A RULE-BASED TEXT-TO-SPEECH SYSTEM FOR PORTUGUESE

Luís C. Oliveira INESC/IST
M. Céu Viana CLUL
Isabel M. Trancoso INESC/IST

INESC, R. Alves Redol 9, 1000 Lisboa, Portugal
CLUL, Av. 5 de Outubro 85-6º, 1000 Lisboa, Portugal

Abstract

This paper describes the latest progress in the development of a text-to-speech system for Portuguese. The system comprises 4 major modules: text normalization, linguistic and phonetic processing, generation of the synthesizer parameters and synthesis. The present rule-based version, based on the Klatt80 formant synthesizer, has achieved promising results, namely in what concerns the performance of stress assignment, phonetic transcription and prosodic parsing. The paper describes each of the major modules, referring some language-dependent issues.

1 Introduction

The goal of this paper is the description of the second version of the DIXI text-to-speech system for Portuguese. Relative to the initial version [8], this system is characterized by an enhanced flexibility, which derives mainly from structuring it into four major modules: text normalization, linguistic and phonetic processing, generation of the synthesizer parameters and synthesis. The dividing line between the first two and the last two modules coincides with the frontier between what is independent and dependent on the type of synthesis (by rule or by concatenation).

The adoption of the synthesis-by-rule approach was motivated by several factors, namely, previous experience with Klatt's manual synthesizer [9] [2], which indicated that good results could be obtained for Portuguese. These results were particularly relevant in what concerns vowel reduction phenomena. In our language, in fact, unstressed vowels can undertake quality changes, shortening, devoicing and deletion. This last case is very frequent, resulting in large consonant clusters. Our second reason for adopting a rule-based system was purely research motivated: we were interested in developing a model which integrated linguistic and phonetic knowledge from the level of an abstract phonological representation to the level of the control of the synthesizer parameters.

The enhanced flexibility and modularity of DIXI's architecture makes this system an extremely important research tool, since it allows the evaluation of the linguistic and phonetic theories of the language, and also serves as a test-bed for assessing different sound production models.

Other major guidelines in the design of this system are its future extension to other varieties of Portuguese such as Brazilian Portuguese and varieties spoken in African countries and the feasibility of real-time implementation, which was taken into account, namely, by using efficient coding and by limiting the dictionary size. The system runs on several platforms including Unix systems (e.g., SPARCstations, DECstations) and PC's running MS-DOS. Due to the fact that the system is all written in C and that it does not need to load files in runtime, it can be easily ported to a dedicated board.

The second and third modules of our system were built using a multi-level rule-compiler (SCYLA - Speech Compiler for Your LAnguage), developed at CSELT [6], combined with a set of auxiliary functions written in the C language. The use of a rule-compiler has the advantage of imposing a more structured rule definition and of enabling the development of the system by non-programmers. SCYLA's choice was motivated by several features: its multi-level structure, which allows each procedure to simultaneously access all previous procedures; its ability to generate portable C code which can be optimized for the hardware where it is going to run; its powerful debugging tools; and, finally, its connection to a conventional procedural language for the operations more easily coded in this form.

In the development of the system a test set of about 25,000 different forms was used for testing procedures operating at the word level or bellow. This constitutes a frequency corpus collected by CLUL, based on oral interviews carried out all over the country. It comprises citation and inflected forms corresponding to

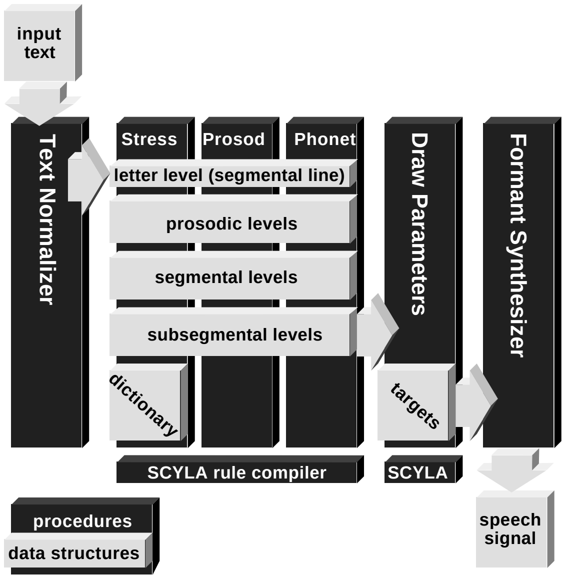

Figure 1: Block diagram of the DIXI system

about 715,000 occurrences and about 3,030,000 segments.

The structure of this paper closely follows the system division into four major modules. The system block diagram and the data-flow are shown in fig. 1 and will be presented in the following sections: text normalization in section 2, the linguistic and phonetic processing module in section 3, the generation of the synthesizer control parameters, in section 4, and finally, the waveform synthesizer, in section 5.

2 Text Normalization

The first procedure in a text-to-speech system is usually the normalization of the input text. In Portuguese, like in many other languages, this module has also to deal with the graphical stress marks in vowels (e.g. à, ê, í, õ) and with the c cedilla (ç). The need for this extra normalization step derives from the fact that, although there is an ISO standard, not all computer manufacturers adopt the same extended eight-bit ASCII code for these characters. Hence, in order to assure portability, we have decided to code these characters into two normal ASCII set characters (e.g. à→`a, ê→e^, í→i´, õ→o~, ç→c¸). This representation is the most commonly adopted by Portuguese users when typing with a keyboard unable to produce vowels with stress marks and c cedilla. In order to process text generated by these keyboards and by other devices or applications that only use the normal seven-bit ASCII character set, the normalizer module also accepts the internal representation format as input. Another problem is the Portuguese ordinal number extension: "º" and "ª" for the male and female gender. The system accepts ".o" and ".a" as equivalent, when located immediately after a number (e.g. 10ª → 10.a). These symbols are also used in abbreviations, as in "n.º" (number) and "Sr.ª" (Mrs.).

The text normalization module has the following steps:

Eight-bit characters: the first step of the normalization module is the conversion of the machine dependent ASCII extended set characters. Some of them are converted into the internal representation and others are replaced by Portuguese words (e.g. £ to libras).

Dates: the system searches the input text for valid dates in numerical format (three numbers separated by "/", "-" or "."). Numbers that are not valid dates (e.g. 30/4/91) will be translated as numerical expressions.

Abbreviations: words containing the symbol "." or "/" are looked-up in the abbreviation dictionary. Currently this dictionary contains 95 expansions. The input string "sr.ª" that was previously converted into "sr.a", is now expanded to "senhora".

Special characters: characters like "#", "/", "%" or "*" have a context-dependent translation. The symbol "/", for instance, can be translated as "a dividir por" (divided by), when surrounded by numbers, or as "barra" (slash), otherwise.

Numerals: the system can translate numbers in several formats: integers (10 → dez), floating point (10,2 → dez vi´rgula dois), scientific notation (1e2 → um vezes dez levantado a dois), money (2$10 → dois escudos e dez centavos), and ordinals (10ª → de´cima).

Stress mark position: when the input text has the stress marks separated from the vowel, it may be necessary to change its position to the internal format (e.g. "p´a" must be changed to "pa´") . This is only performed when there is no ambiguity (e.g. in "ve´u", the stress mark is always considered to be on the "e").

Acronyms: the adopted strategy is to restrict spelling to acronyms with no vowels, and to let the phonetic transcription rules deal with the others.

For efficiency reasons, this module has been directly programmed in the C language, but the normalization patterns have been expressed as regular expressions, which simplifies code writing and legibility [4].

3 Linguistic and Phonetic Processing

DIXI's second module is subdivided into three distinct submodules: lexical stress assignment, syntatic and prosodic parsing and phonetic values assignment.

Prior to the first one, however, a dictionary look-up step takes place, in which each word is searched in the dictionary and, if found, the corresponding entry is associated with it. A small lexicon of about 800 forms is currently being used, each entry having a structure of the form: normalized ortographic form, index of lexical stress position, phonetic transcription, grammatical category and indication of syntactic or prosodic behaviour (e.g. “`{"quatro", 2, "kw\"atru", _qnt, _foc}`”).

Some of the forms in the lexicon are exceptions to the lexical stress assignment and phonetic transcription rules, but it also contains the set of function words and the most frequent heterophonic homographs.

The phonetic transcription assigned to homographic forms is statistically based and temporary. It will be retained if there is lexical ambiguity (e.g. “`bola`” [`bO16`] – ball – and “`bola`” [`bol6`] – sort of pie) or it will be changed later, according to the grammatical category assigned to the item after syntactical analysis (e.g. “`almoço`” [`almosu`] – lunch, Noun – and “almoço” [`almOsu`] – I'm having lunch, Verb).

The system uses an international phonetic alphabet (SAM-PA [10]) and was designed to allow the introduction of applicability conditions at different levels of linguistic processing.

3.1 Lexical stress assignment

The lexical stress assignment module processes every word with more than one syllable, which is not included in the lexicon and lacks stress marks. This module was directly programmed in C for efficiency sake. It can be optionally deactivated for research purposes, by using the SCYLA compiler to test a different set of stress assignment procedures, or by using the SCYLA debugger to directly introduce stress marks in the segmental line.

Primary stress is marked with the SAM-PA symbol (“"”) by a set of 18 rules, basically the same as proposed in [3]. Those rules have been applied to 89.4% of the forms in our test set, producing 99.9% of correct results. Most of the errors are observed for forms which are derived with suffixes that do not trigger stem distressing, and can be overridden by the introduction of the most frequent ones in the lexicon.

3.2 Syntatic and Prosodic Parsing

The first procedure performed with SCYLA fills the first data level, denoted as letter, with the input text and stress marks. This first level is directly taken as the segmental line, without any grapheme-to-phoneme mapping rules. For each item having an associated entry in the lexicon, several other levels are also filled with the corresponding information.

As it is widely accepted, the evolution of prosodic parameters such as fundamental frequency, energy and duration cannot be accurately predicted on syntactical grounds only. There is strong evidence that they are closely related to the utterance phonological structure, which does not necessarily coincide with the syntactic one. Prosodic constituency principles, however, appear to be sensitive to major syntactic boundaries. Thus, the next step consists of a very crude syntactic parsing, based on a limited set of information concerning word class and punctuation marks, and aiming at localizing the verb and identifying sentence boundaries. At this stage, the system also searches for a set of expressions indicating syntactic structures generally related to prosodic focus.

In its current version, the system attempts to mimic the results of a grid and constituent model of prosodic parsing, in order to account for prominence relations ranging from the syllable up to the utterance level. However, no fully metrical analysis is performed. Both the assignment of pure metrical positions, and the first level of prosodic grouping above the syllable, are reconstructed from primary stress location. A bottom-up prosodic parsing is then achieved, based on rhythmic alternation principles.

Short utterances can be produced without any prosodic partition, but long ones need to be phrased. Using a top-down procedure, the system computes a set of possible utterance prosodic partitions, thus guaranteeing that a partition at the Nth level implies a partition at the $N+1th$ level. The selection of the partition level is performed semi-randomly, taking into account different phrasing degrees of probability. These degrees are attributed on the basis of rhythmic principles (e.g. eurhythmic partitions are the most favoured; partitions with the longer prosodic group on the right side are preferred to those having the shorter group in this position). Phrasal boundaries are then assigned and the pauses (if any) are introduced.

The next step is concerned with intonational melodies, represented as tone sequences on a distinct level of the rule system. Tone bearing units are marked, and tone association and spreading are computed on the basis of a set of principles which respect utterance phrasal boundaries at the partition level.

3.3 Phonetic value assignment

Most synthesis by rule or by concatenation techniques require a narrow phonetic transcription as input. DIXI has three different phonetic transcription sets of rules. The first set operates on a left-to-right fashion on the segmental stream, generating a phonetic transcription of each word, as if it was produced in isolation as a citation form. The second set of rules operates on this output stream, accounting for sandhi as well as other prosodic sensitive phenomena.

Altogether, there are about 200 rules, basically the same as in [3], which, when applied to the test set described above, produce 98 and 92% of correct segment and word transcription results, respectively. These results have been obtained without dictionary look up. About 80% of the errors are due to the fact that there is no information about the morphological structure of the forms. Most of the remaining errors are exceptions to several rules. Better scores may be achieved by including erroneously transcribed words in the lexicon.

As the scope of certain features does not always coincide with the segment as a whole, the provided phonetic transcription does not contain all the information needed for the generation of the synthesizer parameters. A third set of rules deals with subsegmental aspects of feature association and spreading, allowing the splitting of segments into different parts. For instance, plosives are decomposed into closure and burst, trills into sequences of obstruent/vowel-like alternations, and nasal vowels into a [-nas] subsegment followed by a [+nas] one.

Although not mandatory for obtaining correct results, this three-step strategy is advantageous from the point of view of rule ordering and verification of applicability conditions. On the other hand, the independency of procedures related to different levels of the linguistic representation results in a greater efficiency and flexibility of the system.

The last step of this module estimates prosodically dependent parametric data, accounting for the relative influence of prosodical as well as segmental properties. Durations are computed first on the basis of speaking rate, prosodic prominences, syllabic structure and segmental context. Reference lines, whose length and slope are also determined by the prosodic structure, are then drawn and used to scale F0 and energy values.

4 Generation of the Synthesizer Parameters

This module draws the variable synthesizer control parameters tracks. In the current version, there are 18 of them: the fundamental frequency, the amplitude of the voiced excitation, the amplitudes of the cascade and parallel model noise excitations, 4 formant frequencies, 3 formant bandwidths, the frequency of the nasal pole, 5 formants amplitudes of the parallel filter, a direct noise bypass amplitude and the nasal zero frequency.

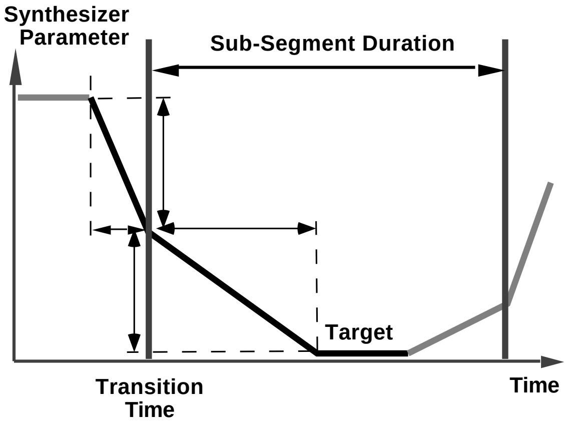

Figure 2: Transition model for the synthesizer parameters

The synthesis strategy is based on a target and transition model similar to the one used in the MITalk system [1]. In our system, however, the drawing of the parameter tracks is done at a subsegmental level. The transition model is shown in fig. 2. The transition variables control the transition shape, the smoothing time constants and the discontinuity values. As subsegmental aspects of feature timing are considered, no exceptional treatment is needed for diphthongs, nor plosives.

The first step of this module is to look up the default target and transition tables for each parameter of the current subsegment. The target and transition modification rules are then applied to adjust these values accounting for feature spreading and overlapping. In diphtongs, for instance, the relative timing of formant movements depends on the position of the glide relative to the vowel, and in nasal vowels, these rules determine the nasal pole-zero pair distance and the amount of nasal murmur.

Using the described model, a set of synthesizer control parameters is then computed every 5 ms by linear interpolation.

5 Waveform Synthesizer

The final module in this TTS system is a test-bed for experimenting different sound production models. As a reference model, the Klatt 80 [5] formant synthesizer has been adopted, given its well known ability to generate natural sounding synthetic speech, when the necessary linguistic and phonetic knowledge is provided.

Our latest research efforts in this area have concentrated on testing different sound production models, from the point of view of signal reproduction. This is done independently of their integration in DIXI, since it may imply the implementation of the two last modules using synthesis by concatenation. This work includes testing vocoder models with different glottal pulse shapes and mixed excitation strategies, multipulse models and harmonic/sinusoidal models.

We have been particularly interested in a recent version of the harmonic synthesizer incorporating narrowband basis functions for the generation of the unvoiced source [7]. The major feature of this class of models which makes it attractive for speech analysis/synthesis research is its direct control over amplitudes and phases of individual harmonics. The objective is also to take into account our knowledge of the human perception mechanism, which can be advantageously characterized in the frequency domain. At this very preliminary stage, however, no results can be reported yet.

6 Conclusions and Future Developments

The DIXI system is, to our knowledge, the only text-to-speech system specifically designed from scratch for European Portuguese, and it results from the cooperation between the speech processing group of INESC and the phonetic group of CLUL. Although still at its earliest stages, very encouraging results have been achieved so far, namely in what concerns the performance of the stress assignment, phonetic transcription and prosodic parsing. In particular, the semi-random selection of the prosodic partition level strongly contributes to the decreased monotonousness of the synthetic speech.

Future work will be carried out, in parallel, in different modules, taking advantage of the flexible structure of the system and aiming, for instance, at achieving a better understanding of the vowel reduction phenomena, integrating new synthesis models and evaluating their performance for Portuguese, and using more powerful syntatic parsers. The realization of intelligibility and comprehension tests is also planned for the near future.

Acknowledgements

We would like to thank D. Klatt and K. Stevens for their strong support and incentive during the preliminary stages of this work.

References

[1] J. Allen, M. S. Hunnicutt, and D. Klatt. *From Text to Speech: The MITalk System.* Cambridge University Press, U.K., 1987.

[2] A. Andrade. Um estudo experimental das vogais anteriores e recuadas em português: Implicações para a teoria dos traços distintivos. ms., 1989.

[3] E. Andrade and M. Céu Viana. Corso I - um conversor de texto ortográfico em código fonético para o português. Technical report, CLUL-INIC, Lisbon, 1985.

[4] P. Carvalho, P. Geada, and P. Lopes. Norm : Normalizador de texto para português. Technical report, INESC, Lisbon, 1991.

[5] D. H. Klatt. Software for a cascade/parallel formant synthesizer. *JASA*, 67:971–995, 1980.

[6] S. Lazzaretto and L. Nebbia. Scyla: Speech compiler for your language. In *Proc. of the European Conf. on Speech Technology*, volume 2, pages 381–384, Edimburgh, Sep. 1987.

[7] J. Marques and L. Almeida. Sinusoidal model of speech: Representation of unvoiced sounds with narrow band basis functions. In J. Lacoume, N. Martin, and J. Malbos, editors, *Signal Processing IV, Theories and Applications*, pages 891–894. North-Holland, 1988.

[8] L. C. Oliveira, M. C. Viana, and I. M. Trancoso. DIXI – Portuguese text-to-speech system. In *Eurospeech*, pages 1239–1242, Genoa, Sep. 1991.

[9] K. N. Stevens, A. Andrade, and M. C. Viana. Perception of vowel nasalization in vc contexts: A cross language study. *JASA*, 82-S119{A}, 1987.

[10] R. Winski, W. J. Barry, and A. Fourcin, editors. *Support Available from SAM Project for other ESPRIT Speech and Language Work.* Esprit Project 2589 (SAM), Multi-Lingual Speech Input/Output Assessment, Methodology and Standardisation, 1989.