AN HMM-BASED BRAZILIAN PORTUGUESE SPEECH SYNTHESIZER AND ITS CHARACTERISTICS

R. Maia, H. Zen, K. Tokuda, T. Kitamura, F.G.V. Resende Jr.

Abstract - Research on speech synthesis area has made great progress recently, perhaps motivated by its numerous applications, of which text-to-speech converters and dialog systems are examples. Several improvements have been reported in the technical literature related to existing state-of-the-art techniques as well as in the development of new ideas related to the alteration of voice characteristics, with their eventual application to different languages. Nevertheless, in spite of the attention that the speech synthesis field has been receiving, the technique which employs unit selection and concatenation of waveform segments still remains as the most popular approach among those available nowadays. In this paper, we report how a synthesizer for the Brazilian Portuguese language was constructed according to a technique in which the speech waveform is generated through parameters directly determined from Hidden Markov Models. When compared with systems based on unit selection and concatenation, the proposed synthesizer presents the advantage of being *trainable*, with the utilization of contextual factors including information related to different levels of the following acoustic units: phones, syllables, words, phrases and utterances. Such information is brought into effect through a set of questions for context-clustering. Thus, both the spectral and the prosodic characteristics of the system are managed by decision-trees generated for each one of the following parameters: mel-cepstral coefficients, fundamental frequency and state durations. As a typical characteristic of the technique based on Hidden Markov Models, synthesized speech with quality comparable to commercial applications built under the unit selection and concatenation approach can be obtained even from a database as small as eighteen minutes of speech. This was tested by a subjective comparison of samples from the synthesizer in question and other systems currently available for Brazilian Portuguese.

Keywords: Speech Processing, Text-to-Speech (TTS) Systems, Speech Synthesis, Hidden Markov Model (HMM).

Resumo - A pesquisa na área de síntese de voz tem alcançado grande progresso recentemente, provavelmente motivada por suas inúmeras aplicações, dentre as quais se pode citar conversores texto-voz e sistemas de diálogo. Muitas melhorias nas técnicas de estado-da-arte existentes, assim como o desenvolvimento de novas idéias relacionadas a alterações das características da voz sintetizada, seguidas por suas respectivas aplicações a diferentes idiomas, são descritos na literatura técnica. No entanto, apesar da atenção que a área de síntese de voz tem recebido, a técnica que consiste na seleção e concatenação de unidades de forma de onda ainda permanece como a mais empregada atualmente. Neste artigo descreve-se a construção de um sintetizador para o português brasileiro, baseado em uma técnica na qual o sinal de voz é gerado por parâmetros diretamente obtidos a partir de Modelos Escondidos de Markov. Quando comparado a sistemas que utilizam o método de seleção e concatenação de formas de onda, o sintetizador em questão apresenta a vantagem de ser *treinável*, com o uso de fatores contextuais que incluem informações referentes aos diferentes níveis das seguintes unidades acústicas: fone, sílaba, palavra, frase e período. Tais informações são efetivadas através de um conjunto de perguntas usadas para uma técnica de agrupamento de contextos. Portanto, as características espectrais e prosódicas do sistema são controladas por árvores-de-decisões correspondentes a cada um dos seguintes parâmetros: coeficientes mel-cepestrais, freqüência fundamental e duração de estados. Como uma propriedade típica do método de síntese de voz baseado em Modelos Escondidos de Markov, pode-se obter voz sintetizada com qualidade comparável à de algumas aplicações comerciais, construídas de acordo com a técnica de seleção e concatenação de unidades, mesmo para uma base de dados tão pequena quanto dezoito minutos de voz. Isto foi testado através de uma avaliação subjetiva de amostras geradas pelo sintetizador em questão e por outros sistemas disponíveis para o português brasileiro.

Palavras-chave: Processamento de Voz, Sistemas de Conversão Texto-Voz (TTS), Síntese de Voz, Modelos Escondidos de Markov (HMM).

1. INTRODUCTION

The speech synthesis area has been stimulating great interest for speech processing researchers in the last years. Aside from topics related to multilingual speech synthesis, with the attempt at designing unified TTS engines which could possibly work on different languages [1], the tendency nowadays has also been driven towards the synthesis of voices with different styles and emotions [2, 3].

Although a few speech synthesis techniques exist, the approach wherein speech is synthesized through the selection and concatenation of waveform units has been largely applied [4, 5]. One of its main advantages when compared with the other techniques is the fact that synthesized speech with high quality can be achieved due to the utilization of natural speech waveforms as units for concatenation, selected

R. Maia is with National Institute of Information and Communications Technology (NiCT), and ATR Spoken Language Communication Laboratories (ATR-SLC), Kyoto, Japan.

H. Zen, K. Tokuda and T. Kitamura are with Dept. of Computer Science, Nagoya Institute of Technology, Nagoya, Japan.

F.G.V. Resende Jr. is with Dept. of Electronic Engineering and Computer Science, and Program of Electrical Engineering, Federal University of Rio de Janeiro, Rio de Janeiro, Brazil.

according to some specific cost functions. Nevertheless, for this technique the synthesis of voices with different styles and emotions, as well as the obtainment of high quality itself requires the availability of large corpora.

Recently, a *trainable* approach in which the speech waveform is synthesized from parameters directly derived from Hidden Markov Models (HMMs) has been reported to work well for some languages [6–8]. One of the main advantages of the referred HMM-based synthesis technique when compared with the unit selection and concatenation method is the fact that voice alteration can be performed with no need of large databases [9–11]. Another advantage is that synthesized speech with *applicability*[1] can be achieved by training the system with a database as small as eighty sentences, as reported in [8]. Besides, still considering small databases (about one hour of speech), HMM-based synthesizers could possibly be competitive in quality to unit selection and concatenation ones [12, 13]. On the other hand, one of the main disadvantages of the referred approach corresponds to the *buzzy* quality of the synthesized speech. This drawback is caused by the source-filter model which is used during the waveform generation stage, which basically consists in a linear predictive vocoder, though in [14] it is reported that the mentioned *buzz* can be removed with the utilization of a mixed excitation scheme. Another approach to solve this problem is shown in [12], which consists in an adaptation from the vocoding method introduced in [15] to HMM-based speech synthesis.

Turning to the Portuguese language case, considerable advance has been achieved on the speech synthesis area for both European (e.g. [16, 17]) and Brazilian (e.g. [18, 19]) dialects. However, the common aspect among most of the contributions given to the Portuguese language is the fact of being related to synthesizers based on the waveform selection and concatenation approach. In this paper, HMM-based Brazilian Portuguese speech synthesis [8, 20] is focused. In a more specific way, the contribution of this paper consists in the description and discussion of topics related to the application of the HMM-based speech synthesis approach to Brazilian Portuguese, namely: determination of a list of contextual factors, definition of an utterance information which enables the derivation of all the specified features, and elaboration of questions which can bring into effect all the factors, according to a tree-based context-clustering algorithm. Finally, in a way to learn the characteristics of the trained system as well as to empirically improve the list of features and questions, a rough inspection of the generated decision-trees is performed.

This paper is organized as follows: in Section 2 all the procedures carried out by the synthesizer engine are described, from the database training to the synthesis of a given utterance. Section 3 describes the aspects of HMM-based speech synthesis applied to Brazilian Portuguese. In Section 4, two subjective tests are presented: the first one concerns the perceptual importance of contextual factors related to syllable, stress, and part-of-speech (POS) [21], whereas the second test corresponds to an evaluation among the synthesizer in question and other systems for Brazilian Portuguese available nowadays. The conclusions are in Section 5.

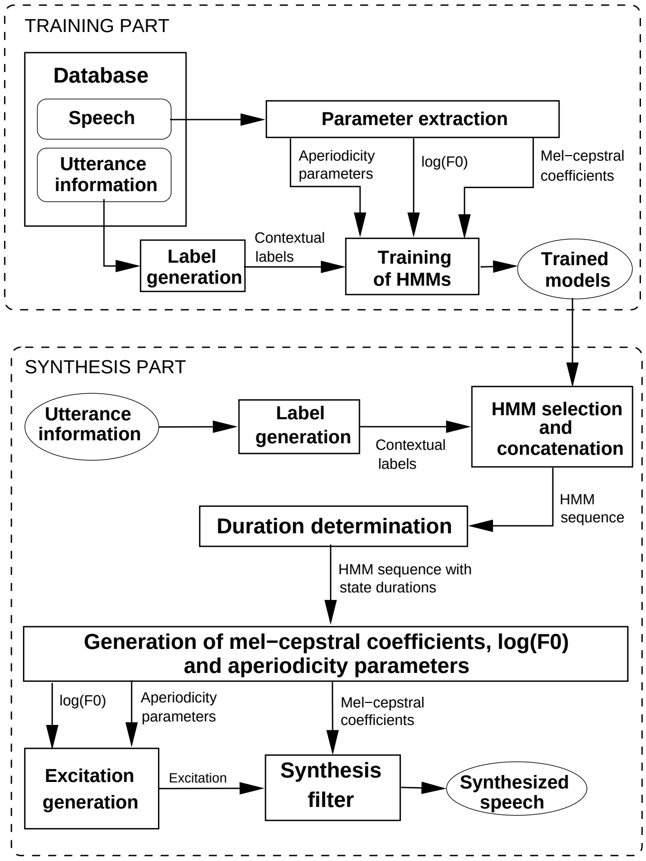

Figure 1. Block diagram illustrating the basic procedures conducted by the speech synthesis engine.

2. ENGINE DESCRIPTION

The procedures of training and synthesis carried out by the speech synthesis engine are depicted in the block diagram of Figure 1. The present engine corresponds to an improved version of the one already described in the literature, e.g. [6–8]. The enhancements correspond to the: (1) application of the high-quality vocoding method described in [15]; (2) utilization of HMMs with explicit state durations (*Hidden Semi-Markov Models*) [22]; and (3) generation of parameters considering global variance [23]. These improvements are described with more details in [12]. In the following sections an outline of the whole engine is given.

2.1 TRAINING PART

The synthesizer is trained through the following steps: (1) speech parameter extraction; (2) label generation; and (3) HMM training.

2.1.1 SPEECH PARAMETER EXTRACTION

The training is started with parameter extraction. In this step, initially a sequence of fundamental frequency logarithms, $\{\log(F0^1), \ldots, \log(F0^N)\}$, including voicing deci-

[1] Meaning that it could possibly be employed by some applications due to the naturalness of the synthesized speech prosody.

sion information (if $F0 = 0$ the frame is considered unvoiced), where N is the total number of frames of all the utterances from the training database, is extracted in a short-time basis. After that, a sequence of mel-cepstral coefficient vectors which represent speech envelope spectra [24], $\{\mathbf{c}^1, \ldots, \mathbf{c}^N\}$, is obtained. Each mel-cepstral coefficient vector, $\mathbf{c}^i = [c_0^i \; \ldots \; c_M^i]^T$, where the superscript i indicates the frame number and $[\cdot]^T$ means transposition, is derived through an M-th order mel-cepstral analysis, taking into account the already extracted sequence of $\log(F0)$ in order to remove signal periodicity [15]. Finally, a sequence of aperiodicity coefficient vectors, $\{\mathbf{b}^1, \ldots, \mathbf{b}^N\}$, is also obtained from all the utterances at the same rate as the mel-cepstral coefficients and $\log(F0)$. Each vector $\mathbf{b}^i = [b_1^i \; \ldots \; b_5^i]^T$ contains the aperiodicity measures for the following frequency sub-bands: 0-1 kHz, 1-2 kHz, 2-4 kHz, 4-6 kHz and 6-8 kHz. The procedures of spectral analysis smoothed by fundamental frequencies and aperiodicity extraction are performed as described in [15].

2.1.2 LABEL GENERATION

In this step, utterance information for all the sentences of the training database are converted into HMM contextual labels. The descriptions of utterance information and contextual label are given in sections 3.2 and 3.3, respectively.

2.1.3 HMM TRAINING

Each HMM corresponds to a no-skip S-state left-to-right model, with $S = 5$. Each output observation vector $\mathbf{o}^i$ for the i-th frame consists of five streams, $\mathbf{o}^i = [\mathbf{o}_1^{i\,T} \; \cdots \; \mathbf{o}_5^{i\,T}]^T$, as illustrated in Figure 2, where:

- stream 1 ($\mathbf{o}_1^i$): vector composed of mel-cepstral coefficients, $\{c_0^i, \ldots, c_M^i\}$, their corresponding delta, $\{\Delta c_0^i, \ldots, \Delta c_M^i\}$, and delta-delta components, $\{\Delta^2 c_0^i, \ldots, \Delta^2 c_M^i\}$;
- streams 2, 3, 4 ($\mathbf{o}_2^i, \mathbf{o}_3^i, \mathbf{o}_4^i$): composed respectively of the fundamental frequency logarithm, $\log(F0^i)$, its corresponding delta, $\Delta \log(F0^i)$, and delta-delta, $\Delta^2 \log(F0^i)$;
- stream 5 ($\mathbf{o}_5^i$): vector composed of aperiodicity coefficients, $\{b_1^i, \ldots, b_5^i\}$, their corresponding delta, $\{\Delta b_1^i, \ldots, \Delta b_5^i\}$, and delta-delta, $\{\Delta^2 b_1^i, \ldots, \Delta^2 b_5^i\}$.

The observation vector $\mathbf{o}^i$ is output by an HMM state s according to a probability distribution given by

$$\beta_s(\mathbf{o}^i) = \prod_{j=1}^{5} \left[\sum_{l=1}^{R_j} \omega_{sjl} \mathcal{N}(\mathbf{o}_j^i; \mu_{sjl}, \mathbf{\Sigma}_{sjl}) \right]^{\gamma_j}, \quad (1)$$

where $\mathcal{N}(\cdot; \mu, \mathbf{\Sigma})$ means Gaussian distribution with mean μ and variance Σ, ω_{sjl} is the weight for the l-th mixture component of the j-th stream vector $\mathbf{o}_j^i$ output by the state s, and γ_j is the output probability weight for the j-th stream, with R_j being the corresponding number of mixture components. The first and fifth stream vectors, $\mathbf{o}_1^i = [\mathbf{c}^{i\,T} \; \Delta \mathbf{c}^{i\,T} \; \Delta^2 \mathbf{c}^{i\,T}]^T$ and $\mathbf{o}_5^i = [\mathbf{b}^{i\,T} \; \Delta \mathbf{b}^{i\,T} \; \Delta^2 \mathbf{b}^{i\,T}]^T$, are modeled by single-mixture continuous Gaussian distributions, where the dimensionality is $3(M+1)$ for $\mathbf{o}_1^i$ and fifteen for $\mathbf{o}_5^i$. For the second, third and fourth scalar streams, $o_2^i = \log(F0^i)$, $o_3^i = \Delta \log(F0^i)$, and $o_4^i = \Delta^2 \log(F0^i)$, the output probability is modeled by multi-space Gaussian distributions [25] with two mixture components.

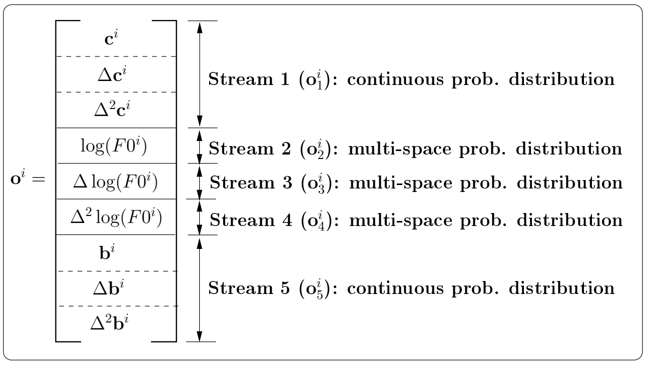
$\mathbf{o}^i =$
$\mathbf{c}^i$
$\Delta \mathbf{c}^i$
$\Delta^2 \mathbf{c}^i$
$\log(F0^i)$
$\Delta \log(F0^i)$
$\Delta^2 \log(F0^i)$
$\mathbf{b}^i$
$\Delta \mathbf{b}^i$
$\Delta^2 \mathbf{b}^i$
Stream 1 ($\mathbf{o}_1^i$): continuous prob. distribution
Stream 2 ($\mathbf{o}_2^i$): multi-space prob. distribution
Stream 3 ($\mathbf{o}_3^i$): multi-space prob. distribution
Stream 4 ($\mathbf{o}_4^i$): multi-space prob. distribution
Stream 5 ($\mathbf{o}_5^i$): continuous prob. distribution

Figure 2. Structure of each feature vector $\mathbf{o}^i$: modeling of mel-cepstral coefficients, $\mathbf{c}^i$, fundamental frequency, $\log(F0^i)$, and aperiodicity parameters, $\mathbf{b}^i$, in a single HMM framework.

For each HMM k, the durations of the S states are considered as vectors, $\mathbf{d}^k = [d_1^k \; \cdots \; d_S^k]^T$, where d_s^k represents the duration of the s-th state. Further, each of the duration vectors, $\{\mathbf{d}^1, \ldots, \mathbf{d}^K\}$, where K is the total number of HMMs representing the database, is modeled by an S-dimensional single-mixture Gaussian distribution. The output probabilities of the state duration vectors are thus re-estimated by Baum-Welch iterations in the same way as the output probabilities of the speech parameters [22].

During the training, a context-clustering technique is applied to the streams of mel-cepstral coefficients, $\log(F0)$ and aperiodicity parameters, as well as to the state duration models. In the end of the process, $3S + 1$ different acoustic decision-trees are generated: S trees for mel-cepstral coefficients (one tree for each state s), S trees for the logarithms of fundamental frequencies (one tree for each state s), S trees for aperiodicity parameters (one tree for each state s), and finally one tree for state duration.

2.2 SYNTHESIS PART

The procedure of synthesis of a given sentence into the corresponding speech is conducted through the following steps: (1) label generation; (2) HMM selection and concatenation; (3) parameter determination; and (4) excitation construction and filtering.

2.2.1 LABEL GENERATION AND HMM SELECTION/CONCATENATION

The synthesis procedure starts with the conversion of the utterance information of a given sentence into contextual labels, which are eventually used to select corresponding leaves from each one of the $3S + 1$ decision-trees generated by the context-clustering procedure in the training stage. In the end of this step, four logical HMM sequences, whose states correspond to the selected leaves, are derived for each of the following parameters: (1) mel-cepstral coefficients; (2) logarithms of fundamental frequencies; (3) aperiodicity coefficients; and (4) state durations. The sequences for items (1), (2) and (3) consist of HMMs with S states, whereas the sequence for state durations comprises single-state HMMs.

2.2.2 PARAMETER DETERMINATION

The four above mentioned HMM sequences are then used to derive mel-cepstral coefficients, $\log(F0)$ and aperiodicity parameters. The whole procedure is conducted as follows. Initially, the duration vectors $\{\mathbf{d}^1, \ldots, \mathbf{d}^K\}$, where K is the number of HMMs in each sequence, are determined from the K S-dimensional Gaussian distributions, defining the state sequence $\mathbf{s} = \{s_1, \ldots, s_L\}$, with L being the number of frames of the utterance to be synthesized and s_i the HMM state wherein the i-th frame belongs to. After that, mel-cepstral coefficient vectors, $\{\mathbf{c}^1, \ldots, \mathbf{c}^L\}$, aperiodicity parameters, $\{\mathbf{b}^1, \ldots, \mathbf{b}^L\}$, and logarithms of the fundamental frequencies, $\{\log(F0^1), \ldots, \log(F0^L)\}$, are determined from each corresponding HMM sequence in a way to maximize their output probability given $\mathbf{s}$, taking into account the delta and delta-delta components, according to the algorithm described in [23].

2.2.3 EXCITATION CONSTRUCTION AND FILTERING

The last step of the synthesis process is divided into two parts. In the first one an excitation signal is derived from the sequences of generated fundamental frequency logarithms, $\{\log(F0^1), \ldots, \log(F0^L)\}$, and aperiodicity parameters, $\{\mathbf{b}^1, \ldots, \mathbf{b}^L\}$, using the same approach described in the high-quality vocoding method of [15], which is based on mixed excitation construction according to frequency subband strengths. In the second part, speech waveform is generated with the utilization of the Mel Log Spectrum Approximation (MLSA) filter [24], whose corresponding coefficients are derived from the sequence of generated mel-cepstral coefficients, $\{\mathbf{c}^1, \ldots, \mathbf{c}^L\}$.

3. ASPECTS OF BRAZILIAN PORTUGUESE SPEECH SYNTHESIS BASED ON HMM

3.1 THE PHONE SET

The synthesizer employs a set of 40 phones - including long and short pause models - as the basic acoustic units, which are shown in Table 1 represented with the use of SAMPA (Speech Assessment Methods Phonetic Alphabet) [26]. Although diphthongs are sometimes considered as independent acoustic units due to their peculiar characteristics [27], where the formants of the initial vowel/semi-vowel are smoothly changed into the formants of the succeeding semi-vowel/vowel, in this work they were not considered as so, following the example of other phone sets reported for Brazilian Portuguese speech synthesis [28, 29].

Table 1. Phone set employed by the synthesizer as the basic acoustic units, with some corresponding examples.

Phone	*Examples*
Oral vowels	
a	jatobá, capacete, cabeça, lua
E	é, pajé, pele, ferro, velho
e	capacete, resolver, respeito
i	justiça, país, lápis, idiota, aqueles, ele
O	ópio, jogos, sozinho, forte
o	jogo, golfinho, cor
u	raul, culpa, baú, cururu, logo
Nasal vowels	
a~	andar, tampar, canção, cama
e~	então, tempo, bem, menos
i~	ninho, tinta, latina, importa
o~	onda, campeões, somos , homem
u~	um, muito, umbigo
Semi-vowels	
w	fácil, voltar, eu, quase
j	pai, foi, caracóis, micróbio
w~	não, cão
j~	muito, bem, parabéns, compõe
Unvoiced fricatives	
f	festa, fanfarrão, afta, afluente
s	sapo, caçar, crescer, sessão, lápis, capaz, casca, excesso
S	chá, xaveco, cachorro
Voiced fricatives	
z	casa, coisa, quase, exato
v	vovó, vamos, avião
Z	geladeira, trovejar
Affricates	
tS	tia, pacote, constituinte
dZ	dia, cidade, disco
Plosives	
b	barba, absinto
d	dados, administrar
t	pato, constituinte
k	casca, quero, quanto
g	guerra, gato, agüentar, agnóstico
p	papai, psicólogo, apto
Liquids	
l	laranja, leitão
L	calhar, colheita, melhor
R	carro, rua, rato, carga, germe
X	casar, certo, arpa, arco
r	garoto, frango, por exemplo
Nasal consonants	
m	mamãe, emancipar
n	nome, atenuar, encanação
J	casinha, galinha
Silences	
sil	beginning and end of utterance
pau	pauses

3.2 DEFINITION OF AN UTTERANCE INFORMATION

For the present synthesizer, utterance information corresponds to the basic text knowledge which is input by the system in order to generate speech. The henceforth defined utterance information is thus composed of the following parts:

- phone part: phone symbol;
- syllable part: syllable transcription and stress indication;

Table 2. Utterance information for "*Leila tem um lindo jardim*" (*Leila has a beautiful garden*).

phone	syll	stress	word	class
sil				
l	lej	1	lejla	content
e				
j				
l	la	0		
a				
t	te~j~	0	te~j~	content
e~				
j~				
u~	u~	0	u~	function
l	li~	1	li~du	content
i~				
d	du	0		
u				
Z	Zax	0	Zaxdi~	content
a				
X				
d	di~	1		
i~				
sil				

- word part: word transcription and POS tag.

Table 2 shows the utterance information for the sentence "*Leila tem um lindo jardim*" (*Leila has a beautiful garden*).

3.2.1 TEXT PROCESSING: UTTERANCE INFORMATION CONSTRUCTION

Utterance information can be derived by a natural language processing (NLP) module. According to the definition of the utterance information, the NLP module is required to perform the following procedures: (1) grapheme-phone conversion; (2) syllabication; (3) stress determination; and (4) POS tagging.

Although NLP is out of scope of this paper, in the following paragraphs the procedures carried out by a text processor specifically designed for the present synthesizer are slightly outlined. Details of the NLP module can be found in [30–32].

Grapheme-phone conversion The grapheme-phone converter is rule-based [30] with a database of word exceptions. Most of these exceptions are composed of terms in which the transcription rules do not cover the problem of open-closed vowel alternation, though [33] presents some directions which could possibly solve this drawback. Special procedures are also applied in order to solve the problem of the homographs [31].

Syllabication and stress determination Even though considerable contributions have been reported concerning automatic syllabication for Portuguese TTS systems, e.g. [34,35], in the present case this task has been performed through the application of orthographic rules to the non-transcribed word tokens [36]. Stress is also determined before grapheme-phone conversion, according to the algorithm presented in [32].

POS tagging The NLP module classifies input words into two groups:

- *content* words (open classes): nouns, verbs, adverbs, adjectives;
- *function* words (closed classes): prepositions, conjunctions, articles, pronouns, interjections and connectives.

The method of classification consists in verifying if the input word belongs to a list of possible function words. If so, the word is classified as *function*, otherwise *content* [37].

3.3 CONTEXTUAL LABEL

3.3.1 THE CONTEXTUAL FACTORS

In speech synthesis, some factors are usually necessary to be taken into account in order to provide a natural reproduction of the prosody. These factors might include context dependent terms, such as preceding/succeeding phone, syllable, word, phrase, etc, and are referred to as *contextual factors* in this paper, though the reference *features* might also be employed [27].

The determination of contextual factors for a particular language is based on prosodic characteristics of the referred language and consequently linguistic assumptions should be considered. Besides this theoretical approach, empirical analysis can also be carried out in order to tune the features, by obtaining related extensions to the factors that are important and eliminating the ones which are not.

The contextual factors listed below, which correspond to the ones employed by the present synthesizer, were firstly derived from those used in HMM-based English speech synthesis [7] and eventually adjusted, through theoretical and empirical approaches, to the characteristics of the Brazilian Portuguese language:

- phone level:
  1. {pre-preceding, preceding, current, succeeding, post-succeeding} phone;
  2. position of current phone in current syllable;
- syllable level:
  1. whether or not {preceding, current, succeeding} syllable is stressed;
  2. number of phones in {preceding, current, succeeding} syllable;
  3. position of current syllable in current word;
  4. number of stressed syllables in current phrase {before, after} current syllable;
  5. number of syllables, counting from previous stressed to current syllable in the utterance;
  6. number of syllables, counting from current to next stressed syllable in the utterance;
- word level:
  1. part-of-speech of {preceding, current, succeeding} word;
  2. number of syllables in {preceding, current, succeeding} word;

3. position of current word in current phrase;
4. number of content words in current phrase {before, after} current word;
5. number of words counting from previous content word to current word in the utterance;
6. number of words counting from current to next content word in the utterance;

- phrase level:

  1. number of {syllables, words} in {preceding, current, succeeding} phrase;
  2. position of current phrase in current utterance;

- utterance level:

  1. number of {syllables, words, phrases} in the utterance.

3.3.2 FORMAT OF THE CONTEXTUAL LABEL

The contextual labels include all the information listed in Section 3.3.1 in a phone-by-phone basis. In other words, for each phone of the input utterance information the whole set of features related to the respective phone is included into the corresponding label. Table 3 describes the label format for the synthesizer.

3.4 CONTEXT CLUSTERING

3.4.1 THE PROBLEM

Since for each phone from the speech database there is a corresponding contextual label which includes all its related features, it can be noticed that there is a wide range of different contextual labels which can result during the training stage of the synthesizer. Thus, it would be impractical to train a large amount of different HMMs for a relatively small database, and consequently the resulting models would not be adequately re-estimated during the training process. This problem can be enlightened through the following example. Considering the utterance information shown in Table 2, the corresponding contextual label for the phone */e/* of the word */lejla/* would be:

```
sil^l-e+j=l/M2:2_2
/S1:y_@y-1_@3+0_@2/S2:1_2/S3:1_8/S4:0_2/S5:0_5/S6:e
/W1:y_#y-content_#2+content_#1/W2:1_5/W3:0_3/W4:0_2
/P1:y_!y-8_!5+y_!y/P2:1_1
/U:8_$5_&1
```

where the letter *y* means *does not apply*. Because of the contextual information attached to the phone */e/*, which goes further to the utterance level, it is probable that only a few examples of the exact same label, if any at all, could be derived from the training corpus, although */e/* is one of the most frequent phones in Brazilian Portuguese.

Furthermore, during the synthesis stage, a given utterance information may generate some contextual labels which do not correspond to any model in the trained set of HMMs.

3.4.2 THE SOLUTION: TREE-BASED CONTEXT-CLUSTERING

In order to solve the problems discussed above, a decision tree-based context-clustering technique is applied [38]. This technique has the property of training models with a proportionally small database - solving the training problem, and constructing unseen models - solving the synthesis problem.

Because the contextual factors are responsible for the spectral and prosodic characteristics of the system, the importance of *how to cluster* these features should be stated. Therefore, the determination of the questions for context-clustering represents an important issue in order to achieve synthesized speech with good quality.

Questions about contextual factors Several questions are applied for each feature listed in Section 3.3.1. As an example of application, the questions for the feature "*position of current syllable in the word*" are listed:

- *Is current syllable in position 1 within the current word?*
- *Is current syllable in position 2 within the current word?*

⋮

- *Is current syllable in position 8 within the current word?*

Questions based on phonetic/phonemic characteristics For the phones, the questions are based on phonetic/phonemic characteristics of the Brazilian Portuguese language [36, 39]:

- voiced phones: vowels and voiced consonants;
- vowels and semi-vowels: anterior, central, posterior, high, middle, non-rounded, reduced, open, closed, oral and nasal;
- consonants: stop, constrictive, convex, concave, fricative, liquid, vibrant, bilabial, labiodental, dental, alveolar, palatal, velar, unvoiced, voiced, oral and nasal.

According to the classification above, some examples of questions are listed below:

- *Is current phone a voiced fricative?*
- *Is pre-preceding phone voiced?*
- *Is succeeding phone an oral semi-vowel?*
- *Is post-succeeding phone a convex alveolar consonant?*

Questions concerning diphthongs Questions regarding diphthongs are determined by considering diphones, in order to take advantage of the possible sequence of a vowel with its succeeding semi-vowel (descendant diphthong), or a semi-vowel with its succeeding vowel (ascendant diphthong). For example, for the question "*Is current phone part of a descendant diphthong?*," it would be true if the current phone was the vowel */e/* and its corresponding right context was the semi-vowel */j/*, forming thus the descendant diphthong */ej/*,

Table 3. Label format for the HMM-based Brazilian Portuguese synthesizer. "/Si:," "/Wi:" and "/Pi:" mean *i-th syllable, word and phrase-based contextual information part*, respectively, whereas "/U:" means *utterance-based contextual information part*.

```
m1^m2-m3+m4=m5/M2:m6_m7
/S1:s1_@s2-s3_@s4+s5_@s6/S2:s7_s8/S3:s9_s10/S4:s11_s12/S5:s13_s14/S6:s15
/W1:w1_#w2-w3_#w4+w5_#w6/W2:w7_w8/W3:w9_w10/W4:w11_w12
/P1:p1_!p2-p3_!p4+p5_!p6/P2:p7_p8
/U:u1_$u2_&u3
```

	Phone part
m1	phone before previous phone
m2	previous phone
m3	current phone
m4	next phone
m5	phone after next phone
m6	position of current phone in current syllable (forward)
m7	position of current phone in current syllable (backward)
	Syllable part
s1	whether previous syllable is stressed or not (0 → no; 1 → yes)
s2	number of phones in previous syllable
s3	whether current syllable is stressed or not (0 → no; 1 → yes)
s4	number of phones in current syllable
s5	whether next syllable is stressed or not (0 → no; 1 → yes)
s6	number of phones in next syllable
s7	position of current syllable in current word (forward)
s8	position of current syllable in current word (backward)
s9	position of current syllable in current phrase (forward)
s10	position of current syllable in current phrase (backward)
s11	number of stressed syllables before current syllable in current phrase
s12	number of stressed syllables after current syllable in current phrase
s13	number of syllables, counting from the previous stressed syllable to the current syllable in the utterance
s14	number of syllables, counting from the current syllable to the next stressed syllable in the utterance
s15	vowel of current syllable
	Word part
w1	part-of-speech classification of previous word
w2	number of syllables in previous word
w3	part-of-speech of classification of current word
w4	number of syllables in current word
w5	part-of-speech classification of next word
w6	number of syllables in next word
w7	position of current word in current phrase (forward)
w8	position of current word in current phrase (backward)
w9	number of content words before current word in current phrase
w10	number of content words after current word in current phrase
w11	number of words counting from the previous content word to the current word in the utterance
w12	number of words counting from the current word to the next content word in the utterance
	Phrase part
p1	number of syllables in previous phrase
p2	number of words in previous phrase
p3	number of syllables in current phrase
p4	number of words in current phrase
p5	number of syllables in next phrase
p6	number of words in next phrase
p7	position of current phrase in the utterance (forward)
p8	position of current phrase in the utterance (backward)
	Utterance part
u1	number of syllables in the utterance
u2	number of words in the utterance
u3	number of phrases in the utterance

like in the word "*comecei*" (*I began*). Some examples of diphthong questions and their corresponding way in which they were applied are listed below:

- *Is current phone the vowel of a diphthong? → Is current phone a vowel and the succeeding or preceding one a semi-vowel?*
- *Does current phone form with right context a descendant diphthong? → Is current phone a vowel and the succeeding one a semi-vowel?*
- *Is right context an ascendant diphthong? → Is succeeding phone a semi-vowel and the post-succeeding one a vowel?*

- *Is left context a descendant diphthong?* → *Is pre-preceding phone a vowel and the preceding one a semi-vowel?*

Questions considering phones under specific contexts Aside from diphthongs, other questions concerning diphones, and even triphones are also taken into account in order to track some peculiar properties of the units under certain contexts. Some of the questions are related to: vowels in the end of utterances (vowel followed by silence), which normally are uttered with lower intonation and energy; inter-word sequence of vowels, which tend to concatenate themselves forming a diphthong, or even an allophonic realization of one of them; and vowels preceded by stops and followed by silences, e.g., the phone /i/ in the end of the word "*qualidade*" (*quality*) pronounced by a native of the northeastern part of Brazil. However, these questions might be effective only if these *special situations* occur in the recorded database since the HMM-based speech synthesis technique tends to mimic the characteristics of the material from which the training is carried out.

3.5 SYSTEM IMPLEMENTATION

3.5.1 THE CORPUS

The text material used to train the synthesizer comprised the 200 phonetically balanced sentences for Brazilian Portuguese spoken in Rio de Janeiro listed in [40], and the 21 phonetically balanced utterances from the joint project reported in [29]. The sentences were recorded by a male Brazilian speaker. The recorded utterances correspond to 18 minutes and 48 seconds of speech including silence regions, where the average duration of each utterance is approximately five seconds with silence regions ranging 1 one to 2 seconds. The database was recorded at a sampling rate of 48 kHz with 16 bits per sample, being posteriorly down-sampled to 16 kHz.

The phonetic labeling of the database was carried out using the phone set shown in Table 1. Time label boundaries were obtained by manual correction of the label boundaries generated by Viterbi alignment. Further, syllable and word labeling as described by the utterance information in Section 3.2 were also manually conducted for each sentence. Thus, the database information was carefully included in a considerably time consuming process.

3.5.2 PARAMETER EXTRACTION

Fundamental frequencies, mel-cepstral coefficients and aperiodicity parameters were extracted from the speech corpus at every 5-ms frames. Mel-cepstral coefficients were obtained through a 39-th order analysis ($M = 39$) with the utilization of 25-ms Blackman windows. The computation of aperiodicity components and smoothing of mel-cepstral coefficients were carried out from speech and $F0$ in a way that the high-quality vocoding technique described in [15] could be applied during the synthesis. The dynamic parameters were derived according to

$$\Delta \mathbf{x}^i = \frac{1}{2}\left(\mathbf{x}^{i+1} - \mathbf{x}^{i-1}\right), \quad (2)$$

$$\Delta^2 \mathbf{x}^i = \mathbf{x}^{i-1} - 2\mathbf{x}^i + \mathbf{x}^{i+1}, \quad (3)$$

where $\mathbf{x}^i$ corresponds to the original feature ($\log(F0)$, mel-cepstral coefficients or aperiodicity parameters) for the i-th frame, and $\Delta \mathbf{x}^i$ and $\Delta^2 \mathbf{x}^i$ are the corresponding delta and delta-delta parameters, respectively. Each HMM had five states ($S = 5$).

Table 4. Number of leaves for each generated decision-tree. The total number of models according to the contextual labels (logical models) is 6273.

Tree	*State*					*Total*	*Reduction*
	1	2	3	4	5		$\left[\frac{physical}{logical}\right]\times 100$
Mel-cepstral	47	43	53	43	47	233	3.7%
$\log(F0)$	82	70	84	80	81	397	6.3%
Duration	-					104	1.7%

Table 5. Number of nodes with questions related to phone, syllable, stress and POS for each generated decision-tree.

Tree	*Total nodes*	*Piece of Information*			
		Phone	Syllable	Stress	POS
Mel-cepstral (total)	233	81.1%	15.9%	2.6%	0.4%
Mel-cepstral S_1	47	40	5	0	1
Mel-cepstral S_2	43	37	5	0	0
Mel-cepstral S_3	53	41	11	4	0
Mel-cepstral S_4	43	34	7	2	0
Mel-cepstral S_5	47	37	9	0	0
$\log(F0)$ (total)	397	58.9%	29.5%	9.8%	4.3%
$\log(F0)$ S_1	82	53	15	3	5
$\log(F0)$ S_2	70	44	20	7	2
$\log(F0)$ S_3	84	48	29	12	1
$\log(F0)$ S_4	80	40	28	11	7
$\log(F0)$ S_5	81	49	25	6	2
State duration	104	79.8%	15.4%	5.8%	2.9%
Total	734	68.9%	23.2%	6.9%	2.5%

3.5.3 GENERATED DECISION-TREES

Figure 3 and Figure 4 show respectively the top part of the decision-trees generated for mel-cepstral coefficients and $\log(F0)$ (for the third HMM state, i.e., $s = 3$) whereas Figure 5 shows the decision-tree generated for state duration, when the training procedure was concluded. Table 4 shows the number of leaves derived in the end of the process for each tree as well as the model reduction rate, which corresponds to the ratio between of number of states after and before performing context-clustering.

By observing Figure 3, Figure 4 and Figure 5, and based on the assumption that top nodes are more important considering the parameter which is being clustered, one can notice that knowledge regarding syllable, word, phrase and utterance are more crucial for $\log(F0)$ and state duration. On the other hand, questions related to phones are more significant to the tree of mel-cepstral coefficients.

Table 5 presents for each generated decision-tree the number of nodes concerning the information input to the synthesizer, namely: phone, syllable, stress, and POS. It should be

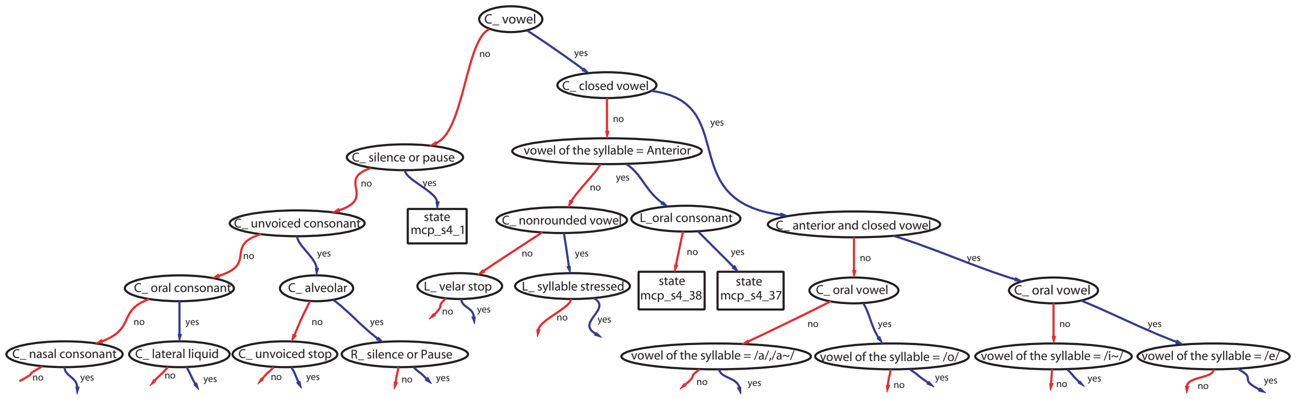

Figure 3. Top of the decision-tree constructed to cluster the third HMM state for mel-cepstral coefficients. The terms "C_", "L_" and "R_" stand for *current*, *left*, and *right* contexts, respectively.

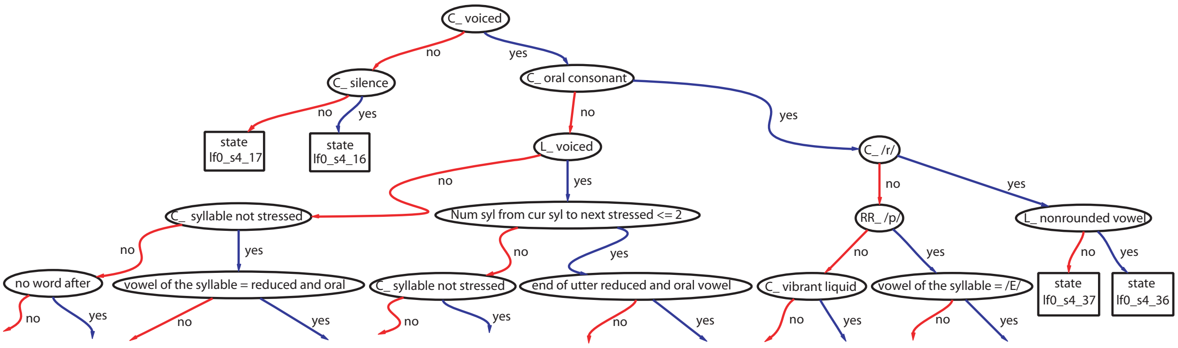

Figure 4. Top of the decision-tree constructed to cluster the third HMM state for $\log(F0)$. The terms "C_", "L_", "R_" and "RR_" stand for *current*, *left*, *right* and *after the right* contexts, respectively.

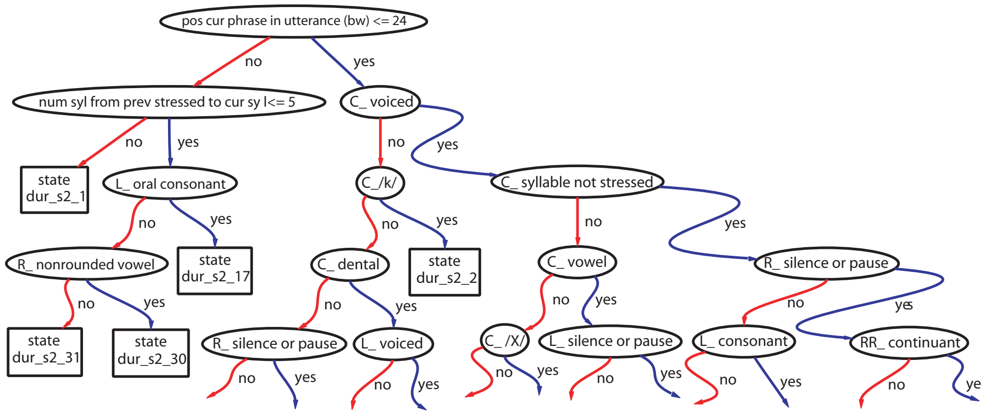

Figure 5. Top of the decision-tree constructed to cluster the distribution for state durations. The terms "C_", "L_", "R_" and "RR_" stand for *current*, *left*, *right* and *after the right* contexts, respectively.

noted that nodes regarding phrase, utterance and word (aside from POS) are not listed. From Table 5 the importance of phone, syllable, stress and POS for the quality of the synthesized speech can be figured out. Thus, assuming the total number of questions related to each of these pieces of information as the evaluation parameter, the following order of importance can be retrieved: (1) phone; (2) syllable; (3) stress; and (4) POS.

3.5.4 EXAMPLE OF SYNTHESIS

Figure 6 shows the spectrograms for the natural utterance "*Quando eu vim para cá, eu sempre gostei de jogar futebol*" (*Since I came here, I have always enjoyed playing football*) and its synthesized version. It should be noted that the referred sentence was not part of the training database[2]. Aside from the reproduction of the phones, it can also be observed from Figure 6 that the synthesized version presents speaking

[2] More synthesized samples can be found at `http://kt-lab.ics.nitech.ac.jp/˜maia/demo`.

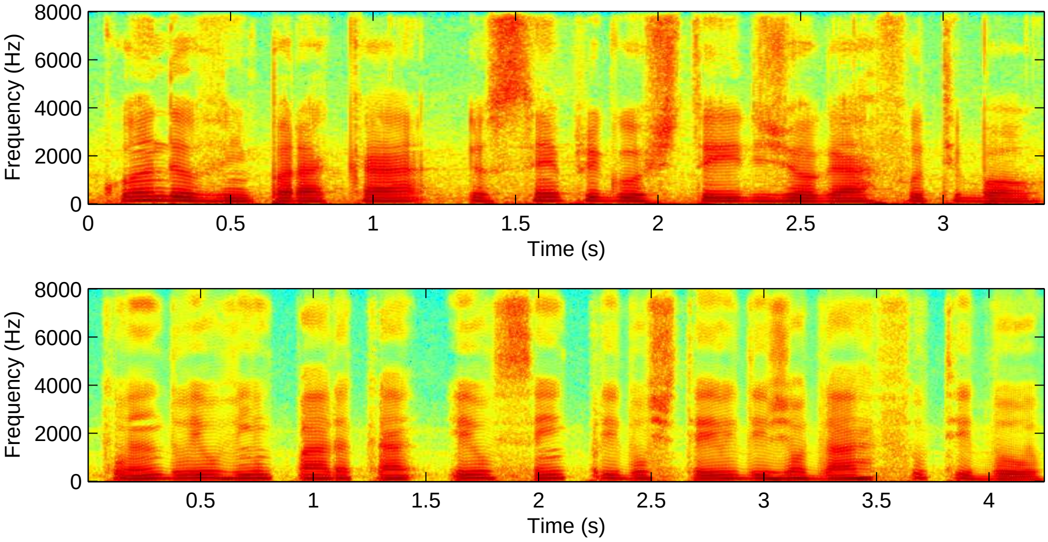

Figure 6. Examples of spectrograms for the utterance "*Quando eu vim para cá, eu sempre gostei de jogar futebol*" (*Since I came here, I have always enjoyed playing football*). Top: natural speech; bottom: synthesized speech.

rate similar to the natural speech case. This represents an important characteristic of the HMM-based speech synthesis approach: the ability to mimic the prosody of the speech corpus used to train the system.

Although it has been reported that even with a database as small as eighty utterances it is possible to synthesize speech [8], the lack of database strongly affects the quality. Once the HMMs do not track properly the characteristics of the several contextual labels derived from the training database, inconsistent parameters might be generated during the synthesis part, consequently resulting into synthesized speech with poor quality. Because of the trade-off between number of contextual labels and speech material, the severity of this problem might depend on the language. For example, for Japanese, which has a small phonetic system, intelligible synthesized speech (although with a badly reproduced prosody) can be achieved even by the training of sixty utterances.

4. SUBJECTIVE TESTS

4.1 INFLUENCE OF SOME CONTEXTUAL FACTORS ON THE SYNTHESIZED SPEECH

A subjective evaluation was conducted in order to investigate the importance for the synthesized speech of: (1) POS; (2) syllable; and (3) syllable stress. In addition to the empirical improvement on the speech quality by tuning the contextual factors, the analysis had also contributed to the development of the NLP module which has been applied to the synthesizer.

The tests were divided in two in order to not cause fatigue to the listeners who took part in it. The decision of which factors should be evaluated in the first and second tests was taken so as to match the difficulty levels of the NLP module. Hence, the first test attempted to verify the importance of carrying out or not POS tagging and/or syllabification. The second test, which concerns the importance of syllable stress, was conducted in order to investigate the need of proceeding or not beyond syllabification.

4.1.1 INFLUENCE OF POS AND SYLLABLE

To evaluate the influence of POS and syllable on the synthesized speech, the synthesizer was built under the conditions of no POS and no syllable information. This was achieved by training the original system with the exclusion of all the questions for context-clustering which concerned POS and syllable, respectively. In this way, all the features related to POS and syllable were automatically not taken into account during the training and synthesis procedures.

The perceptual evaluation corresponded to a forced AB comparison test whose procedure is described as follows. Each test sentence was synthesized into three different utterances: original, no POS, and no syllable information versions. The resulting utterances were combined in pairs so that three different test pairs for each test sentence were obtained. Each subject had to listen, for each test sentence, to the three pairs of utterances, eventually giving for each pair his/her preference concerning which utterance presented better quality. The order of the pairs as well as the order of the utterances within the respective pairs were randomly chosen for each listener. In case of undistinguished quality the subjects were instructed to choose the first synthesized utterance out of the corresponding test pair, so that options A and B could have equal probability of being chosen.

To perform the test, a total of ten sentences which were not used to train the system were randomly chosen from a list of twenty phrases, that were selected from a newspaper database through a genetic algorithm [41]. Thus, each listener had to listen to thirty different pairs of utterances. Among the eleven Brazilian listeners who participated in the test, four of them

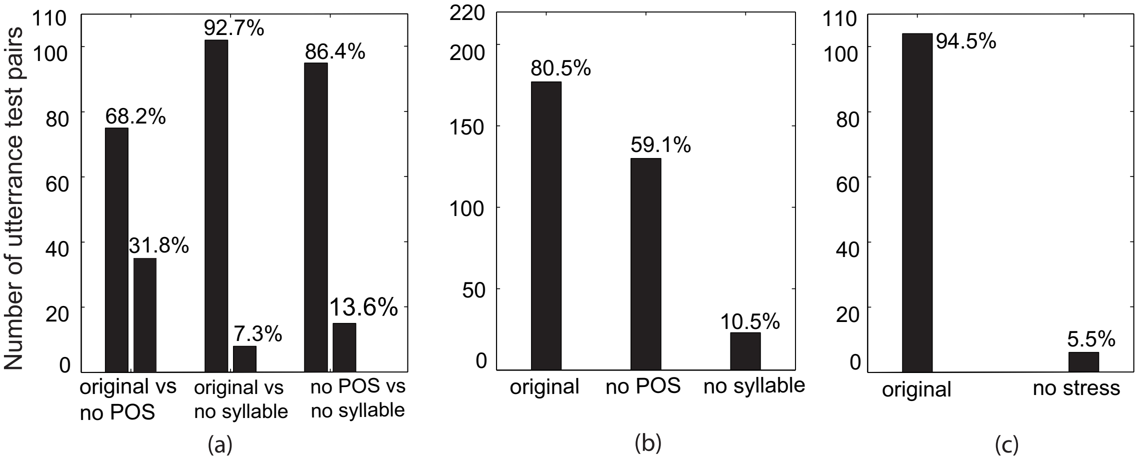

Figure 7. Results of the comparison tests where: (a) for the first test, listener's preferences according to each test pair; (b) for the first test, listener's preferences according to each synthesized version; and (c) for the second test, listener's preferences according to each synthesized version.

were speech processing specialists.

Figure 7(a) shows the choices of the listeners according to each test pair, whereas Figure 7(b) presents the overall preference. It can be observed that the lack of information related to syllable and POS degrades the quality of the synthesized speech, with the absence of syllable information being more severely sensed.

4.1.2 INFLUENCE OF SYLLABLE STRESS

For the evaluation of the influence of syllable stress on the speech quality, the original synthesizer was trained with the exclusion of all the questions for context-clustering related to syllable stress.

The subjective test was performed in the same way as the previous test in which the influence of POS and syllable was verified. However, for the present case only two utterance versions, namely, original and no stress information, were compared. The same listeners who participated in the previous test took part in this one.

Figure 7(c) shows the preference of the listeners for this case, where it can be seen that the lack of features related to syllable stress information strongly degrades the quality of the synthesized speech.

4.1.3 DISCUSSION

Comparing the results shown in Figure 7 with the number of nodes of the generated decision-trees presented in Table 5 it is possible to connect the correlation between the influence of POS, syllable and stress on the synthesized speech with the amount of nodes. However, it should be stated that a more precise evaluation should also take into account the fact of which nodes belong to the topper or lower parts of the generated trees.

By observing Figure 7(a) and Figure 7(c), another fact which might arise a discussion is that stress apparently degrades more the synthesized speech than syllable, though stress information corresponds to a sub-set of syllable according to the way in which the experiment was carried out. This difference might have occurred due to the conditions in which the tests were conducted, since syllable information was tested jointly with POS in the first evaluation, whereas stress was separately evaluated in a test where there was only one pair of utterances for each sentence.

4.2 COMPARISON OF THE CURRENT SYNTHESIZER WITH OTHER SYSTEMS

A Mean Opinion Score (MOS) test was conducted among the HMM-based synthesizer and other four systems available for Brazilian Portuguese.

4.2.1 THE SYNTHESIZERS

The synthesizers considered in the subjective test were:

- the Brazilian Portuguese engine of the MBROLA Project [42];
- the optional free TTS engine which can be used in the DOSVOX system, i.e., the Lernout & Hauspie engine [43];
- two commercial systems (Commercial System 1 and Commercial System 2).

All the systems described above are based on the method of unit selection and concatenation. The first synthesizer is part of the MBROLA Project, which intends to provide free speech synthesis engines for several languages. The DOSVOX system corresponds to a free operating system for the visually impaired. It includes its own speech synthesizer and offers the possibility of using other engines. For this test, it was considered the Lernout & Hauspie engine for Brazilian Portuguese, which has also been applied to other multilingual applications based on TTS. Commercial System 1 and Commercial System 2 were chosen to be kept without identification to avoid any sort of implication.

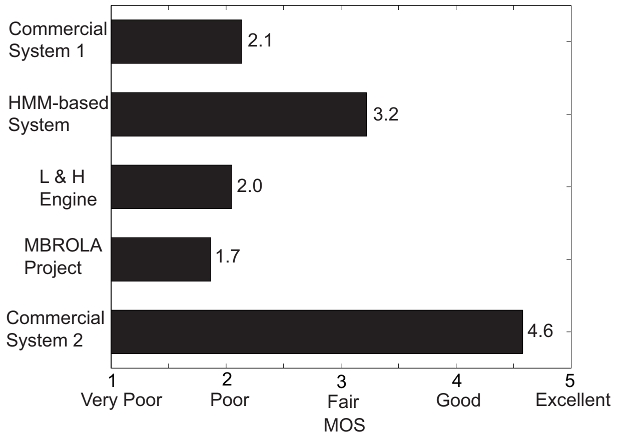

Figure 8. Overall result of the MOS test comparing the synthesizer with four other systems.

4.2.2 THE SENTENCES

For the test, the same twenty sentences selected from newspaper database used in the test of Section 4.1 were used. It should be noted that each utterance information produced by the NLP modules connected to the HMM-based and MBROLA synthesizers was manually corrected in order to avoid transcription and/or stress related errors on the synthesized speech. However, no sort of manual correction was carried out for the other systems due to the impossibility of accessing the intermediate parts of their respective TTS engines. Nevertheless, it was observed that they were able to perform text processing with no apparent errors according to listening tests, therefore ensuring a fair comparison among the synthesizers.

4.2.3 THE SUBJECTS

A total of twenty Brazilian subjects participated in the test. Since the intention was to evaluate the overall quality of the synthesizers from the viewpoint of the general user, the chosen listeners had no training and were not familiarized with the speech processing area.

4.2.4 THE RESULTS

Figure 8 shows the overall MOS obtained for each synthesizer, where it can be seen that except for Commercial System 2 the HMM-based synthesizer outperforms the other systems.

Although the HMM-based system achieved the second score in terms of overall quality, the difference between the referred synthesizer and Commercial System 2 is considerable. The quality of the database seemed to be crucial for the decision of the listeners. Among the evaluated synthesizers, Commercial System 2 was the only engine which presented female voice. Furthermore, the referred system seems to be designed from a speech inventory with considerable size, since in most of the times it was hard to detect discontinuity distortions on the synthesized speech. These sort of artifacts are typical from unit selection and concatenation-based systems derived from limited corpora, although it is possible to reduce their effect through some prosody modification techniques (e.g [44]).

5. CONCLUSION AND FUTURE WORK

The description of a Brazilian Portuguese speech synthesizer with its corresponding characteristics was performed in this paper. The system is based on a technique wherein the speech waveform is generated from parameters directly derived from HMMs. The main advantages of this approach, when compared with the other techniques, corresponds to the possibility of obtaining synthesized speech with good quality and the modification of voice characteristics/styles, even from relatively small speech databases. It was shown that with the proper application of questions for context-clustering and determination of the list of features, the HMMs, through the generated decision-trees, were able to track the characteristics of the Brazilian Portuguese language. In order to investigate the value of the information input to the synthesizer, it was verified according to some subjective tests that POS tags, syllable and stress are important for the quality of the synthesized speech. Also, according to a MOS test performed with listeners not familiarized with the speech processing area, the synthesizer in question performed well when compared to other systems based on the unit selection and concatenation method.

Future work concerns the utilization of larger corpora (at least one hour of speech) and experiments related to speaker adaptation as well as synthesis of voices with different styles and expressions (beyond *read speech*).

6. ACKNOWLEDGEMENT

The authors would like to thank the reviewers for their comments and suggestions, which resulted in a significant improvement of this paper.

REFERENCES

[1] W. B. Kleijn and K. K. Paliwal, Eds., *Speech coding and synthesis*. Amsterdam, The Netherlands: Elsevier, 1995.

[2] A. Black, "Unit selection and emotional speech," in *Proc. of the European Conf. on Speech Communication and Technology (EUROSPEECH)*, 2003.

[3] N. Campbell, "Towards synthesizing expressive speech; designing and collective expressive speech data," in *Proc. of the European Conf. on Speech Communication and Technology (EUROSPEECH)*, 2003.

[4] A. Black and N. Campbell, "Optimising selection of units from speech databases for concatenative synthesis," in *Proc. of the European Conf. on Speech Communication and Technology (EUROSPEECH)*, 1995.

[5] A. Hunt and A. Black, "Unit selection in a concatenative speech synthesis system using a large speech database," in *Proc. of the IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, 1996.

[6] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "Simultaneous modeling of spectrum, pitch and duration in HMM-based speech synthesis," in *Proc. of the Euro-*

pean Conf. on Speech Communication and Technology (EUROSPEECH), 1999.

[7] K. Tokuda, H. Zen, and A. W. Black, "An HMM-based speech synthesis applied to English," in *Proc. of IEEE Workshop in Speech Synthesis*, 2002.

[8] R. Maia, H. Zen, K. Tokuda, T. Kitamura, and F. G. Resende, "Towards the development of a Brazilian Portuguese text-to-speech system based on HMM," in *Proc. of the European Conf. on Speech Communication and Technology (EUROSPEECH)*, 2003.

[9] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "Speaker interpolation in HMM-based speech synthesis," in *Proc. of the European Conf. on Speech Communication and Technology (EUROSPEECH)*, 1997.

[10] K. Shichiri, A. Sawabe, T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "Eigenvoices for HMM-based speech synthesis," in *Proc. of the Int. Conf. on Spoken Language Processing (ICSLP)*, 2002.

[11] J. Yamagishi, M. Tachibana, T. Masuko, and T. Kobayashi, "Speaking style adaptation using context clustering decision tree for HMM-based speech synthesis," in *Proc. of the IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, 2004.

[12] H. Zen and T. Toda, "An overview of Nitech HMM-based speech synthesis for Blizzard Challenge 2005," in *Proc. of the European Conf. on Speech Communication and Technology (EUROSPEECH)*, 2005.

[13] C. L. Bennett, "Large scale evaluation of corpus-based synthesizers: results and lessons from the Blizzard Challenge 2005," in *Proc. of the European Conf. on Speech Communication and Technology (EUROSPEECH)*, 2005.

[14] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "Mixed-excitation for HMM-based speech synthesis," in *Proc. of the European Conf. on Speech Communication and Technology (EUROSPEECH)*, 2001.

[15] H. Kawahara, I. Masuda-Katsuse, and A. de Cheveigné, "Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction: possible role of a repetitive structure in sounds," *Speech Communication*, vol. 27, no. 3-4, pp. 187–207, Apr. 1999.

[16] D. Freitas and D. Braga, "Towards an intonation module for a Portuguese TTS system," in *Proc. of the Int. Conf. on Spoken Language Processing (ICSLP)*, 2002.

[17] H. Fujisaki, S. Narusawa, S. Ohno, and D. Freitas, "Analysis and modeling of F0 contours of Portuguese utterances based on the command-response model," in *Proc. of the European Conf. on Speech Communication and Technology (EUROSPEECH)*, 2003.

[18] P. A. Barbosa, F. Violaro, E. C. Albano, F. Simões, P. Aquino, S. Madureira, and E. Françozo, "Aiuruetê: a high-quality concatenative text-to-speech synthesis system for Brazilian Portuguese with demisyllabic analysis-based units and a hierarchical model of rhythm production," in *Proc. of the European Conf. on Speech Communication and Technology (EUROSPEECH)*, 1999.

[19] S. G. Kafka, F. S. Pacheco, I. C. Seara, S. Klein, and R. Seara, "Utilização de segmentos transicionais homorgânicos em síntese de fala concatenativa," in *Anais do Congresso Brasileiro de Automática (CBA)*, 2002.

[20] R. Maia, H. Zen, K. Tokuda, T. Kitamura, and F. G. Resende, "On the application of HMM-based speech synthesis to Brazilian Portuguese," in *Proc. of Autumn Meeting of the Acoustical Society of Japan*, 2003.

[21] ——, "Influence of part-of-speech tagging, syllabication, and stress on HMM-based Brazilian Portuguese speech synthesis," in *Proc. of Spring Meeting of the Acoustical Society Japan*, 2004.

[22] H. Zen, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "Hidden semi-Markov model based speech synthesis," in *Proc. of the Int. Conf. on Spoken Language Processing (ICSLP)*, 2004.

[23] T. Toda and K. Tokuda, "Speech parameter generation algorithm considering global variance for HMM-based speech synthesis," in *Proc. of the European Conf. on Speech Communication and Technology (EUROSPEECH)*, 2005.

[24] T. Fukada, K. Tokuda, T. Kobayashi, and S. Imai, "An adaptive algorithm for mel-cepstral analysis of speech," in *Proc. of the IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, 1992.

[25] K. Tokuda, N. Miyazaki, and T. Kobayashi, "Multi-space probability distribution HMM," *IEICE Transactions on Information & Systems*, vol. E85-D, no. 3, pp. 455–464, Mar. 2002.

[26] "Speech Assessment Methods Phonetic Alphabet (SAMPA)," http://www.phon.ucl.ac.uk/home/sampa/home.htm.

[27] A. W. Black, P. Taylor, and R. Caley, "The Festival Speech Synthesis System," http://www.festvox.org/festival/.

[28] E. C. Albano and A. A. Moreira, "Archisegment-based letter-to-phone conversion for concatenative speech synthesis in Portuguese," in *Proc. of the Int. Conf. on Spoken Language Processing (ICSLP)*, 1996.

[29] "SPOLTECH: Advancing human language technology in Brazil and the United States through collaborative research on Portuguese spoken language systems," Final Report, July 2001, federal University of Rio Grande do Sul, University of Caxias do Sul, Colorado University, and Oregon Graduate Institute.

[30] F. Barbosa, G. Pinto, F. G. Resende, C. Gonçalves, R. Monserrat, and R. Rosa, "Grapheme-phone transcription algorithm for a Brazilian Portuguese TTS," in *Proc. of 6th. Workshop on Computational Processing of Written and Spoken Portuguese (PROPOR)*, 2003.

[31] F. Barbosa, Ferrari, and F. G. Resende, "A distinção entre homógrafos heterófonos em sistemas de conversão texto-fala," in *Processamento da Linguagem, Cultura e Cognição: estudos de lingüística cognitiva*, 2003.

[32] D. C. Silva, A. de Lima, R. Maia, D. Braga, J. F. de Morais, J. A. de Morais, and F. G. V. Resende Jr., "A rule-based grapheme-phone converter and stress determination for Brazilian Portuguese natural language processing," in *Proc. of IEEE Int. Telecomm. Symposium (ITS)*, 2006.

[33] I. C. Seara, S. G. Kafka, S. Klein, and R. Seara, "Alternância vocálica das formas verbais e nominais do português brasileiro para aplicação em conversão texto-fala," *Revista da Sociedade Brasileira de Telecomunicações*, vol. 17, no. 1, pp. 79–85, June 2002.

[34] R. Seara, Jr., I. C. Seara, S. Kafka, F. S. Pacheco, R. Seara, and S. Klein, "Parâmetros lingüísticos utilizados para a geração automática de prosódia em sistemas de síntese de fala," in *Anais do Simpósio Brasileiro de Telecomunicações (SBrT)*, 2004.

[35] C. Oliveira, L. C. Moutinho, and A. Teixeira, "On European Portuguese automatic syllabification," in *Proc. of the European Conf. on Speech Communication and Technology (EUROSPEECH)*, 2005.

[36] E. Bechara, *Moderna Gramática Portuguesa*. Rio de Janeiro, RJ, Brazil: Lucerna, 2002.

[37] F. Barbosa, R. Maia, and F. G. Resende, "Análise comparativa do impacto da classe gramatical em sistemas TTS baseados em HMMs," in *Anais do Simpósio Brasileiro de Telecomunicações (SBrT)*, 2004.

[38] J. J. Odell, "The use of context in large vocabulary speech recognition," Ph.D. dissertation, Cambridge University, Mar. 1995.

[39] E. Lopes, *Fundamentos da Lingüística Contemporânea*. São Paulo, SP, Brazil: Cultrix, 1975.

[40] A. Alcaim, J. A. Solewicz, and J. A. de Morais, "Freqüência de ocorrência dos fones e listas de frases foneticamente balanceadas para o português falado no Rio de Janeiro," *Revista da Sociedade Brasileira de Telecomunicações*, vol. 7, no. 1, pp. 23–41, Dec. 1992.

[41] R. J. R. Cirigliano, C. Monteiro, F. L. de L. Barbosa, F. G. V. Resende Jr., L. R. Couto, and J. A. de Morais, "Um conjunto de 1000 frases para o português brasileiro obtido utilizando a abordagem de algoritmos genéticos," in *Anais do Simpósio Brasileiro de Telecomunicações (SBrT)*, 2005.

[42] "The MBROLA Project," http://tcts.fpms.ac.be/synthesis/mbrola.html.

[43] "Projeto DOSVOX," http://intervox.nce.ufrj.br/dosvox.

[44] D. O'Brien and A. Monaghan, "Concatenative synthesis based on a harmonic model," *IEEE Transactions on Speech and Audio Processing*, vol. 9, no. 1, pp. 11–20, Jan. 2001.

Ranniery da Silva Maia was born in Natal, Brazil. In 1998 he graduated in Electrical Engineering from the Federal University of Rio Grande do Norte (UFRN), Natal, Brazil. In 2000 he received the M.Sc. degree in Electrical Engineering from the Federal University of Rio de Janeiro (COPPE/UFRJ), Rio de Janeiro, Brazil, and in 2006 the Dr.Eng. degree in Computer Science and Engineering from the Nagoya Institute of Technology (NIT), Nagoya, Japan. He is currently a researcher at the National Institute of Information and Communications Technology (NiCT) and ATR Spoken Language Communication Laboratories (ATR-SLC), Kyoto, Japan. His research concerns speech synthesis, spoken dialog systems, speech coding and natural language processing.

Heiga Zen was born in Osaka, Japan in 1979. He received the A.E. degree in electronic and information engineering from Suzuka National College of Technology, Suzuka, Japan, in 1999, and the B.E., M.E., and Dr.Eng. degrees in computer science, electrical and computer engineering, and computer science and engineering from Nagoya Institute of Technology, Nagoya, Japan, in 2001, 2003, and 2006, respectively. During 2003, he was an intern researcher at ATR Spoken Language Translation Research Laboratories (ATR-SLT), Kyoto, Japan. From June 2004 to May 2005, he was an intern/co-op researcher in the Human Language Technology group at IBM T.J. Watson Research Center, Yorktown Heights, NY. He is currently a postdoctoral fellow at Nagoya Institute of Technology. His research interests include statistical speech recognition and synthesis.

Keiichi Tokuda received the B.E. degree in electrical and electronic engineering from the Nagoya Institute of Technology, Nagoya, Japan, the M.E. and Dr.Eng. degrees in information processing from the Tokyo Institute of Technology, Tokyo, Japan, in 1984, 1986, and 1989, respectively. From 1989 to 1996 he was a Research Associate at the Department of Electronic and Electric Engineering, Tokyo Institute of Technology. From 1996 to 2004 he was a Associate Professor at the Department of Computer Science, Nagoya Institute of Technology as Associate Professor, and now he is a Professor at the same institute. He is also an Invited Researcher at ATR Spoken Language Translation Research Laboratories, Japan and was a Visiting Researcher at Carnegie Mellon University from 2001 to 2002. He is a co-recipient of the Paper Award and the Inose Award both from the IEICE in 2001, and the TELECOM System Technology Prize from the Telecommunications Advancement Foundation Award, Japan, in 2001. He was a member of the Speech Technical Committee of the IEEE Signal Processing Society. His research interests include speech speech coding, speech synthesis and recognition, and statistical machine learning.

Tadashi Kitamura received the B.E. degree in electronics engineering from Nagoya Institute of Technology, Nagoya, in 1973, and M.E. and Dr.Eng. degrees from Tokyo Institute of Technology, Tokyo, in 1975 and 1978, respectively. In 1978 joined the Research Laboratory of Precision Machinery and Electronics, Tokyo Institute of Technology, as a Research Associate. In 1983 he joined Nagoya Institute of Technology, as a Assistant Professor. He is currently a Professor of Graduate School of Engineering of Nagoya Institute of Technology. His current research interests include speech processing, image processing and sign language recognition. He is a member of IEEE, ISCA, IEICE, IPSJ, ASJ and ITE.

Fernando Gil Vianna Resende Junior received the B.Sc. degree from Military Institute of Engineering (IME), Brazil, in 1990, and the M.Sc. and Ph.D. degrees from Tokyo Institute of Technology (TIT), Japan, in 1994 and 1997, respectively, all in electrical engineering. Since 1998 he has been with the Department of Electronic Engineering and Computer Science, Polytechnic School, Federal University of Rio de Janeiro (UFRJ), as Associate Professor. Also, since 2003 he has been with the Program of Electrical Engineering, COPPE/UFRJ. His research interests are in the areas of natural language processing, speech synthesis, speech and speaker recognition, and speech coding.