural intonation. Indeed, intonation is the acoustic correlate of high-level linguistic functions; variations in phoneme duration and fundamental frequency convey focus, boundaries between meaning units, hierarchical relations between concepts, theme/rheme distinction in discourse structure, extra-textual messages about speaker's emotions and judgements. As all what a text-to-speech system can exploit must be traced inside the written text, the problem is to identify the lowest level of textual analysis providing the richest information relevant to intonation.

The text-to-speech system for the Italian language, developed at the Olivetti Speech and Language Laboratory, at present generates prosody on the basis of a limited linguistic analysis of the input text (1). Its prosodic module takes as input the phonetic transcription of the original input text, marked with stresses and punctuation, and assigns to each phoneme its duration and its pitch value. Punctuation should be present in the original text, while stress marks are inserted by the linguistic module, which solves stress ambiguities and deaccentuates function words relying on local contextual syntactic analysis (2). The resulting intonation is not satisfactory. Punctuation is too rare in texts to be the only cue to prosody. The possibility has been provided to the user of enriching the

text with prosodic markers, corresponding to intonation contours which are not represented by punctuation: parenthetical, emphatic, suspensive, continuative, list contour. But a more systematic approach to prosody is in order. Left aside implementation problems (such as memory size and real time requirements) a general solution should be figured out.

The experimental work described in the following was undertaken to investigate the text-to-prosody question, having in view as a first concrete goal the realization of a pre-processing system automatically inserting prosodic markers on the basis of a linguistic analysis of text. The hypothesis leading the work is that the prosodic structure of sentences can be to a large extent traced back to their syntactic structure. Such an assumption is being verified by analyzing human speech (par. 4), while the abilities of an automatic syntactic parser are being tested (par. 5) to the end of implementing a syntactico-prosodic preprocessor.

2. THEORETICAL BACKGROUND

The existence of a syntax-to-prosody mapping has been debated in the context of recent studies concerning suprasegmental aspects of phonology, especially with respect to the English language. A view of the question, developed in the framework of generative phonology (3) but largely shared in its most general assumptions, is reported in the following.

The elements of the phonological representation of sentences are arranged in two distinct sorts of hierarchical organization: prosodic constituent structure and rhythmic structure, both determined by the regular occurrence of silences, accents (fundamental frequency peaks together with vowel lengthening), and intonation contours (typical pitch movements). The units of the prosodic structure are: syllables, prosodic words (i.e. syllable sequences with a single accent peak), intonational phrases (i.e. prosodic word sequences characterized by final lengthening, possible final silence, intonation contour on the last prosodic word). Their correspondence with syntax is strict, at least in read sentences, where the logical structure of discourse is mediated by the rigorous syntax typical of the written text rather than being directly expressed by prosody together with speaker's emotions and pragmatical intentions. A prosodic word corresponds to one or more

lexical words, generally one stressed content word, preceded by deaccentuated function words. An intonational phrase corresponds to a "sense unit" which in turn may correspond to a syntactic phrase or clause, the nature of the intonation contour and the duration of final syllables and final silence depending on the depth of the syntactic boundary, i.e. on the position of the phrase in the syntactic tree. The soundness of such a theoretical framework with respect to the Italian language has been partially verified by experiments reported in the literature, aiming to answer specific debated questions.

The work here described is an attempt to collect systematic data about the relations between syntax and phonology in Italian, within a text-to-speech oriented approach. The work plan consists of the following steps: digitally recording human speech; segmenting it into phonetic units and extracting fundamental frequency; correlating pitch and duration values with different levels of linguistic description (phonetic, lexical, grammatical, syntactical, rhythmical). As we are not concerned with spontaneous speech, but rather with the emulation of human correct reading of written texts, the choice has been to build up a corpus of syntactically representative sentences to be read by a professional speaker. Special care has been devoted to sentence design, aiming to represent the most relevant syntactic constructs which are supposed (by the literature and by native speaker intuition) to have a prosodic correlate. Only regular syntactic structures have been considered, with little regard to non-standard discourse structure and complex dislocation and clefting phenomena (4). The text corpus is described in par. 3. Work is in progress now to analyze the speech data base. First results concerning the position and duration of pauses are reported in par. 4. Once the analysis is performed, the detected syntax-to-prosody correspondences will be expressed by formal rules and embedded in the grammar of a syntactic parser. A prototype of a syntactico-prosodic parser has been implemented, which applies a preliminary set of syntax-to-prosody rules. Its performances are discussed in par. 5.

3. TEXT CORPUS DESIGN

The text corpus, which consists of 200 sentences amounting to about 2,500 words and 14,000 characters, is arranged in four sections.

1. Isolated words.

Words occurring in the other sections are read in isolation for reference and comparison with their pronunciation in continuous speech. They are chosen so as to be representative of Italian lexicon with respect to: phonological structure, length, lexical stress position, grammatical class, frequency of use.

2. Simple sentences.

2.1 Intended for investigating prosodic word formation. Sentence design should reveal how deaccentuation is affected by:

lexical features (see above), rhythm (e.g. inter-stress distance), sentence structure (e.g. phrase structure, function word sequencing).

2.2 Some examples of peculiar or non-standard structures: e.g. quotations vs. parentheses, questions, alternative assertions, exclamative sentences, clefting, proper nouns, numbers, acronyms.

3. Complex sentences.

Intended for investigating intonational phrase formation and its dependence on syntactic structure. Two parallel sequences of sentences instantiate the same sequence of syntactic trees (ordered by complexity), differing in vocabulary and topic/style: essay style (very long words), fiction style (relatively short and usual words).

Sentences inside each group are designed so as to be comparable: the most complex sentences (more than 70 word long) are obtained from the simplest one (two-word long) through a sequence of enrichments, adding new words and complicating the syntactic structure. Almost identical word sequences may occur in different sentences, placed in different positions or playing, at the same place, different syntactic roles (adverb vs. prepositional phrase vs. subordinate clause; attribute vs. relative clause; restrictive vs. non-restrictive relative clause). Alternative attachments are compared: e.g. (NP (PP VP)) vs. ((NP PP) VP); (N (PREP (N (PREP N)))) vs. ((N (PREP N)) (PREP N))

4. Paragraph.

The reading of a paragraph (extracted from a book, essay style) is compared with its sentences read in isolation. Intended for investigating supra-sentential prosody and the intonational correlates of discourse structure.

Section 3 of the corpus has been chosen as the first to be analyzed in order to extract macroscopic prosodic features (see par. 4) and as a training text for the syntactico-prosodic parser (see par. 5).

4. ANALYSIS OF THE SPEECH DATA BASE

The text corpus described in the previous section has been read aloud by a professional speaker and the resulting speech signal has been acquired and stored digitally using a 12 bits A/D converter and a 16 KHz sampling frequency. All the recorded speech material (200 files) has been automatically analyzed to derive pitch and the voiced/unvoiced information. The pitch tracking algorithm used is based on cepstral peak detection. The work of segmenting and phonetically labelling the complete speech data base is now in progress. The final representation of the data base will associate to each uttered phoneme its duration and pitch values as well as a way of accessing the information about its phonetic, lexical, and syntactical context of occurrence.

As a first analysis on acquired data, an investigation has been carried out on the speech material corresponding to section 3 of the text corpus, detecting the position and duration of pauses (si-

lences). The hypothesis to be tested was that pauses are likely to occur at relatively regular intervals in correspondence of syntactic boundaries, their magnitude depending on the depth of the syntactic break and possibly on the duration of the adjacent phrases, whose length in number of phonemes may affect articulation rate. The total duration of the analyzed speech material, including pauses, is 9.1 minutes. The uttered sentences are 75, amounting to 1,180 words and 6,491 phonemes. Sentence duration ranges from 1.3 to 32.6 seconds. The number of speech segments, delimited by pauses (or enclosed within beginning of sentence and pause or pause and end of sentence) is 189.

Statistical results obtained on the described speech material are reported in the figures.

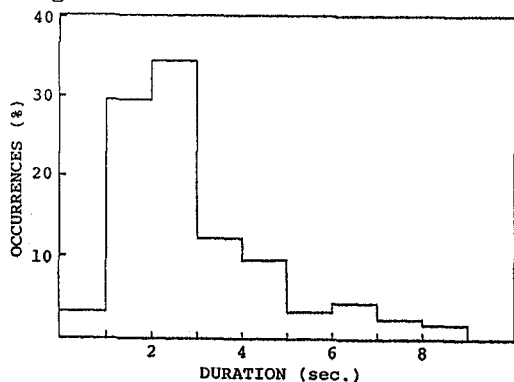


Fig. 1 Distribution of the duration of speech segments.

The distribution of the duration of speech segments (in sec.) is presented in the histogram of Fig. 1, where the abscissae represent time in sec. and the ordinates are the percentages. The histogram shows that about 80% of the segments do not exceed a 4 sec. duration, the mean value being 2.89 sec.

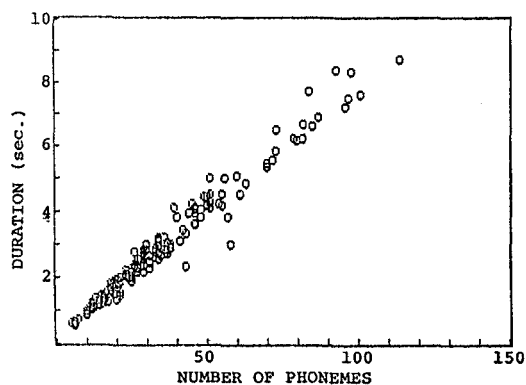


Fig. 2 Duration of speech segments as a function of the length in phonemes.

The relation between the duration of speech segments and their length in number of uttered phonemes is reported in the scatter diagram of Fig. 2. A strong linear relation holds between the two variables, the Bravais linear correlation coefficient being greater than 0.98. Contrary to

expectations, there is no evidence, from these data and for that speaker, of increasing articulation rate due to the length of phrase to be uttered.

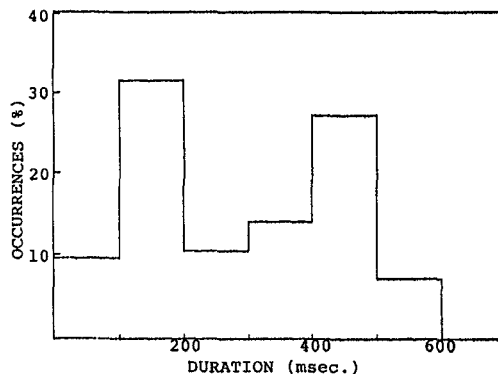


Fig. 3 Distribution of pause duration

The distribution of the duration of the 114 pauses, is presented in the histogram of Fig 3, where the abscissae represent time in msec. and the ordinates are the percentages. This distribution is clearly bimodal showing that pauses may be classified in two categories: short pauses lasting less than 300 msec with a mean duration of 140 msec. and long pauses lasting more than 300 msec. with a mean duration of 430 msec.

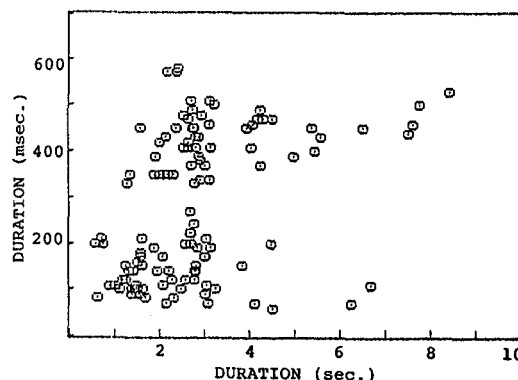


Fig. 4 Duration of pauses vs. duration of the preceding speech segment

Finally, the scatter diagram of Fig. 4 reports the durations of the pauses (ordinates in msec.) as a function of the duration of the preceding speech segment (abscissae in sec.). Two comments arise from the observation of Fig. 4: the first is that there is no evidence of correlation between the duration of a speech segment and the duration of the following pause; the second is that also this figure shows that pauses may be roughly split into two categories. A deeper inspection of the data base shows that long pauses have been detected only in sentences longer than 9 sec. (as a correlate of their syntactic complexity).

Looking for correlates of pause position and duration, the syntactic trees of the analyzed sentences have been inspected. As expected, all the pauses occur at

syntactic boundaries and never fall inside a phrase. Pauses delimit subordinate clauses and parentheses, follow prepositional phrases, separate subject and predicate as well as coordinate clauses and elements of lists. Typically "short" pauses occur between subject and predicate, while "long" pauses precede or follow a subordinate clause or delimit a parenthesis.

5. SYNTACTICO-PROSODIC PARSING

Were the function syntax-to-prosody known and expressed by rules, the problem would still be open of reaching syntax from text. A morphological lexicon and a syntactic chart parser are available in the Laboratory (6): in order to test the feasibility of a syntactico-prosodic preprocessor, the grammar of the parser has been modified to the effect of inserting prosodic markers at syntactic boundaries in the text. The considered markers are those recognized by the text-to-speech system (see par. 1): "continuation", "suspension", "parenthesis", "list". Syntax-to-prosody rules have been guessed, as far as possible reliable and general: in such a first-attempt approach, only most frequent phenomena are covered and errors are to be expected because of the lack of subtler distinctions. The rules have been expressed in the formalism of the augmented context-free grammar of the parser:

(PHRASE-CAT (pattern) (conditions)
(assembling instructions) PROS-MARK)

As an example, the marker CONTINUATION will be inserted in the input text between the subject and the predicate, when the parser builds the node S (sentence) of the syntactic tree by applying the rule:

(S (NP VP)(cond.)(instr.) CONTINUATION)

Rules may require restrictions: those which are syntactic (e.g. concerning the inner structure of an item in the pattern) can be expressed as "conditions", while others (e.g. rhythmical constraints such as length or inter-stress distance) should be imposed on the marker-augmented text by post-processing. The syntactico-prosodic parser prototype has been refined so as to obtain satisfying performances on a "training text" (section 3 of the corpus). The inserted markers (all and only those which were expected) were 150 on 1,452 words, that amounts, together with the 123 existing punctuation marks, to more than a prosodic marker every six words. The results are congruent with those obtained by analysing human reading of the same text (see par. 4).

The performances of the prototype have been tested on texts extracted from a newspaper, amounting to 2,000 words. The results are reported in Table 1, showing that 91.84% of expected markers were correctly inserted, while 8.16% are missing and 9.09% of inserted markers are wrong (mostly because of confusions between list and parenthesis). It is worth noting that

wrong markers are never inserted inside phrases, resulting in some way prosodically acceptable.

TABLE 1

RESULTS OF SYNTACTIC ANALYSIS ON NEWSPAPER TEXT

PROSODIC MARKER	CORRECT	WRONG	MISSING
Continuation	33	-	-
Suspension	27	1	-
Parenthesis	21	6	4
List	9	2	4

Error analysis reveals, beside syntactic parsing mistakes and insufficiencies of the implemented prosodic rules, some intrinsic limitations of the approach, which assumes the congruence of syntactic and prosodic structures, the graph of the second supposed included in that of the first. The assumption suffers from counterexamples. Syntactically coordinated phrases included between commas should be read as parenthetical, resulting as prosodically subordinate. More generally, syntax may be ambiguous in cases in which prosody instead is explicit: typical cases are prepositional attachment, or subject/object inversion. The human reader, indeed, chooses the alternative that is semantically plausible or, if semantics too is ambiguous, the one which is suggested by the pragmatic context or is congruent with the standard word order. Some knowledge about most likely (standard) attachment could be provided to the parser and information carried by commas should be exploited, forcing attachments, but some amount of syntactic ambiguity cannot be avoided.

6. CONCLUSIONS

A strong relation between the syntactic structure of sentences and the position and duration of pauses was revealed by the analysis of speech data. The performances of a syntactico-prosodic parser prototype were rather encouraging. Such preliminary results enforce our purpose of realizing a prosodic preprocessor, which is likely to strongly improve the quality of synthetic speech.

REFERENCES

- (1) E. Vivalda, "Italian Text-to-Speech Synthesis: the Linguistic Processor", Olivetti Res. & Tech. Review 7, 1987
- (2) S. Quazza and E. Vivalda, "Contextual Syntactic Analysis for Text-to-Speech Conversion", Proc. European Conf. on Speech Technology, Edinburgh, 1987
- (3) E. O. Selkirk, "PHONOLOGY AND SYNTAX", The MIT Press Cambridge Ma., 1986
- (4) R. Del Monte, "A Grammatical Component for a Text-to-Speech System", Proc. IEEE I.C.A.S.S.P., Tokio 1986
- (5) D. Cericola et al., "Morpho-Syntactic Tools for Speech Processing", in this issue