Translating Text to Phonemes for the Portuguese Language

Geraldo Lino de Campos

Dimas Trevisan Chbane

Departamento de Engenharia de Computação e Sistemas Digitais

Escola Politécnica da Universidade de São Paulo

Caixa postal 8174 - São Paulo 01065-970

BRASIL

RTC@FPSP.FAPESP.BR

Abstract:

A system for text-to-speech conversion for the Portuguese language is described. The system includes a 320,000 words vocabulary, encompassing all valid Portuguese words. Using information in this vocabulary and a set of rules, the system is able to translate a general text in the corresponding phoneme sequence.

1 Introduction

The use of speech as a form of computer output has growing importance as part of user-friendly man-machine interfaces and for specific applications. However, its usefulness has been impaired by the lack of synthesis systems able to generate speech from an unrestricted text, providing a natural and pleasant sounding.

The process of voice synthesis from text is comprised of the following steps:

1 - Text to phonemes translation. In this context, the word phoneme comprises phonemes and diphones, including their durations. Diphones are peculiar sequences of phonemes that, due to coarticulation effects, are very difficult to simulate precisely when the phonemes are considered alone. It happens more with plosive-vowel sequences. This translation is most a problem of lexical analysis, although some rudimentary syntactical analysis is required, since a relatively large class of words has different pronunciation when they belong to different grammatical classes. This happens more frequently with verbs and substantives (go[o]sto, substantive, and go[ ]sto, verb, for instance).

2 - Pitch contour determination. Speech with constant pitch, resulting in a robot like voice, is very irritating and a serious system must generate a pitch contour such that each phrase sounds as naturally as possible. Since the pitch conveys non-syntactical and emotional information, it is not generally possible to produce really natural sounding utterances, but acceptable pitch contour can be generated based on the general grammatical structure of a phrase. The result of this phase can further change the duration of the phonemes already determined in phase 1. Much research is still needed, but present results are encouraging; results will be published in the future.

3 - Speech synthesis. This phase can be implemented with an improved LPC synthesizer[Sam78], yielding medium quality voice. The improvement used is to change the impulsive excitation by a triangularly shaped excitation function. A better technic currently under study is the use of a more complex excitation function, using a technic somewhat similar of what is done on the successful CELP vocoders. The results will be published elsewhere.

The synthesis process is shown in figure 1.

This paper concentrates on algorithms and techniques for the phase 1, and is organized in 5 sections. Section 2 presents some general considerations; section 3 deals with the characteristics and internal structures used by the vocabulary. Section 4 discuss the specifics of the problem of text-to-phonemes conversion for the Portuguese language, and the rules adopted. Section 5 presents some concluding remarks.

Through this document, symbols between brackets are phonemes represented according to the IPA (International Phonetic Alphabet). The word phoneme is used in a broad sense, including diphones as well.

2 General considerations

Converting an unrestricted text into speech requires the capacity of converting any sentence in the language; therefore, every word of the language must be recognized. This requires the availability either of a general algorithm able to convert any word into a sequence of phonemes, or a vocabulary comprising all the words of a language, or some suitable combination.

Practice had shown that the first approach is impossible for most languages, including English and Portuguese. Most systems use the third approach, based on the assumption that a large vocabulary is difficult to generate, to maintain and, above all, to access.

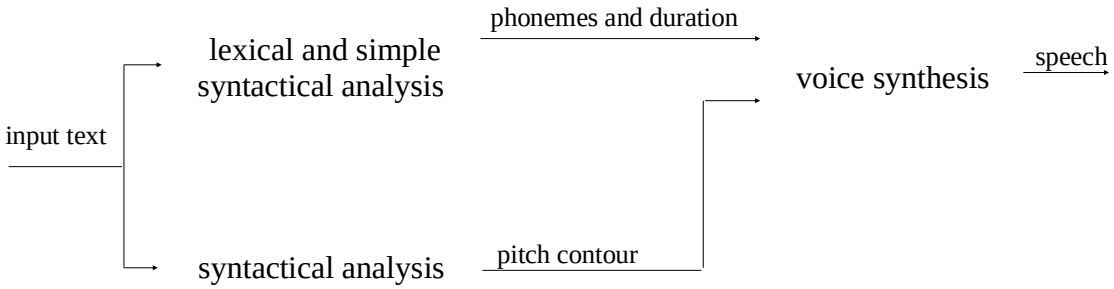

Figure 1 - Modules of the text-to speech system

The present system uses the third approach, referring all words to a vocabulary. In Portuguese, the problem of conversion from text to phonemes is simpler than in English, since most letters can be unambiguously converted to a phoneme, but there are exceptions that requires using a dictionary, since these letters can be mapped to a small (2 to 4) set of phonemes, in most cases by historical reasons. Once a dictionary is required, it is worthwhile to consider the problems usually associated with the use of a full vocabulary.

Contrary to the general opinion, it is easier to generate a full vocabulary, provided there is a standard one. All that is required is to transcribe it to a machine readable form, either with OCR equipment or by hand transcription; in any case, a mechanical operation. To use only the exceptions requires the determination of the exceptional cases, which is a tedious but intellectual operation. Maintenance and access issues are related: if there is an automatic procedure that converts the transcription of the vocabulary to a compact representation, and fast access procedures, these issues can not be considered a problem. Such procedures exists, and consist of converting the vocabulary to a (possibly augmented) finite state automaton, [Gro89], [Luc93]; in the latter, a representation was obtained with a compactation in the range of then 1 bit per word, and very fast accesses.

Taking this in consideration, it was decided to use a full representation of the vocabulary. Section 3 details the characteristics and implementation of the vocabulary.

The vocabulary is not enough, however, since the uttering of a word frequently depends on context. There are three conceptually different dependency categories. The first is essentially grammatical, and has to do with individual words; it occurs when the same word may represent different grammatical entities, like being a substantive or a verb. It can be somewhat easily handled by a simple grammatical analyzer.

The other categories deal with the sentence level, and determine the pitch contour of the utterance. The second can be related with the structure of the sentence, and can roughly be determined by a more elaborate syntactical analysis. The third category depends on semantics: a

state	attribute	link		letter	state	letter	state
0		→		C	1	R	9
1		→		O[o]	2		
2		→		N	3		
3		→		T	4		
4		→		O[o]	5	O[ ]	10
5		→		R	6		
6		→		N	7		
7		→		O[o]	8		
8	substantive, masculine, singular						
9		→		E[e]	3		
10		→		R	11		
11		→		N	12		
12		→		O[o]	13		
13	verb, present, singular, 1st person						

Figure 3 - Automaton corresponding to the words in figure 2

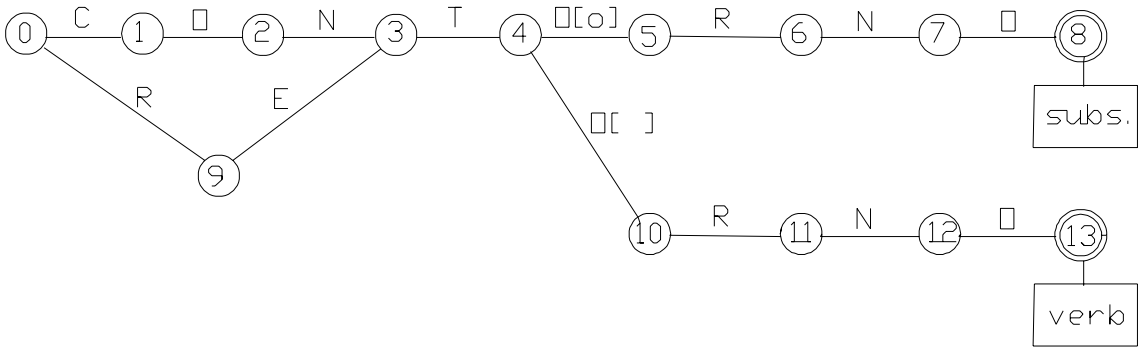

Figure 2 - Example of an augmented automaton
This automaton recognizes the words reto[o]rno (substantive), reto[ ]rno (verb), conto[o]rno (substantive), conto[ ]rno (verb). See text for details.

speaker can change the meaning of an utterance by changing the stress and intonation of certain words; it is impossible to determine differences in this category using only the text.

3 Vocabulary structure

The vocabulary was produced by transcription of the "Vocabulário Ortográfico da Língua Portuguesa", which is the official list of the Brazilian Portuguese language. It contains the word and grammatical category (categories if the word belongs to more then one). For some words, it contains also phonetic information about some vowels that can be open or closed. As usually happens in dictionaries produced by hand, there is a lot of difficulty in using these information, since there are many deviations from the standard syntax, specially lack of punctuation and comments in unexpected places. Overall, it contains about 320.000 entries, including only the infinitive of 22,600 verbs, that in Portuguese change with tense and person. If expanded, this would add more than a million entries; fortunately, most verbs follow a regular pattern of change, and this is not necessary. This verbs can be automatically expanded during their inclusion in the automaton representing the vocabulary.

Internally, the vocabulary will be represented by an augmented finite state automaton. The language accepted by the automaton will be the set of all words present in the Vocabulary, plus expanded verbal forms. Figure 3 presents an example automaton. In this figure, ignoring for now the boxes, the state numbered 0 is the initial state and the states marked with a double circle are possible final states. This representation is quite efficient, since it offers an implicit way of sharing prefixes and suffixes [Luc93]. It also offers the possibility of representing new words produced by proper derivation. An example automaton is shown in figure 2.

The automaton is augmented by grammatical and phonetic information. This allows the conversion of the input string as the automaton is making transitions (possibly non-deterministically). When a letter may have multiple phonetic transcriptions, the incoming symbol is expanded to include all the phonetic alternatives, denoted by the phonetic symbol appended to the letter. In this case, the automaton follows all the alternatives simultaneously. Each final state is augmented by the grammatical category of the recognized word. In figure 2, the boxes marked "verb" and "subs." (substantive) show examples of this situation.

When a letter may represent more than one phoneme, the preceding state will have as many transitions as the number of possible phonemes. In the example of figure 2, state 4 has two transitions associated with the letter O; one corresponds to the phoneme [o], and the other to [ ]. The other transitions with the letter O are unique, since all of then corresponds to the phoneme [o].

In the internal data structure, every state is represented by a tuple (*state*, *attribute*, *pointer*). The field *state* contains the number of the state; in the final minimized version, it is determined by the relative position of the other fields, and does not exist explicitly. The *attribute* field is also one byte long, and contains grammatical or other complementary information. The field *pointer* is three byte long and points to a transition list, formed by pairs (*letter*,*state*) indicating the next state corresponding to each incoming letter. In this table, the *letter* field is one byte long, and contains the letter, augmented by the phonetic variant, if the letter may correspond to several

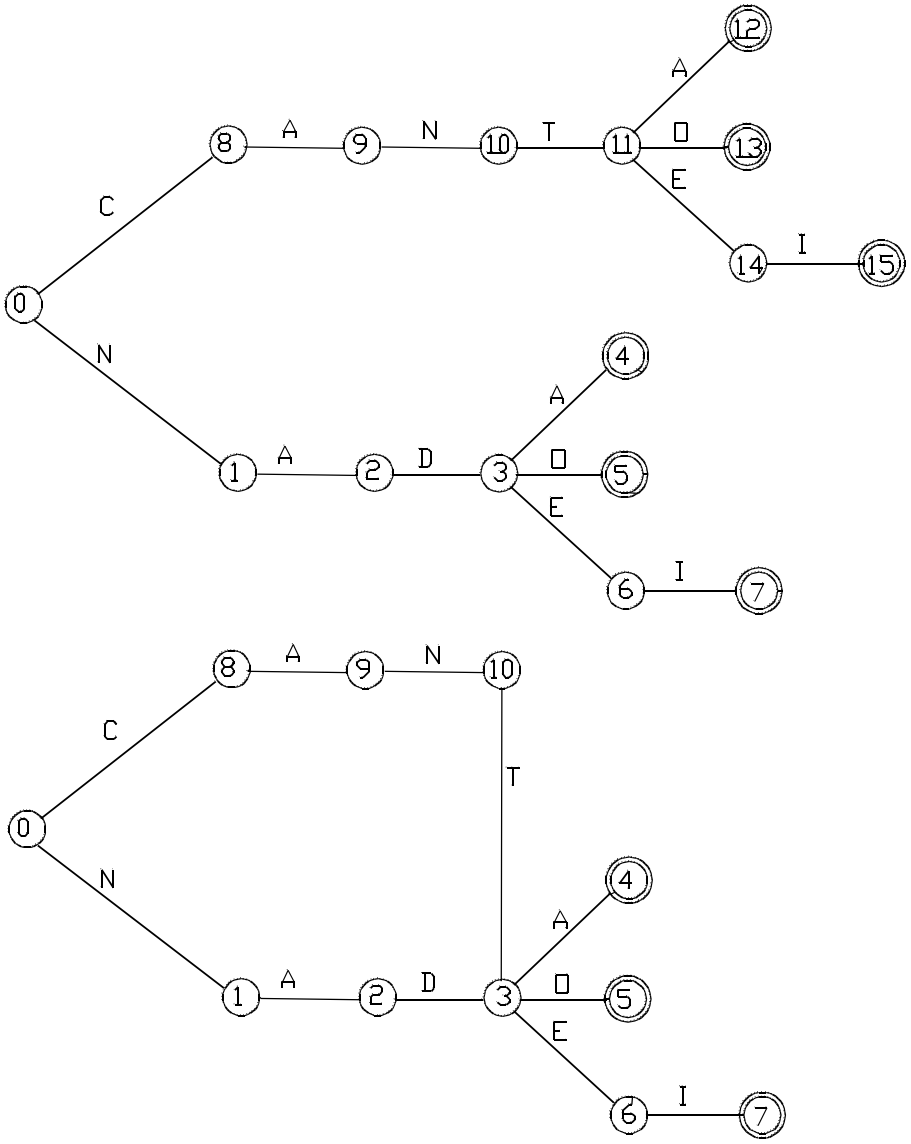

Figure 4 - Example of minimization
This figure shows an automaton recognizing the words canto, canta, cantei, nado, nada e nadei, both in the expanded and in the simplified forms.

phonemes (in Portuguese, there are at most four variants; this occurs with the letter X). Figure 3 shows the structure of the automaton corresponding to example in the figure 2.

The algorithm mounting the automaton makes initially a trie (a special form of a tree of letters, described in [Fre60]; the name trie comes from reTRIEval). After constructing it, formal methods are applied for minimizing the resulting automaton, by the reduction of equivalent states, as shown in figure 4.

Words are included in the automaton directly from the transcription of the "Vocabulário Ortográfico", except in the case of verbs, that are automatically conjugated according to the usual rules of the Portuguese grammar. The representation of letters is always phonetic, and a word appears as many times as its possible variants, either in phonetic representation or grammatical category. The phonetic translation of each word is determined either from information contained in the vocabulary or by the application of the rules presented in the next section.

4 Conversion rules

Although there are many works on text-to-speech or text-to-phonemes systems for English [Ain73] [All87] [Coi89] [Elo76] [Kla87], we couldn't identify previous work in the formal conversion from written words to sequences of phonemes for the Portuguese language.

Although this is much simpler than the analogous problem for English, since in Portuguese most letters have a fixed mapping to a phoneme, there are a few difficult points:

1- Letters E and O

These letters can be pronounced in open [ε], [ ] or closed [e], [o] variants when they are at the stressed syllable. There are no fixed rules, but the "Vocabulário Ortográfico" usually contains this phonetic indication. The main difficulty here is that the "Vocabulário Ortográfico" contains, for each word, only one indication for this case; nevertheless, many occurrences of the affect letter may occur in the word. This happens because only one of the letter can have the possibility of different phonetic translation. The determination of which letter is obvious by inspection, but a formal approach is somewhat difficult, and some heuristics are required.

2- Letter X

This letter represents one of four phonemes: [s], [ks], [s] and [z]. Here, there are some rules:

[s] beginning of words;
after "N"
after ai, ei, ou

[s] if followed by a consonant

[z] words beginning with EX, followed by a vowel

These rules do not encompass all cases, so the phonetic information present in the vocabulary is also used.

3- Nasalization

Nasalization may occur with the vowel a [a], which becomes [ã] in the following situations:

- when stressed and before a nasal consonant [m], [n];

- always before a sequence of a nasal consonant and another consonant,

- always before a nasal consonant in the end of a word.

This seems simple, but there are conflicts with the following rule, and in some cases it might depend on the grammatical category of the word. It is necessary to distinguish, for instance ca[ã]minha, which is a substantive, and ca[a]minha, a verbal form. In these cases, the two alternatives are placed in the automaton; the selection will depend on a syntactical analysis, albeit naive, of the input sentence.

4 - Derivation with suffixes.

Some very frequent suffixes, like "-(z)inha" and "-(z)inho", making diminutives, and "-mente", making adverbs from adjectives, and a few less frequent, change the behavior of the last vowel of the radical in two important ways:

- Rule 1 is applied to this last vowel, independent of stressing.

- Rule 3 above is applied independently of the syllable be stressed or not;

The vocabulary has errors and omissions in the part of phonetic specifications, but this seems to happen in a small number of words infrequently used. Errors are being corrected as detected, but no effort are being undertaken to quantify them.

5 Conclusions

This paper presents a system for converting written text to phonemes in Portuguese. Although the conversion is simpler than in some other languages as English, it shows some complexity and a vocabulary is essential for a large number of words. The system will be used as part of a text-to-speech system under development.

It is important to note that the automaton, augmented by the phonetic and grammatical information described in this paper, can be used also for voice recognition, since the phonetic information is an integral part of the state changing mechanism.

The transcription of the vocabulary and its codification is an important part of the work. This vocabulary will be put in the public domain once certain legal aspects are clarified.

Bibliography

Ain73 Ainsworth, W. A., A system for Converting English Text into Speech. IEEE Trans. on Audio and Electroacoustics, **21(3)**:288-240, June 1973.

All87 Allen, J., Synthesis of Speech from unrestricted text, Proc. of the IEEE, **64(4)**:433-422, April 1987.

Cal90 Callou, D. and Leite, Y.,Iniciação à Fonética e à Fonologia, Jorge Zahar Editor, Rio de Janeiro, 1990.

Coi89 Van Coile, B. M., The DEPES Development system for Text-to Speech Synthesis, in Proc. of the Inter. Conf. on Acoustics, Speech and Signal Processing, 250-253, 1989.

Elo76 Elovitz, H. S. et al, Letter-to-sound Rules for Automatic Translation of English text to Phonetics, IEEE Trans. on Acoustics, Speech and Signal Processing, ASSP **24(6)**:446-459, Dec. 1976

Fre60 Fredkin, E., Trie Memory, Comm. of the ACM, **3(9)**:490-500, Sept 1960.

Gro89 Gross, M., The Use of Finite Automata in the Lexical Representation of Natural Language, in M. Gross and D. Perrin, editors, Electronic Dictionaries and Automata in Computational Linguistics, 34-50, Springer-Verlag, Berlin, 1989. Lecture Notes in Computer Science, vol 377.

Her87 Hertz, S. R. et al, The Delta Rule Development System for Speech Synthesis from Text. Proc. of the IEEE, **73(3)**:737-793, Sept. 1987

Kla87 Klatt, D. H., Review of test-to-Speech Conversion for English, Journal of the Acoustical Society of America, **82(3)**:737-793, Sept 1987

Luc93 Lucchesi, C. L., and Kowaltowsky, T., Applications of Finite Automata Representing Large Vocabularies, Software - Practice and Experience, **33(1)**, Jan 1993.

Sam78 Samburg, M.R. et al, On reducing the buzz in LPC synthesis, J. Acoust. Soc. Am. **63(3)**, Mar 1978