Prosodic Control in a Text-to-Speech System for Italian

Silvia Quazza, Pier Luigi Salza , Stefano Sandri, Alberto Spini
CSELT - Centro Studi e Laboratori Telecomunicazioni
Via G.Reiss Romoli 274, 10148 Torino, Italy

ABSTRACT

The paper illustrates a TTS-oriented model of prosody generation for Italian, which tries to account for the relations linking the prosodic parameters to relevant linguistic features. A proper rule development system allows to express contextual rules and to maintain the alignment between different linguistic representation levels.

INTRODUCTION

The prosody generation model of the TTS system for Italian developed at CSELT (for more details see Balestri et al., 1993) tries to account for the relations linking phoneme duration and fundamental frequency to phonetic structure, stress positions, syntax, and application-dependent pragmatic requirements of the message. High-level linguistic information is provided to the prosodic processor by the TTS module devoted to text processing, or by the application software which may select the appropriate prosodic style or insert syntactico-prosodic markers in the text of the message. By analyzing the surface structure of the text and by making use of a lexicon, the text processor is able to perform the following steps: text segmentation according to punctuation marks; translation of numbers and abbreviations in letters; recognition of the grammatical category for verbal forms and functional words; lexical stress assignment; syntactic boundary detection. Acronym pronunciation rules and phonetic transcription rules are then performed. The TTS prosodic processor marks any syllable carrying lexical stress, functional words excluded, and groups words into prosodic phrases relying on punctuation marks and NP/VP boundaries. So far, only one stress level is considered, whereas two different strength categories are assigned to phrase boundaries, depending on the presence/absence of pauses. Weak phrase boundary is recognized between short NP and VP, before numerical expressions and between the two words of a double surname, and involves no pause. Strong phrase boundary requires the insertion of a short pause (50-300 msec.) and is located in correspondence with comma or colon and anyway between long NP and VP. Sentence boundary corresponds to semicolon, full stop, exclamative and question marks, and is always accompanied by a long pause (500-1000 msec.). Finally, the prosodic processor distinguishes six sentence melodic modalities.

The rules managing such information are written in an abstract formalism and translated into C procedures by SCYLA (Lazzaretto and Nebbia, 1987), a specially designed rule development system which allows: to express contextual relations between sequentially organized elements by means of variables affecting different descriptive linguistic levels, hierarchically organized; to maintain data structures alignment between the different descriptive representations of the sentence; to associate duration and f0 values to each phoneme through a sequence of steps. As long as the position of a given segment is crucial for effectively determining prosodic parameter variations, suitable

counters are activated which detect step-by-step the units' location within the higher level unit they belong to, i. e. phonemes into syllable, syllables into phrase and sentence (counting stressed syllables is equivalent to counting phonological words), phrases into sentence. The variables pertaining to the several descriptive levels can be indifferently used for specifying the context in any procedure. A proper audio-visual debugger is available for rule testing, allowing to scan up to ten parameters for each phoneme, to interactively change parameter's values in every point of the phoneme string and to listen to the speech output.

DURATION RULES

The linguistic model underlying the duration rules (Salza and Sandri 1986, Salza 1988), is based on the principle of superposition of effects and on the possibility of modifying speaking rate, as previously illustrated (Nebbia, 1990). Standard rate corresponds to that used by speakers in aloud reading (about 5 syllables per sec.), but it can be continuously varied from -50% to +50%. If the sentence length is <= 5 syllables an automatic control slows the rate by 20%. Further specifications have been recently included into the general model coming from the analysis of dedicated natural speech corpora for the application of TTS to the reading of surnames, acronyms and addresses in the automation of telephone reverse directory service. New rules proved to contribute dramatically to the improvement of TTS intelligibility (Balestri et al., 1992).

Three separate procedures are devoted to assign, respectively, Intrinsic Duration of phonemes (Di), Phonetic Context Coefficients (Cph) and Syntactic Context Coefficients (Csy) to a target which is usually the single phoneme, rarely a whole class of speech segments. The context is expressed by means of the variables listed below, affecting five descriptive levels. In parentheses are signed the procedures involved.

Level Phoneme: individual phoneme label; phoneme articulatory (manner and locus) classes; voicing; intrinsic length, i. e. full vowel, semiconsonant, reduced vowel, single consonant, geminated consonant (Di, Cph).

Level Syllable: stressed/unstressed; open/closed (Cph, Csy).

Level Word: grammatical category (Cph, Csy).

Level Phrase: unit length, i. e. short phrase, full phrase; boundary strenght, i. e. weak phrase boundary, strong phrase boundary, sentence boundary (Cph, Csy).

Level Sentence: melodic modality, i. e. affirmative, interrogative, etc.; further specification, i. e. isolated word, long sentence, reverse directory service (Cph, Csy).

By making efficient use of SCYLA facilities, the maximum context extension now reached by the rules is 5 units, target phoneme included.

The most important phonetic context (microprosodic) phenomena covered by rules are the following. Both stressed and unstressed vocalic segments appear longer in vowel clusters than in diphthongs and in interconsonantal position. Increasing of consonant duration is apparent when they are in clusters. Stress realization involves: vowel lengthening from 50% to 150%, according to the surrounding phonetic context, semiconsonants longer in stressed diphthongs than in unstressed ones, slight duration increasing of intervocalic consonants. Like in many european languages, vowel duration is fairly influenced by voicing and articulatory class of the following consonant.

As for so called syntactic context effects, adjacency to sentence and phrase boundaries determines a generalized duration lengthening of the involved syllables, on the average more noticeable in final than in initial position, larger in final position for

vowels, in initial position for consonants. Sentence/phrase final segments lengthens from 20% to 180%, stressed segments less than unstressed ones. In presence of weak phrase boundary rules assign to segments in pre- and post-boundary positions slight lengthenings (from 20% to 40%). Heavier increasings are assigned in presence of strong phrase boundary. Moreover, the final vowel of the first word in long NP lengthens by 60%. Sentence boundary determines even more consistent duration lengthenings on adjacent segments. In isolated words every variation is stronger than in long sentences.

INTONATION RULES

Intonation rules have been developed relying on a multi-layered description of intonation (Avesani, 1990). An abstract phonological description of the linguistically relevant f0 modulations, which can be traced back to Pierrehumbert intonation model (Pierrehumbert, 1980), is well suited to represent the correspondence between tonal elements and structural events such as stressed syllables or syntactico-prosodic boundaries. A less abstract approach ('t Hart et al., 1990) has been adopted both as a methodological guideline for the experimental analysis of natural f0 curves and as a framework for rule implementation. This detailed phonetic and acoustic interpretation of tones is perspicuous in drawing the exact shape of pitch variations, which seems to be perceptually relevant and should be reproduced by the TTS system. Finally, the actual computation of f0 values takes into account those (physiological) features, such as pitch range and declination, which can be mathematically modeled as global trends of the f0 curve.

The intonation rules interpret the abstract tonal elements as pre-defined configurations of stylized 'pitch movements', classified according to their direction, dimension and timing with respect to stressed syllables and boundaries. Separate procedures, to be sequentially applied to the whole sentence, respectively:

- represent sentence intonation as a sequence of pitch configurations determined by the syntactico-prosodic structure;
- convert each configuration into a sequence of pitch movements by associating to the proper phonemes the targets of pitch variations, expressed as differences (in semitones) with respect to a reference baseline;
- combine targets with the current baseline, which depends on pitch range, declination and reset points;
- linearly interpolate target f0 values in order to obtain the complete f0 curve.

Declarative sentences are realized as a sequence of 'pointed hat' configurations aligned with stressed syllables, with a 'continuation rise' on the last syllable preceding a phrase boundary and with a 'flat hat' final configuration on the last two words. Phonologically, declarative contours can be represented as follows:

((H*) L H%) (H*) H* H+L* L L%

Phonetically, H* is interpreted as a rise starting on the pretonic syllable and reaching a peak at the end of the stressed vowel, followed by a fall on the postonic syllable. The continuation rise L H% is realized as a low rise on the last vowel before the boundary. The flat hat H* H+L* L L% is a rise on the pretonic and stressed syllables of the penultimate word, followed by a gradual slight fall till the last stressed syllable, a steep fall on the last stressed vowel and a further slight fall till the end of the sentence. Special

cases of declarative sentences are isolated words, which are realized with a high plateau followed by a steep fall on the stressed vowel and a further fall on the postonic syllables.

Interrogative sentences are realized with peculiar initial and final configurations and with 'reduced pointed hat' configurations on intermediate stressed syllables. As a final configuration a fall is realized on the syllables preceding the last stressed vowel, followed by a steep rise on the stressed vowel and a high plateau till the end of the sentence (higher in yn-questions). Different initial configurations are assigned to yn- and wh-questions. A rise-fall is realized on the first two words of yn-questions, reaching its peak at the end of the first stressed vowel and coming back to the baseline after the second stressed syllable. Two kinds of wh-questions are distinguished. If the wh-word is unstressed a rise is realized reaching its peak at the beginning of the first stressed vowel of the sentence, followed by a steep fall on the stressed vowel and the postonic syllables. If the wh-word is stressed a gradual fall is realized on it, starting well over the baseline at sentence beginning and reaching the baseline on the pretonic syllable of the second word.

CONCLUSION

Current research aims at realizing a more sophisticated text-analyzer which would give a richer and more reliable representation of the prosodic structure of sentences. Experimental analyses are also in progress in order to enhance the modelling of duration and intonation in long sentences, accounting for: destressing and graduation of stress levels, contextual speaking rate variations, identification of semantic focus, closer correlations among variations of different parameters for realizing prominence. As an alternative to rule systems, adaptive systems are also under investigation. Some encouraging results have already been obtained in the application of Neural Nets and CART to the prediction of phoneme duration.

REFERENCES

C. Avesani (1990), "A contribution to the synthesis of Italian Intonation", *Proc. ICSLP '90, Kobe, November 1990*, Vol. 2, pp. 833-836.

M. Balestri, E. Foti, L. Nebbia, M. Oreglia, P.L. Salza and S. Sandri (1992), "Comparison of natural and synthetic speech intelligibility for a reverse telephone directory service", *Proc. ICSLP '92, Banff, October 1992*, Vol. 1, pp. 559-562.

M. Balestri, S. Lazzaretto, P.L. Salza and S. Sandri (1993), "The CSELT system for Italian text-to-speech synthesis", to be presented at *EUROSPEECH '93, Berlin, September 1993*.

S. Lazzaretto and L. Nebbia (1987), "SCYLA: Speech Compiler for Your LAnguage", *Proc. EUROSPEECH '87, Edinburgh, September 1987*, Vol. 2, pp. 381-384.

L. Nebbia, "Text-to-Speech Synthesis System for Italian: an Overview", *Proc. VERBA '90, Rome, January 1990*, pp. 326-333.

J. Pierrehumbert (1980), *The Phonology and Phonetics of English Intonation*, Ph.D. dissertation (MIT, Cambridge).

P.L. Salza (1988), "Durations of Italian diphthongs and vowel clusters", *Language and Speech*, Vol. 31, Part 2, pp. 97-113.

P.L. Salza and S. Sandri (1986), "Microprosodic timing rules for consonant clusters in Italian", *Proc. ICASSP '86, Tokyo, April 1986*, Vol. 3, pp. 2035-2038.

J. 't Hart, R. Collier, A. Cohen (1990), *A Perceptual Study of Intonation* (Cambridge University Press, Cambridge).