DIXI – PORTUGUESE TEXT-TO-SPEECH SYSTEM

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 software architecture of the Portuguese text-to-speech system DIXI[1]. The system has three major modules. The first one contains the text normalizer and searches each word in the lexicon. The second one is a multi-level rule based module for lexical stress assignment, orthographic to phonetic transcription, metrically based prosodic patterning and for generating the evolution of the synthesizer parameters. The final module is the Klatt 80 formant synthesizer. The paper describes each of these main modules, emphasizing the particularities of text-to-speech synthesis in the Portuguese language.

Keywords: Speech Synthesis; Text-to-speech Systems; Portuguese Language; Synthesis-by-rule.

1 Introduction

The DIXI project is the result of the cooperation between the speech processing group of INESC and the phonetic group of CLUL and is, to our knowledge, the first text-to-speech system specifically designed for European Portuguese, from scratch.

Several guidelines were adopted in the design of this system. One of the priorities was to have a modular and flexible structure in order to allow its use as a tool for linguistic and phonetic research, and the development and evaluation of new models of sound wave production. The future extension of this system to other varieties of Portuguese, such as Brasilian Portuguese and varieties spoken in African countries was another major guideline. The system was also designed bearing in mind its real-time implementation, namely by using efficient coding and by limiting the dictionary size. It runs on several platforms including Unix systems (e.g., VAXstations, DECstations, Suns, Alliant) and PC's running MS-DOS. Due to the fact that all the system can be transcribed into the C language and that it does not need to load files in runtime, it can be easily ported to a dedicated board.

For procedures applied at word level or bellow, a test set of about 25000 different forms was used. This constitutes a frequency corpus collected by CLUL for other purposes, comprising citation as well as inflected forms, and corresponding to about 715000 occurrences.

[1] Latin expression used at the end of a public speech

Figure 1: Block diagram of the DIXI system

The three major modules of the DIXI system are depicted in the block diagram of fig. 1 and will be separately discussed in the following sections: text pre-processing in section 2, linguistic and phonetic processing in section 3, and finally, the formant synthesizer in section 4.

2 Linguistic Pre-Processing

This first module performs the input text normalization and searches each word in the dictionary.

For efficiency reasons, the module is programmed directly in the C language, using functions for compiling and matching regular expressions, which simplifies code writing and legibility.

The first step in the normalization procedure is the conversion

the eight-bit characters to an internal representation in seven-bit characters. This is particularly important for the Portuguese language, since it uses the c cedilla (ç) and graphical stress marks in vowels (e.g. à, ê, í, õ) which are usually coded in the extended ASCII code using the eight-bit representation. Although there is an ISO standard for this extended set, it is not respected by all manufacturers, which led us into adopting two seven-bit characters for these symbols (e.g. `c, 'a e^ i' o~`). There are also other symbols that must be replaced by Portuguese words (e.g. £ to `libras`) or by internal representation (e.g. º to `.o`).

In the next step, the system searches the input string for dates in numerical format (e.g. 28/2/91, 28-2-91). Only valid dates are transcribed, in order to reduce the risk of translating a numerical expression.

The system contains a small dictionary of 95 abbreviation expansions which is searched when the current word ends with the symbols "." or "/" eventually followed by an extension (e.g. the Portuguese abbreviation for engineer – engº – which was previously normalized to `eng.o`, is now expanded to `engenheiro`).

The following step in the normalization procedure is the translation into words of all the characters that are not letters nor punctuation marks (like #, $, %, *). Some of these characters have context dependent translations for instance "*" can be translated to `asterisco` (star) or to `vezes` (times) in the middle of a mathematical expression.

The translation of numerals is a common procedure in all text normalizers. The DIXI system can translate both ordinal numbers (e.g. 101 – `cento e um`, 101º – `cente'simo primeiro`, 101ª – `cente'sima primeira`) as well as cardinal numbers in integer, fixed or floating point format.

Since not all keyboards can produce the Portuguese characters, the normalizer also accepts the stress marks separated from the vowels, as in `'a` or `a'`, and the cedilla separated from the c. This is specially useful for processing Unix electronic mail messages which do not allow eight-bit characters and it is also by far the most common way adopted by Portuguese users when typing on a foreign type of keyboard. Whenever necessary, the text normalizer changes the position of the mark or cedilla to the internal format position.

The last step of the normalization procedure is the processing of acronyms. The adopted strategy is to restrict spelling to acronyms with no vowels, and to let the phonetic transcription rules take care of the others.

After input text normalization each word is searched in the dictionary and, if the search is successful, the entry is associated with it. In the current version, the system uses a small dictionary, containing the index of the word stress vowel, the phonetic transcription and the grammatical category of each form. The dictionary is used for exceptions to the phonetic transcription rules and for syntactic parsing of the utterance.

3 Linguistic and Phonetic Processing

Although a text-to-speech system can be seen as an attempt to model the linguistic and phonetic knowledge needed to produce natural speech from an abstract phonological representation, this is only partly true for the present version of DIXI. In fact, a complete model would require a much deeper understanding of some of the language specific phenomena in Portuguese. On the other hand, more pragmatic approaches can be justified in some parts for efficiency sake.

The system uses an international alphabet (SAM-PA [10]), and was designed to allow the introduction of applicability conditions at the different levels of the linguistic processing. The two factors are important for its use as a research tool and for future extensions to other varieties of Portuguese.

With the exception of lexical stress assignment , the linguist and phonetic module was built using a rule compiler 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 [6] and enabling the system development by researchers with less programming skills.

SCYLA, Speech Compiler for Your LAnguage, the rule compiler developed by CSELT [7], was selected because of three basic features of its multi-level structures, allowing each procedure to access simultaneously all the previous procedures results; its ability to generate portable C code which can be optimized for the hardware where it is going to run; and, finally, its connection to a conventional procedural language for the operations more efficiently coded in this form.

3.1 Lexical stress assignment

Lexical stress assignment is one of the most important factors for a correct reading of European Portuguese, since stress dependent vowel reduction is one of its most striking characteristics. Unstressed vowels can undergo quality change, shortening, devoicing and deletion.

This assignment is a necessary step for words not included in the dictionary, without a graphical stress mark (` , ´ or ^) and with more than two letters.

The stress vowel is marked with the SAM phonetic alphabet symbol for primary stress (") and is located by a set of 18 rules which are basically the same as described in [3].

For efficiency sake, we have decided to write these rules directly in the C language instead of using the rule compiler. Otherwise, stress could have been assigned by the same set of rhythmic rules that describe the relative prominence of syllables within a word.

In our test set, 88% of the forms need the stress vowel marking rules. The general rule is applied for 71% of the cases, and each one of the remaining rules never exceeds an application rate of 10%.

3.2 The segmental line

The first procedure of the rule system fills in the first level with the input text and the marks on the stress vowel. A number of different levels is also filled with the dictionary information for the words with an associated entry.

The first level, letter, is taken directly as the segmental line, without any grapheme-to-phoneme mapping rules. This option

is motivated by the regularity of Portuguese orthography, based mainly on phonological criteria. In the cases of e and o, where the same orthographic symbol can be associated with two different phonological representations, the low vowel is assumed. This approach, taken for rule simplicity as well as statistical reasons, handles homographs (e.g. pega [p"Eg6] – magpie – and pega [p"Eg6] – a handle) as well as ambiguous word endings (e.g. maravilhosa [m6r6viL"Oz6] – marvelous – where osa is a suffix, and raposa [R6p"oz6] – fox – where it is not).

3.3 The syntactic parsing

A limited syntactic parsing is a common procedure in several text-to-speech systems for other languages [1] [9] [8]. In the DIXI system, a very crude syntactic analysis is performed by means of identifying punctuation marks and a set of grammatical words to which certain syntactic structures are normally associated. This step aims to identify the clause and sentence boundaries, modality and verb localization. This type of information is essential for a good performance of the prosodic parsing and phonetic transcription procedures, described below.

At this level, the program also searches a set of expressions indicating syntactic structures, always associated in Portuguese with a prosodic focus (e. g. até – even, o próprio – himself). If no such structures are found, a focus marker can be assigned to the first or last constituent of the sentence by a random process.

3.4 Prosodic parsing

In Portuguese, as well as in many other languages, the different syllables within a word are structured according to a rhythmic principle of alternation between strong and weak beats, the same kind of principles being used to group words into larger units. Using a grid and constituent model to account for the internal organization of syllables within words, a close relationship was found between the degrees of prominence attributed to the syllables by the model and their relative durations [4].

Those degrees of prominence also account for most of the variance observed in the relative durations of the segments within the syllables. Extending this type of analysis to prosodic domains above the word level, it is possible to account for the relative prominences in connected, speech and predict its main prosodic properties.

As temporal and spectral reduction of unstressed vowels, as well as lengthening of stressed ones, play a central role on the naturalness and fluency of spoken Portuguese, DIXI performs a prosodic parsing of each utterance to be synthesized.

Based on rhythmic principles, a bottom-up prosodic parsing is made: segments are grouped into syllables, syllables into words, words into prosodic phrases and prosodic phrases into prosodic groups up to the level of the utterance. This gradual grouping into larger and larger units is achieved respecting the broad syntactic hierarchy, and prominences are attributed at each level of the analysis procedure.

Next, a top-down procedure of prosodic partition of the utterance is adopted. Relatively short utterances can be produced without any prosodic partition, but very long ones cannot. The existence of a partition at level N_x implies a partition at the level N_{x+1}. The set of possible partitions is computed and different degrees of probability are assigned to each of them (e.g. eurhythmic partitions are the most favored; partitions with the longer prosodic group on the right side are preferred to those having the shorter group in this position). A random selection of the partition level is then made and the pauses (if any) are introduced.

The next step consists of a series of procedures for melody assignment and tonal association. The tonal features are considered as behaving independently, and are thus represented in distinct autosegmental lines, synchronized with the segmental line. The critical associations are marked and tonal features spread over different domains that can intersect on the segmental line.

The prosodic module generates an abstract representation in terms of phonetic features and prominence relations that determines the prosodic properties of the utterance.

3.5 Orthographic-phonetic transcription

Binary-valued features are associated to each token of the segmental line and the phonetic transcription of the utterance is performed. Some of the rules considered at this level are purely phonetically motivated, while others are sensitive to prosodic boundaries. A small set is also sensitive to the word grammatical category. The complete set of rules, in a number of about 200, is basically the same as in [3] and accounts for allophonic variation within and across word boundaries. Re-syllabification is automatically triggered off by certain rules, namely by those involving vowel deletion or diphthongization across word boundaries.

Using the test set referred above, 92% of the words were correctly transcribed without resorting to a dictionary. In the remaining cases, a single error per word was generally detected. About 80% of the errors are due to the fact that there is no information on the morphological structure of the forms. Most of the remaining errors are exceptions to several rules. Both need to be included in the dictionary.

3.6 Phonetic values assignment

The phonetic transcription is rather broad and does not contain all the information needed to drive the control parameters of the synthesizer. Other domains of feature association need to be taken into account. This is motivated by the fact that while certain features associate with the segment as a whole, others do not. They are associated to parts of segments and spread to adjacent parts of other segments (e. g. nasality in vowels).

The notion of subsegment is, thus, explored and the first procedure at this level does the splitting of a segment into different parts. For instance, plosives are decomposed into closure and burst, and trills in sequences of obstruent/vowel-like alternations, whose number and order are context-dependent. Binary-valued features are then transformed into n-ary ones.

Tables of default transition duration and target values for each one of the variable control parameters of the synthesizer are searched and synchronized at this subsegmental level. A set of

target modification rules are then applied. In the case of nasal vowels, for instance, these rules determine the nasal pole-zero pair distance and the amount of nasal murmur.

3.7 Synthesizer control

The synthesis strategy is similar to the one described in [1], that is, a target and transition model is used to draw all the control parameter tracks for the synthesizer. Transition variables define the transition type, the smoothing time constants and the discontinuity values.

The last step of this module estimates the parametric data, accounting for the relative influence of prosodical as well as segmental features. Reference lines, whose length and slope are determined by the prosodic structure, are used to scale prosodic properties such as F0 and energy. Segmental durations are also calculated on the basis of prosodic prominences, syllabic structure and segmental properties.

Finally, the synthesizer control parameters are computed in 5 milliseconds interval, by linear interpolation. In the current version, this module controls 18 of the synthesizer parameters.

4 The Formant Synthesizer

The DIXI system uses the Klatt80 [5] formant synthesizer with small changes in the voicing source.

One of the main reasons for this choice was the well known ability of this synthesizer to generate speech close to a human-like quality for different types of voices, provided that the necessary linguistic and phonetic knowledge is adequately embedded in the system, as shown by the performances of MITalk and DECtalk systems.

Previous work using Klatt80 to generate stimuli for perceptual experiments [11] [2], also showed that the acoustic patterns observed for Portuguese can be successfully imitated at natural-speech quality.

Furthermore, Portuguese vowel reduction and large consonant clusters resulting from vowel deletion, are easier to produce using this model than by concatenation of segmental units.

For those reasons the Klatt80 synthesizer seemed like the natural choice for our system.

5 Conclusion

Although the DIXI system is still in an experimental stage, the results achieved so far can be considered encouraging, namely: the performance of the stress assignment and phonetic transcription modules, the versatility of the target and transition model and the results of prosodic parsing. Moreover, the intensity and fundamental frequency modulations, based on the random selection of a partition level, strongly contribute to a much less monotonousness of the synthetic speech.

Future work will include the realization of intelligibility and comprehension tests, the implementation of a more powerful syntactic parser and a better understanding of vowel reduction and related phenomena.

References

[1] Allen, J., Hunnicutt, M. S., Klatt, D. (1987). *From Text to Speech: The MITalk System*, Cambridge Univ. P., Cambridge, U.K.

[2] Andrade, A., (1989). *Um Estudo Experimental das Vogais Anteriores e Recuadas em Português: Implicações para a Teoria dos Traços Distintivos*, Diss. CLUL-INIC, ms.

[3] Andrade, E., Viana, M. C. (1985). "Corso I - Um conversor de texto ortográfico em código fonético para o português", *Tech. Rep.*, CLUL-INIC.

[4] Andrade, E., Viana, M.C. (1988). "Ainda sobre o ritmo e o acento em português", Actas do 4º Encontro da Associação Portuguesa de Linguística, Lisboa, 1988, pp. 3-15.

[5] Klatt, D. H. (1980), "Software for a cascade/parallel formant synthesizer", *Journal of the Acoustical Society of America*, 67, 971-995.

[6] Klatt, D. H. (1987), "Review of Text-to-Speech Conversion for English", *Journal of the Acoustical Society of America*, 82(3), 737-793.

[7] Lazzaretto, S., Nebbia, L., (1987). "SCYLA: Speech Compiler for Your Language", *Proc. of the European Conf. on Speech Technology*, Edimburgh, September 1987, 2, 381-384.

[8] O'Shaughnessy, D.D. (1989). "Parsing with a small dictionary for applications such as text to speech", *Computational Linguistics*, 15, 97-108.

[9] Sorin, Ch., Larreur, D., Llorca, R. (1987). "A rhythm based prosodic parser for text-to-speech systems in French". *Proceedings of the XIth ICPhS*, 1:125-128.

[10] Winski, R., Barry, W. J., Fourcin, A., (ed.s) (1989) *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.

[11] Stevens, K. N., Andrade, A., Viana, M. C., "Perception of Vowel Nasalization in VC Contexts: A Cross Language Study", *Journal of the Acoustical Society of America*, 82, 1987, S119{A}.