Towards the Development of a Brazilian Portuguese Text-to-Speech System Based on HMM

R. da S. Maia[1], *H. Zen*[1], *K. Tokuda*[2], *T. Kitamura*[2], *F. G. V. Resende Jr.*[3]

[1]Dept. of Electrical and Computer Engineering, Nagoya Institute of Technology, Nagoya, Japan
[2]Dept. of Computer Science, Nagoya Institute of Technology, Nagoya, Japan
[3]Dept. of Electronics and Computer Science/EPoli, Federal University of Rio de Janeiro, Rio de Janeiro, Brazil
{maia,zen,tokuda,kitamura}@ics.nitech.ac.jp, gil@lps.ufrj.br

Abstract

This paper describes the development of a Brazilian Portuguese text-to-speech system which applies a technique wherein speech is directly synthesized from hidden Markov models. In order to build the synthesizer a speech database was recorded and phonetically segmented. Furthermore, contextual informations about syllables, words, phrases, and utterances were determined, as well as questions for decision tree-based context clustering algorithms. The resulting system presents a fair reproduction of the prosody even when a small database is used for training.

1. Introduction

Text-to-speech synthesis (TTS) is an area which stimulates great interest for speech processing researchers nowadays. Although most of TTS schemes are based on the selection and concatenation of speech waveforms, to change voice characteristics for this technique a large amount of speech database is required. On the other hand, TTS systems where speech can be generated directly from hidden Markov models (HMM) [1] enable the alteration of voice characteristics without the necessity of large databases by applying, e.g., a speaker interpolation technique [2] or eigenvoices [3]. Although having been originally implemented for Japanese language, the HMM-based speech synthesis (HSS) approach has also been applied to other languages, e.g., English [4], because input contextual labels and questions for context clustering are the only language dependent topics in the HSS scheme.

In order to put HMM-based Brazilian Portuguese TTS to work it was necessary to record a speech database which was eventually segmented and annotated with contextual informations about syllable, word, phrase, and utterance that could possibly have influence in the proposed TTS. The implemented system produces synthetic speech with a fair reproduction of the prosody even when a small amount of 80 utterances is used for HMM training.

This paper is organized as follows: Section 2 briefly describes the HSS technique; Section 3 concerns to the development of the Brazilian Portuguese TTS; Section 4 discusses the experiments which were conducted; and Section 5 presents the conclusion.

2. HMM-based speech synthesis technique

The HMM-based speech synthesis technique comprises training and synthesis parts, as depicted in Figure 1.

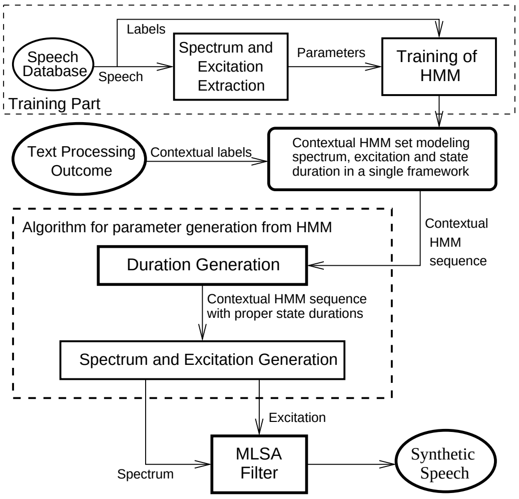

Labels
Speech Database
Speech
Spectrum and Excitation Extraction
Parameters
Training of HMM
Training Part
Text Processing Outcome
Contextual labels
Contextual HMM set modeling spectrum, excitation and state duration in a single framework
Contextual HMM sequence
Algorithm for parameter generation from HMM
Duration Generation
Contextual HMM sequence with proper state durations
Spectrum and Excitation Generation
Excitation
Spectrum
MLSA Filter
Synthetic Speech

Figure 1: *Diagram of the HMM-based speech synthesis technique.*

2.1. Training part

In the training part, spectrum and excitation parameters are extracted from the speech database, and HMM training is carried out. In HSS, each HMM corresponds to a left-to-right no-skip model where each output vector is composed of two streams: spectrum part, represented by mel-cepstral coefficients and their related delta and delta-delta coefficients; and the excitation part, represented by excitation parameters and the corresponding dynamic features delta and delta-delta. In order to reproduce temporal structure of speech, the HMM include density distributions for state durations. Thus, spectrum, excitation, and state duration are jointly modeled in a single HMM framework.

2.1.1. *Spectrum modeling*

In this work spectrum is modeled by mel-cepstral coefficients which can directly synthesize speech using the MLSA (mel log spectrum approximation) filter [5]. The mel-cepstral coefficients are extracted from the speech database through a 24^{th} order mel-cepstral analysis using 25-ms Blackman windows with 5-ms shifts. Output probability for the mel-cepstral coefficients corresponds to multivariate Gaussian distributions.

2.1.2. *Excitation modeling*

The excitation parameters in this work are composed of logarithm of the fundamental frequency F_0, $\log(F_0)$, and its corresponding delta and delta-delta coefficients. F_0 extraction from the speech database is carried out at every 5 ms using the ESPS (Entropic Signal Processing System) package [6]. Since the observation sequence of F_0 (or $\log(F_0)$) is usually composed of one-dimensional continuous values and a discrete symbol which represents the unvoiced regions of the F_0 contour, continuous multivariate probability distributions are not able to model F_0 sequences properly. In order to solve this problem, multi-space probability distribution is applied for F_0 pattern modeling [7].

2.1.3. Duration modeling

The state durations of each HMM are modeled by a multivariate Gaussian distribution. The dimensionality of the density for each HMM is equal to the number of states in the corresponding HMM, where the n^{th} dimension of state duration density corresponds to the n^{th} state of the respective HMM since left-to-right no-skip models are employed.

2.1.4. Decision tree-based context clustering

During HSS training, the speech database may not include enough samples of contextual models to permit a good HMM parameter estimation. In addition, as for the synthesis, sometimes a contextual label to be synthesized does not have a corresponding HMM in the trained model set. In order to solve these two problems, a decision tree-based context clustering technique is applied to the distributions of spectrum, F_0 and state duration in the same manner as HMM-based speech recognition. Furthermore, since spectrum, F_0 and state duration have their own influential contextual informations, they are clustered independently [1].

2.2. Synthesis part

The synthesis part of HSS is processed as follows. First, a given text to be synthesized is converted into a contextual label sequence. Then, according to such label sequence, an HMM sequence is constructed by concatenating context-dependent HMM. After this, state durations for the HMM sequence are determined so that the output probability of the state durations are maximized. From the HMM sequence with the proper state durations included, a sequence of mel-cepstral coefficients and $\log(F_0)$ values, including the voiced/unvoiced decisions, are generated using Case 2 of the algorithm presented in [8]. Finally, speech waveform is synthesized directly from the generated mel-cepstral coefficients and F_0 values by using the MLSA filter.

3. Developing Brazilian Portuguese TTS

3.1. Database recording, labeling and segmentation

A database composed of 613 sentences, including 200 phonetically balanced sentences for Brazilian Portuguese spoken in Rio de Janeiro [9], was recorded from a male Brazilian Portuguese native speaker at a sampling rate of 48 kHz, with 16 bits per sample. The database was posteriorly down sampled to 16 kHz.

The phonetic labeling procedure for the database was performed in two steps: (1) text-to-phoneme transcription through a phonetic transcriber software [10]; (2) label correction of the transcriptions obtained in the previous step. The latter was performed due to the differences between the phoneme set employed in [10] and the one employed in this work. In the present case the phoneme set consists in 43 phonemes, including one silence and one pause model. Despite the absence of diphthongs in some known Brazilian Portuguese phoneme sets employed for TTS purposes [11, 12], in the present case some diphthongs were considered because of the usual hard task of separating the vowels from semi-vowels during the phonetic segmentation. Table 1 shows the phoneme set which was employed to label the database, excluding silence and pause models.

Table 1: *Phoneme set employed to label the database.*

Vowels	a, E, e, i, O, o, u, a˜, e˜, i˜, o˜, u˜
Semi-vowels	j, w
Diphthongs	aj, aw, ej, ew, oj, ow, wa
Consonants	f, s, S, z, v, Z, b, d, t, k, g, p, l, L, R r, x, m, n, J

Time label boundaries (phonetic segmentation) for the database were created using HMM Toolkit (HTK) [13]. First, a flat-start training was applied to the database since time label boundaries are not necessary. Using the flat-start trained HMM set, a time alignment was performed through the Viterbi algorithm determining, thus, rough time label boundaries for the database. At first, 50 sentences were manually corrected. These 50 sentences were used to re-train the database using the bootstrap method. From the new re-trained models more sentences were corrected. This method of correcting and posteriorly re-training was employed instead of correcting all the sentences from the flat-start training because at each application of re-training using bootstrap the posterior alignment procedure tends to improve the time label boundaries, and thus decreasing the hard work of the manual correction. A total of 80 sentences were segmented.

3.2. Contextual information

One of the language dependent issues of HSS corresponds to the input contextual labels which are used to determine the corresponding HMM in the models set. Thus, contextual informations which are fully represented in such contextual labels were necessary to be considered in order to obtain a good reproduction of the prosody. Whenever a given contextual label does not have a corresponding HMM in the models set, decision tree-based context clustering is applied to generate the respective unseen model. In this work, the following contextual informations were considered:

- phoneme:
  - {preceding, current, succeeding} phonemes;
  - position of current phoneme in current syllable;
- syllable:
  - whether or not {preceding, current, succeeding} syllables are stressed;
  - number of phonemes in {preceding, current, succeeding} syllables;
  - position of current syllable in current word;
  - number of stressed syllables in current phrase {before, after} current syllable;
  - number of syllables, counting from previous stressed to current syllable in the utterance;
  - number of syllables, counting from current to next stressed syllable in the utterance;
- word:
  - part-of-speech of {preceding, current, succeeding} words;

  - number of syllables in {preceding, current, succeeding} words;
  - position of current word in current phrase;
  - number of content words in current phrase {before, after} current word;
  - number of words counting from previous content word to current word in the utterance;
  - number of words counting from current to next content word in the utterance;
  - interrogative flag for the word;

- phrase:
  - number of {syllables, words} in {preceding, current, succeeding} phrases;
  - position of current phrase in current utterance;

- utterance:
  - number of {syllables, words, phrases} in the utterance;

3.3. Utterance information

Text processing for the present TTS is composed of two parts. In the first part a given text to be synthesized is converted into an *utterance information* which concerns phoneme, syllable and words, namely:

- phoneme level:
  - phoneme symbol;
- syllable level:
  - syllable in which the current phoneme is inserted;
  - stress indication for the syllable (0 → not stressed, 1 → stressed);
- word level:
  - word in which the current phoneme is inserted;
  - part-of-speech of the word, as shown in Table 2;
  - indication if the word has the typical intonation of a word in the end of an interrogative phrase (0 or nothing → no, 1 → yes).

During the text processing, the pause model "lp" indicates phrase separation, including period (.), question mark (?), exclamation mark (!), comma (,), colon (:), semicolon (;), and dash (-), whereas the silence model "sil" indicates the beginning and the end of the text to be synthesized. As an example, Table 3 shows the utterance information for ”Leila tem um lindo jardim”, which means ”Leila has a beautiful garden”.

Utterance informations for the training database were manually included because this first part of text processing is still in development. The starting point for the utterance information inclusion was the database labeling outcome which preceded the segmentation. Despite the hard task, manual inclusion usually tends to achieve better results because even the best text processor is not able to track the differences which eventually might exist between what is prompted by the sentences and what is really recorded, such as concatenation of words, recording mistakes, etc.

In the second part of the text processing procedure, which is not as hard as the first part, utterance information like the one showed in Table 3 is converted into contextual labels which include all the informations listed in Section 3.2.

Table 2: *Parts-of-speech (word classes) and their respective symbols*

Class	Symbol	Class	Symbol
Verb	ver	Noun	subs
Adjective	adj	Adverb	adv
Article	art	Pronoun	pro
Preposition	prep	Numeral	num
Conjunction	conj	Interjection	int
Contraction: preposition + article			part

Table 3: *Utterance information for ”Leila tem um lindo jardim”, which means ”Leila has a beautiful garden”.*

phoneme	syll	stress	word	class	intr
sil					
l	lej	1	lejla	subs	
ej					
l	la	0			
a					
t	te˜j	0	te˜j	ver	
e˜					
j					
u˜	u˜	0	u˜	art	
l	li˜	1	li˜du	adj	
i˜					
d	du	0			
u					
Z	Zax	0	Zaxdi˜	subs	
a					
x					
d	di˜	1			
i˜					
sil					

3.4. Context clustering

In order to carry out decision tree-based context clustering, some questions were determined to cluster the phonemes. Afterwards these questions were extended to include all the contextual informations, i.e., syllable, word, phrase and utterance. The questions were derived according to phonetic characteristics of vowels, semi-vowels, diphthongs, and consonants [14]. The following high and low level classifications for the phonemes corresponded to some questions applied to generate the decision trees:

- silence/pause;
- vowel → anterior, central, posterior, open, closed, oral, nasal;
- semi-vowel;
- diphthong → ascendant, descendant;
- consonant → stop, fricative, liquid, vibrant, bilabial, labiodental, dental, alveolar, palatal, velar, unvoiced, voiced, oral, nasal.

4. Experiments

The training of the system was conducted using the HMM-based speech synthesis toolkit (HTS) [15] applied to 80 utterances phonetically segmented and annotated with contextual informations. Synthetic speech, even for such small database has a fair reproduction of the prosody. Some samples can be verified at

`http://kt-lab.ics.nitech.ac.jp/˜maia/demo`

Figure 2 shows a comparison between one F_0 pattern generated from HMM for a given sentence which is not included in the

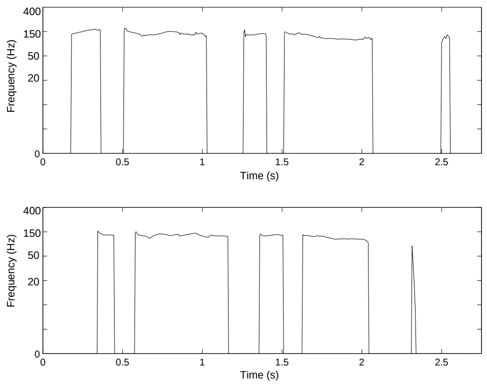


Figure 2: *Examples of F_0 patterns: (a) extracted from the uttered sentence (natural); (b) generated from HMM.*

training database, and the F_0 pattern extracted from the same sentence, uttered by the speaker who recorded the database. It can be noticed that the generated F_0 contour is close to the natural pattern. Figure 3 shows examples of decision trees constructed for spectrum, F_0, and state durations for a given contextual HMM. It can be seen from the decision-trees that phoneme contexts are more important for spectrum whereas informations about syllable, word and phrase mainly affect F_0 and state durations.

5. Conclusion

This paper presents the current development of a Brazilian Portuguese text-to-speech system based on HMM. Database recording and labeling were necessary to be carried out in order to build the synthesizer. Further, contextual informations and questions for decision tree-based context clustering were determined. The system achieves a fair reproduction of the prosody even when the small amount of 80 sentences is used for HMM training. Future work for the present TTS includes the conclusion of the first part of text processing, segmentation of more utterances to increase the training database, and improvement on the contextual informations.

6. Acknowledgments

The authors would like to thank Filipe L. de F. Barbosa, from Federal University of Rio de Janeiro, for his contribution.

7. References

[1] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "Simultaneous modeling of spectrum, pitch and duration in HMM-based speech synthesis," in *EUROSPEECH*, 1999.

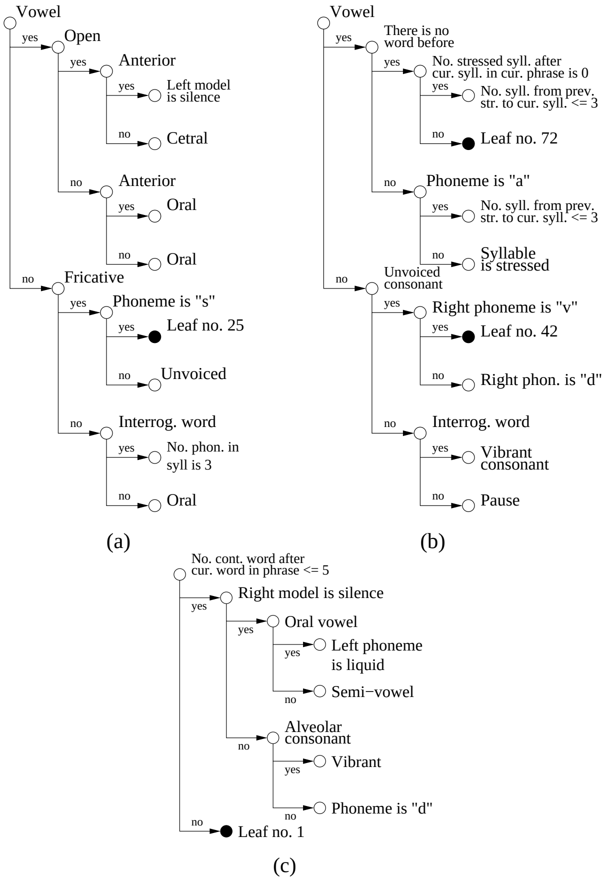
[2] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "Speaker interpolation in HMM-based speech synthesis," in *EUROSPEECH*, 1997.

[3] K. Shichiri, A. Sawabe, T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "Eigenvoices for HMM-based speech synthesis," in *ICSLP*, 2002.

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

[5] T. Fukada, K. Tokuda, T. Kobayashi, and S. Imai, "An adaptive algorithm for mel-cepstral analysis of speech," in *ICASSP*, 1992.

[6] `http://www.entropic.com/esps.html`.



(a)

(b)

(c)

Figure 3: *Example of decision-trees constructed: (a) for spectrum (3th state); (b) for F_0 (3th state); and (c) for state duration.*

[7] K. Tokuda, N. Miyazaki, and T. Kobayashi, "Multi-space probability distribution HMM," *IEICE Trans. Inf. & Syst.*, vol. E85-D, Mar. 2002.

[8] K. Tokuda, T. Yoshimura, T. Masuko, T. Kobayashi, and T. Kitamura, "Speech synthesis generation algorithms for HMM-based speech synthesis," in *ICASSP*, 2000.

[9] A. Alcaim, J. A. Solemicz, and J. A. de Morais, "Freqüêcia 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, Dec. 1992.

[10] G. O. Pinto, F. L. F. Barbosa, and F. G. V. Resende Jr., "A Brazilian Portuguese TTS based on HMMs," in *Int. Telecomm. Symposium (ITS)*, 2002.

[11] E. C. Albano and A. A. Moreira, "Archisegment-based letter-to-phone conversion for concatenative speech synthesis in Portuguese," in *ICSLP*, 1996.

[12] "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.

[13] `http://htk.eng.cam.ac.uk`.

[14] L. A. Sacconi, *Nossa Gramática: teoria e prática*. São Paulo, SP, Brazil: Atual, 1994.

[15] `http://hts.ics.nitech.ac.jp`.