An HMM-based phonetic vocoder using mixed excitation

R. da S. MAIA[1], R. J. da R. CIRIGLIANO[2], D. ROJTENBERG[2], F. G. V. RESENDE Jr.[2]

PEE/COPPE[1], DEL/EE[2]
Universidade Federal do Rio de Janeiro
PO Box 68504, Rio de Janeiro, RJ, 21945-970
BRAZIL
{maia, rjcirig, rojtenberg, gil}@lps.ufrj.br

Abstract: - This work describes a phonetic vocoder which uses HMM-based speech recognition/synthesis techniques. Mixed excitation is applied during the synthesis procedure in order to solve the problem of having unnatural synthetic speech when the binary excitation pulse train/random noise is applied. Experiments have shown that a good degree of intelligibility is obtained for an average bit rate of 780 bps whereas the mixed excitation approach efficiently produces more natural speech.

Key-words: - Speech coding, speech recognition, speech synthesis, HMM, MELP.

1 Introduction

Nowadays, speech coding techniques which can efficiently represent digital speech using bit rates under 2 kbps are important for many applications, including transmission and storage. Some coders have been reported to reach good quality around these bit rates [1, 2]. However, when the objective is to obtain lower bit rates (under 1.0 kbps), usually specific techniques that depend on the language are applied [3, 4]. Among these techniques, the phonetic vocoders are those which usually segment the speech signal into a sequence of speech models (like monophones or triphones) using a recognition technique, transmitting such speech models to the decoder jointly with excitation information. The synthesis of the signal is made by concatenating the speech models according to the information transmitted.

This paper presents a phonetic vocoder for the Brazilian Portuguese language using speech recognition on the analysis part, and speech synthesis from Hidden Markov Models (HMM) on the synthesis. In order to produce more natural synthetic speech, mixed excitation based on the Federal Standard Mixed Excitation Linear Prediction (MELP) speech coder [5] is applied on the synthesis part instead of the traditional excitation where pulse train is applied for voiced segments and random noise for unvoiced segments. Vocoders which synthesize speech from HMM have been reported in the literature [4]. However, the present vocoder applies mixed excitation in the synthesis part, and a different quantization for the speech models and state durations during the analysis, since these quantizations are supposed to be applied in a Brazilian Portuguese database.

This work is organized as follows: in Section 2 the phonetic vocoder is described; Section 3 gives information concerning the HMM models training; Section 4 determines the average bit rate of the vocoder; Section 5 describes the experiments performed; and finally Section 6 shows the conclusions.

2 Vocoder description

Figure 1 shows the block diagram of the phonetic vocoder. In sub-sections 2.1 and 2.2 the encoding and decoding processes are respectively described.

2.1 Encoder

In the encoder part two procedures are carried out separately: speech recognition and excitation analysis, as it can be noticed in Figure 1(a).

2.1.1 Speech recognition

Fist, an M^{th} order mel-cepstral analysis is performed on the original speech at every 5 ms, using 25 ms Hamming windows centered on the corresponding frames, wherein mel-cepstral coefficients $\{c_0, \ldots, c_M\}$ which have the property of representing speech spectrum envelope are extracted [6]. After that, speech recognition is performed where the observation sequence $\mathbf{O} = [\vec{o}_1, \ldots, \vec{o}_N]$ has each one of its observation vectors given by $\vec{o}_i = [\vec{c}_i^T \quad \Delta\vec{c}_i^T \quad \Delta^2\vec{c}_i^T]$, with $\vec{c}_i$, $\Delta\vec{c}_i$ and $\Delta^2\vec{c}_i$ being the mel-cepstral coefficients vector and their related dynamic features delta and delta-delta, respectively, for the frame i - the superscript T indicates transposition. These last two

vectors are obtained from the former through

$$\Delta\vec{c}_i = \frac{1}{2}(\vec{c}_{i-1} + \vec{c}_{i+1}), \quad (1)$$

$$\Delta^2\vec{c}_i = \frac{1}{4}(\vec{c}_{i-2} + \vec{c}_{i+2}) - \frac{1}{2}\vec{c}_i. \quad (2)$$

For the present case, $M = 12$, so that each observation vector $\vec{o}_i$ contains 39 elements.

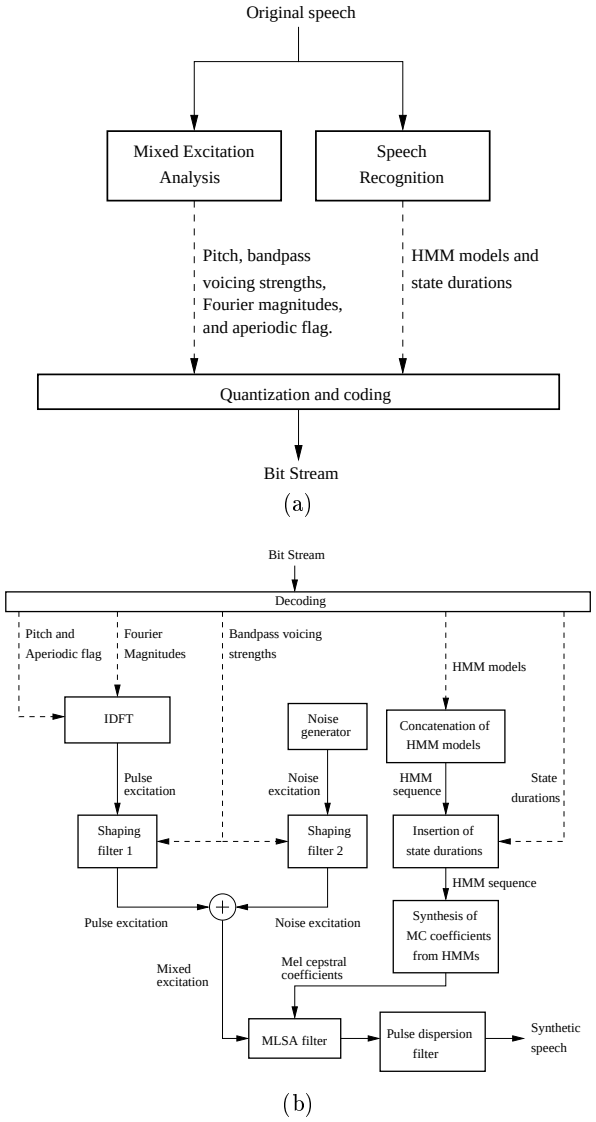

Figure 1: Block diagram of the phonetic vocoder: (a) encoder; (b) decoder.

2.1.2 Excitation analysis

In order to extract excitation parameters, the original speech is divided into 22.5 ms frames. After that, speech analysis is performed in a similar way to the MELP algorithm [5]. However, for the present case the goal is to determine:

- pitch period;
- bandpass voicing strengths;
- Fourier magnitudes;
- jitter.

These informations are quantized in a frame-by-frame basis. The details about how to determine the above parameters can be found in [5].

2.1.3 Quantization and coding

As for the quantization of the speech models, since there are 50 monophones that can compose the triphones, each model is quantized with 6 bits, and the related indices represent the model information.

The state durations for each model are regarded as 3-dimensional vectors (each model is represented by a 3-state HMM) and vector quantization (VQ) is applied. Codebooks were designed with 128, 64, 32, 16, 8 and 4 entries according to the triphone model, and training was conducted by the LBG algorithm. During the search procedure for the state durations, the best state duration vector $\hat{\vec{k}}_o$ is the one which minimizes the distortion

$$D_k = \sqrt{\sum_{i=1}^{3} k_i - \hat{k}_i}, \quad (3)$$

under the restriction

$$|\sum_{i=1}^{3}(k_i - \hat{k}_i)| \leq 1. \quad (4)$$

This restriction produces a seeking-alignment for the models, avoiding loss of model parts caused by an inaccurate state duration.

The excitation parameters listed in Section 2.1.2 are quantized using the same approach employed by the MELP algorithm.

2.2 Decoder

The decoder receives excitation information, model indices and state durations indices from the encoder.

2.2.1 Mixed excitation generation

The excitation is formed like in the MELP coding case, except for the adaptive spectral enhancement and gain adjustment. In the present case, different from the original MELP, there is no adaptive enhancement filter since this filter is implemented using linear prediction coefficients. Also, the gain adjustment is not applied. Figure 1(b) details how the mixed excitation is built.

2.2.2 Mel-cepstral coefficients extraction

First, the information of model indices are used to concatenate a sequence of HMMs. After that, the state durations for each model of the HMM sentence are inserted. Having the HMM sequence with the proper state durations inserted, mel-cepstral coefficients are extracted from this sentence using the algorithm described in [7].

2.2.3 Speech synthesis

Speech is synthesized passing the mixed excitation through the Mel Log Spectrum Approximation (MLSA) filter [6], using the extracted mel-cepstral coefficients. The transfer function of this filter is given by

$$D(z) = \exp \sum_{m=0}^{M} b_m \Phi_m(z), \tag{5}$$

where

$$\Phi_m(z) = \begin{cases} 1, & m = 0, \\ \frac{(1-\alpha^2)z^{-1}}{1-\alpha z^{-1}} \tilde{z}^{-(m-1)} & m \geq 1, \end{cases} \tag{6}$$

with z^{-1} being an all-pass transfer function defined by

$$\tilde{z}^{-1} = \frac{z^{-1} - \alpha}{1 - \alpha z^{-1}}, \quad |\alpha| < 1, \tag{7}$$

and α the corresponding all-pass constant. The coefficients $\{b_0, \ldots, b_M\}$ are obtained from mel-cepstral coefficients $\{c_0, \ldots, c_M\}$ through

$$b_m = \begin{cases} c_M, & m = M \\ c_m - \alpha b_{m+1}, & 0 \leq m < M. \end{cases} \tag{8}$$

The MLSA filter is implemented in as shown in [6], where the α parameter is set to 0.31 for the sampling frequency of 8 kHz.

3 Bit rate

According to experiments, the triphone models can be extracted from the speech signal at an average rate of 12.8 models/s. Six bits are used to quantize each triphone as mentioned in Section 2.1.3. The durations for the 3 states of each triphone are regarded as 3-dimensional vectors, and vector quantization is applied with 2, 3, 4, 5, and 6 bit codebooks, according to the triphone. Experiments have shown that in average 5.6 bits/model are necessary to quantize the state durations properly. Table 1 shows the bit allocation and bit rate for the models and state durations.

Table 1: Bit allocation and average bit rate for the triphones and state durations.

Parameter	Bits/model	Model/s	Bits/s
Triphones	6	12.8	76.8
State durations	5.6	12.8	71.68
Total	148.48 bits/s		

As explained before, the excitation parameters are quantized using the same quantization method employed by the MELP algorithm. The difference lies on the fact that the protection and error correction bits are not transmitted. Due this fact, for unvoiced frames only the bits corresponding to pitch and overall voicing decision are transmitted. According to experiments, 20% of the frames are unvoiced so that the number of bits for the voiced and unvoiced cases are weighted, respectively, by 0.8 and 0.2 for sake of average bit rate computation. Table 2 shows the bit allocation and corresponding bit rate for the excitation parameters.

Table 2: Bit allocation and bit rate for the excitation parameters.

Parameter	Bits/frame		Bits/s
	voiced	unvoiced	
Pitch and overall voicing	7	7	311.11
Fourier magnitudes	8	-	284.44
Aperiodic flag	1	-	35.56
Total	16	7	631.11

The total average bit rate (BR) for the coder is, consequently,

$$\begin{aligned} BR &= \underset{\text{(Models, state durations)}}{\text{Spec. parameters}} + \underset{\text{(Pitch,BPVS,FM,AF)}}{\text{Excitation}} \\ &= 148.48 + 631.11 \approx 780 \text{ bps}. \end{aligned}$$

4 HMM models training

Two databases were used to train HMM models. The first, composed by 313 sentences spoken by a male speaker in the Brazilian Portuguese language, sampled at 8 kHz, was used on the A/B decision test. The second database, composed by 200 phonetically balanced sentences spoken by the same speaker of the first database, sampled at 8 kHz, was used on the intelligibility test. Speech signals were windowed by 25 ms Hamming windows with 5 ms shifts, and mel-cepstral coefficients were obtained through the mel-cepstral analysis technique shown in [6]. The feature vectors consisted of 13 mel-cepstral coefficients, including the 0^{th}, and their

corresponding delta and delta-delta coefficients calculated as shown in (1) and (2).

Each model corresponded to triphones modeled by 3-state left-to-right HMMs. The output probability of the states were modeled by single Gaussian distributions with diagonal covariance. A total of 3218 triphones were modeled from the first database and 1783 from the second.

5 Experiments

Two different subjective tests were performed in order to evaluate the intelligibility of the phonetic vocoder, and the efficiency of the mixed excitation.

In the first test, six sentences from the second training database were analyzed concerning the intelligibility, where eight listeners gave their opinion. Table 5 shows the results for this test. As it can be seen, the HMM-based phonetic vocoder reaches a good degree of intelligibility.

Table 3: Results for the test where 8 listeners gave their opinions about the degree of intelligibility for 6 sentences synthesized by the phonetic vocoder.

Sentence	Intelligibility (%)
1	96.4
2	96.4
3	100
4	92.5
5	85.4
6	100
Mean	95

In the second test, an A/B decision was performed with nine listeners. Six sentences from the first training database were synthesized using both mixed excitation and binary excitation pulse train/random noise. For the binary case, pitch period was extracted at every 5 ms from the original speech signal using the autocorrelation method with 30 ms Hamming windows centered on each 5 ms frame. Pitch quantization at every 5 ms using 7 bits (128 levels = from 20 to 146, 0 for unvoiced) would cause a bit rate of 7 bits/5 ms = 1.4 kbps, therefore increasing the bit rate when compared with the mixed excitation (632 bps). Table 5 shows the results for this test. It can be seen that mixed excitation was preferred in almost all the cases.

6 Conclusion

In this work, an HMM-based phonetic vocoder for the Brazilian Portuguese language which uses mixed excitation during the synthesis process was presented.

Table 4: Results for the test where 9 listeners gave their preferences between sentences synthesized by mixed and traditional pulse train/random noise excitations.

Sentence	Mix. (%)	Similar (%)	Trad. (%)
1	33	22	45
2	45	33	22
3	45	45	10
4	90	10	-
5	68	10	22
6	45	33	22
Mean	54	26	20

The encoder carries out HMM-based speech recognition using triphones as 3-state left-to-right models. Triphones and state durations indices are sent to the decoder, jointly with the excitation parameters to compose the mixed excitation from the MELP algorithm. In the decoder the phonemes and states durations are used to concatenate a sequence of HMMs, from which mel-cepstral coefficients that can represent speech spectrum envelope are extracted. Finally, mixed excitation is used to produce speech in the output of the MLSA filter, using the mel-cepstral coefficients extracted from the models. Experiments have shown that the vocoder at 780 bps reaches good degree of intelligibility, and the use of mixed excitation improves the quality when compared with the traditional excitation pulse train/random noise.

References

[1] A. McCree and J. C. De Martin, "A 1.7 kb/s MELP coder with improved analysis and quantization," in *Proc. ICASSP*, 1998.

[2] T. Wang, K. Koishida, V. Cuperman, A. Gersho, and J. S. Collura, "A 1200 bps speech coder based on MELP," in *Proc. ICASSP*, 2000.

[3] C. Ribeiro and I. Trancoso, "Phonetic vocoding with speaker adaptation," in *Proc. EUROSPEECH*, pp. 1291–1294, 1997.

[4] T. Masuko, K. Tokuda, and T. Kobayashi, "A very low bit rate speech coder using HMM with speaker adaptation," in *Proc. ICSLP*, 1998.

[5] McCree, K. Truong, E. George, T. Barnwell, and V. Viswanathan, "A 2.4 kbits/s MELP coder candidate for the new U.S. Federal Standard," in *Proc. ICASSP*, pp. 200–203, 1996.

[6] T. Fukada, K. Tokuda, T. Kobayashi, and S. Imai, "An adaptive algorithm for mel-cepstral analysis of speech," in *Proc. ICASSP*, pp. 137–140, 1992.

[7] K. Tokuda, T. Masuko, T. Yamada, T. Kobayashi, and S. Imai, "An algorithm for speech parameter generation from continuous mixture HMM with dynamic features," in *Proc. EUROSPEECH*, 1995.