MIXED-EXCITED PHONETIC VOCODING AT 265 BPS

R. da S. Maia[1]*, R. J. da R. Cirigliano*[2]*, D. Rojtenberg*[2]*, F. G. V. Resende Jr.*[2]

[1]Electrical Engineering Program/COPPE, [2]Dept. of Electronics and Computer Science/EPoli
Federal University of Rio de Janeiro
PO. Box 68504, 21945-970, Rio de Janeiro, RJ, Brazil

ABSTRACT

In this paper a phonetic vocoder which synthesizes speech using mixed excitation is presented. The encoder carries out HMM-based speech recognition and pitch analysis, whereas the decoder performs parameter extraction from HMM and builds a mixed excitation using pitch and bandpass voicing strengths. The vocoder at an average bit rate of 265 bps reaches good degree of intelligibility, while the use of mixed excitation significantly improves the speech quality with no increase of bit rate when compared with the conventional binary excitation pulse train/random noise.

1. INTRODUCTION

Speech coding techniques which can efficiently represent digital speech using bit rates under 2 kbps are important for many applications, e.g. transmission and storage. Although some coders have been reported to reach good performance around these bit rates [1], when the goal is to work at lower bit rates, namely under 1.0 kbps, usually specific techniques that depend on the language are applied. Among these techniques, the phonetic vocoders [2, 3] are those which usually segment the speech signal into a sequence of speech models (like monophones) using a recognition technique, transmitting such speech models to the decoder jointly with prosodic information. The decoding process is usually made by concatenating these models to compose the spectral parameters, whereby jointly with prosodic information produce synthetic speech.

This work presents a phonetic vocoder which uses speech recognition on the encoder part and parameter synthesis from Hidden Markov Models (HMM) in the decoder [3]. In order to produce more natural synthetic speech, mixed excitation based on the Federal Standard Mixed Excitation Linear Prediction (MELP) speech coder [4] is applied instead of the traditional excitation wherein pulse train is applied for voiced segments and random noise for unvoiced segments. We have already proposed a phonetic vocoder which uses mixed excitation [5], where the speech quality was improved when compared with the binary excitation at the cost of a significant increase of bit rate. In the present work, mixed excitation is applied with no bit rate increment by modeling bandpass voicing strengths coefficients (BPVC) from MELP jointly with mel-cepstral coefficients in a single HMM framework. Experiments have shown that speech quality is still significantly improved with the use of this approach.

This work is organized as follows: in Section 2 a description of the current phonetic vocoder is presented; Section 3 concerns to the bandpass voicing strengths modeling by HMM; in Section 4, the performance evaluation of the proposed vocoder is considered; and the conclusions are in Section 5.

2. VOCODER DESCRIPTION

2.1. Encoder

The encoder inputs digital speech sampled at 8 kHz and outputs monophone indices, state durations and pitch through two main procedures: speech recognition and pitch analysis, as shown in Figure 1(a).

2.1.1. Speech recognition

Speech recognition is conducted by an HMM continuous speech recognizer, where each 3-state no-skip HMM represents one monophone with its respective left and right contexts (triphone). The output probabilities are modeled by single Gaussian distributions with diagonal covariances. The feature vectors comprise mel-cepstral coefficients which can represent speech spectrum [6], and their related delta and delta-delta coefficients. These last two parameters are computed from the former through

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

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

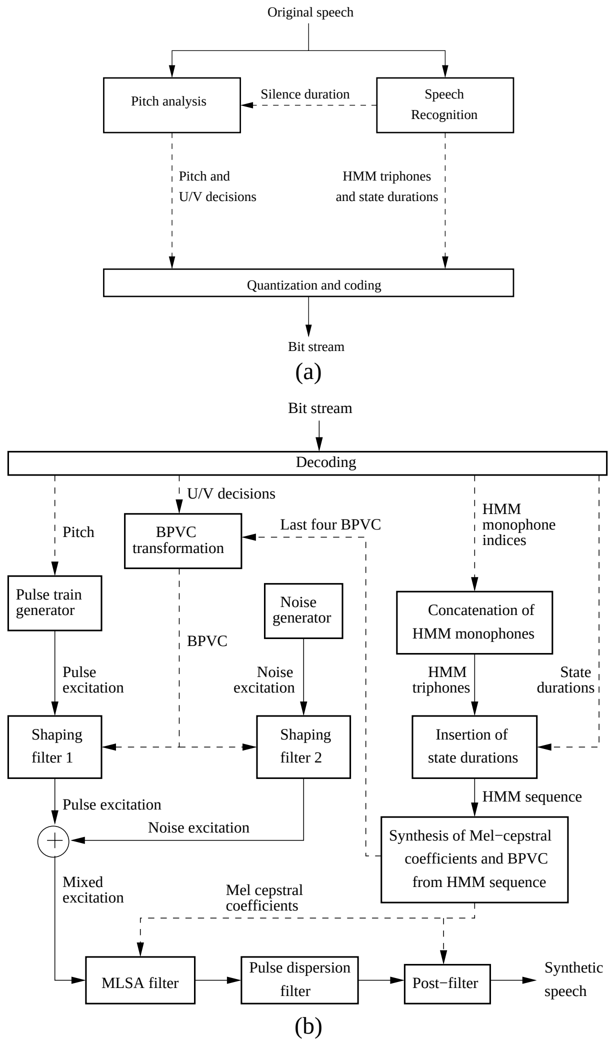
where $\vec{c}_i = [c_0 \ldots c_M]^T$, $\Delta\vec{c}_i = [\Delta c_0 \ldots \Delta c_M]^T$, and $\Delta^2\vec{c}_i = [\Delta^2 c_0 \ldots \Delta^2 c_M]^T$ represent the mel-cepstral coefficients vector, and its related delta and delta-delta vectors for the i^{th} frame, respectively. T indicates transposition. A total of 13 mel-cepstral coefficients ($M = 12$) are extracted from the speech signal at every 5 ms using 25-ms Hamming windows centered on the corresponding frames.

A database composed of 160 phonetically balanced sentences, approximately 10 min, spoken by a male speaker in Brazilian Portuguese language sampled at 8 kHz was used to train the recognizer. At first, 49 monophones plus one silence were modeled. Afterwards, the monophones were cloned and the transition matrices were tied in order to create triphone models. A total of 2175 triphones plus one silence were modeled.

In addition to mel-cepstral coefficients, BPVC were also used to train the HMM models. These coefficients plus their respective delta and delta-delta parameters were used as a second stream during the training part of the speech recognizer. However, for speech recognition only the mel-cepstral coefficients are considered. Section 3 describes with more details the modeling of BPVC.

2.1.2. Pitch analysis

Pitch analysis is performed at every 20 ms on the speech signal after silence duration in the beginning and in the end of the sentence



(a)

(b)

Fig. 1. Diagram of the phonetic vocoder: (a) encoder; (b) decoder.

has been determined by the speech recognizer. This silence information is important to set synchronism between the pitch analysis and the speech recognition procedures.

In order to compute pitch period, the autocorrelation method based on the method employed by the MELP vocoder is used. Firstly, an integer pitch period P_{int} is computed from the input speech signal low-pass filtered at 1 kHz. Secondly, a fractional pitch refinement is taken using the input speech signal low-pass filtered at 0.5 kHz $s_{LP}(n)$, where the fractional pitch P_{frac} is determined in the interval $[P_{int}-5; P_{int}+5]$. The final pitch P is calculated from the low-pass filtered residual signal - obtained by inverse filtering the input speech signal $s_{LP}(n)$ through the inverse linear prediction filter - performing an integer pitch search in the interval $[P_{frac}-5; P_{frac}+5]$, with P_{frac} rounded to the nearest integer; and finally a fractional pitch refinement is once more applied to obtain the final pitch value P.

2.1.3. *Quantization and coding*

As for the quantization of the recognized triphones, since there are 50 monophones each model would be quantized with 6 bits. However, based on the fact that some monophones have more occurrences than others, Huffman coding is applied, giving rise to an average rate of 4.61 bits/monophone instead of 6 bits/monophone.

Table 1. Joint quantization of pitch and overall voicing decisions for each 80-ms super-frame.

U/V modes	3-bit CB	Additional bits	Total
UUUU	000	no additional bits	3
UUUV UUVU UVUU VUUU	001	2 bits → mode selection + 6 bits → scalar quantization	11
UUVV VVUU	010	1 bit → mode selection + 7 bits → 2-d VQ	11
VUUV UVUV VUVU UVVU	011	2 bits → mode selection + 6 bits → 2-d VQ (UVVU), or 3-d VQ (UVUV and VUVU), or 4-d VQ (VUUV)	11
UVVV VUVV VVUV VVVU	100	2 bits → mode selection + 7 bits → 3-d VQ (UVVV and VVVU), or 4-d VQ (VUVV and VVUV)	12
VVVV	101 110 111	9-bit 4-d VQ with CB No. 1 9-bit 4-d VQ with CB No. 2 9-bit 4-d VQ with CB No. 3	12

The state durations for each model are regarded as 3-dimensional vectors and vector quantization (VQ) is performed. One 128-entry codebook was designed using the LBG algorithm for all monophones, where Huffman coding is also applied, decreasing the average number of bits necessary to quantize the durations for each monophone from 7 bits to 4.58 bits.

The logarithmic pitch and overall voicing decisions are quantized in a super-frame basis, based on the method presented at [1]. However, for the present scheme each super-frame comprises four consecutive 20-ms frames, resulting in 80 ms. If only one frame in the super-frame is voiced, scalar quantization is performed for the pitch lag with a 64-level quantizer. For the remaining configurations, VQ with different codebook sizes and dimensions is performed, as shown in Table 1. It can be noticed, for instance, that 3 bits/super-frame are necessary to perform quantization in the UUUU mode, whereas 12 bits are necessary in the VVVV mode. In this last case, which is the most critical one, three 512-entry codebooks are applied, giving rise to 1536 possibilities. For every super-frame the following distortion measure is used for VQ

$$D_p = \sum_{i=1}^{N} w_i (P_i - \hat{P}_i)^2 + \delta \sum_{i=1}^{N} w_i (\Delta P_i - \Delta \hat{P}_i)^2, \tag{3}$$

where P_i and $\hat{P}_i$ are respectively the i^{th} original and quantized pitch values in the pitch vector, the weights w_i are 1 for voiced and 0 for unvoiced frames, and N is the dimension of the VQ. The pitch differential ΔP_i is given by

$$\Delta P_i = P_i - P_{i-1}, \tag{4}$$

whereas $\Delta \hat{P}_i$ is obtained substituting P_i and ΔP_i in (4) by $\hat{P}_i$ and $\Delta \hat{P}_i$, respectively. The δ factor is used to control the contribution of pitch differentials in order to track the pitch trajectory, and for the present case this parameter is set to 0.75. The codebooks were designed by the LBG algorithm using the same database applied to train the HMM speech recognizer.

2.2. Decoder

The decoder receives pitch information, monophone indices and state durations indices from the encoder, as shown in Fig. 1(b). Speech is synthesized at every 20-ms frame, which are divided in four 5-ms sub-frames. Pitch period is the same for each frame whereas BPVC and mel-cepstral coefficients change at every sub-frame.

2.2.1. Mel-cepstral coefficients and BPVC extraction

Mel-cepstral coefficients and BPVC are synthesized as follows: firstly, the information of monophone indices are used to concatenate a triphone sequence of HMM. Secondly, the state durations for each triphone from the formed HMM sequence are inserted. Having the HMM sequence with the proper state durations inserted, mel-cepstral coefficients and BPVC are extracted at every 5 ms from this sequence, using the algorithm for feature generation from HMM described in [7].

2.2.2. BPVC transformation

The BPVC effectively employed during the synthesis $\{v_1, \ldots, v_5\}$ are obtained so that: if the current frame is unvoiced, the BPVC for all sub-frames in the current frame are set to zero; otherwise, for all the sub-frames, the first BPVC v_1 is set $v_1 = 1$ and the remaining ones, for $2 \leq i \leq 5$, are given by

$$v_i = \begin{cases} 1, & \text{if } \hat{v}_i > 0.6, \\ 0, & \text{Otherwise,} \end{cases} \qquad (5)$$

where $\{\hat{v}_2, \ldots, \hat{v}_5\}$ are the last four BPVC synthesized from the HMM models.

2.2.3. Excitation generation

To build mixed excitation, pulse and noise excitation should be added together. The initial pulse excitation corresponds to a pulse train whose period is the linear interpolated pitch between the previous and current frame. If the frame is unvoiced, a default pitch value of 50 samples is used [4]. The initial noise sequence is obtained by a Gaussian random noise generator with zero mean and unity variance. The initial pulse and noise excitation are filtered by the shaping filters $H_p(z)$ and $H_n(z)$, whose transfer functions are given by

$$H_p(z) = \sum_{j=1}^{5} v_j \sum_{i=0}^{64} b_{i,j} z^{-i}, \qquad (6)$$

$$H_n(z) = \sum_{j=1}^{5} (1 - v_j) \sum_{i=0}^{64} b_{i,j} z^{-i}, \qquad (7)$$

where $b_{i,j}$ represents the i^{th} coefficient for the j^{th} band from the synthesis filter bank, and v_j represents the BPVC for the j^{th} band. After filtering, pulse and noise excitations are added to compose the mixed excitation. The synthesis filter bank corresponds to a 5-band 64-order FIR filter bank with the following configuration: 0-0.5 kHz, 0.5-1 kHz, 1-2 kHz, 2-3 kHz, and 3-4 kHz.

Table 2. Average bit rate for the vocoder.

Parameter	Bits/model	Model/s	Bits/s
Triphones	4.61	13.76	63.39
State durations	4.58	13.76	63.01
Pitch	-	-	137.6813
Total	264.08 bits/s		

2.2.4. Speech synthesis

Speech is synthesized passing the mixed excitation through the Mel Log Spectrum Approximation (MLSA) filter [6], since the mel-cepstral coefficients $\{\hat{c}_0, \ldots, \hat{c}_M\}$ synthesized from the HMM sequence can model speech spectrum envelope. The pulse dispersion filter from MELP is also used, and finally a post-filter is applied to improve speech quality. The later is implemented using the MLSA filter with the coefficients $\{\hat{b}_0, \ldots, \hat{b}_M\}$ obtained from the mel-cepstral coefficients, for $0 \leq i \leq M$, through

$$\hat{b}_i = \begin{cases} 0, & i = 1 \\ \beta \hat{c}_i, & \text{Otherwise,} \end{cases} \qquad (8)$$

where the parameter β was set to 0.5.

2.3. Bit rate

For the train database, in average, a rate of 13.76 monophones/s was verified whereas pitch VQ, also applied to the train database, produced a bit rate of 137.6819 bps. Table 2 highlights the average bit rate for the vocoder, considering the average number of bits requested to quantize monophone indices and state durations.

3. BPVC MODELING

In [5], BPVC, Fourier magnitudes, and jitter were applied for phonetic vocoding in order to solve the problem of having unnatural synthetic speech. Nevertheless, according to subjective experiments Fourier magnitudes and jitter do not make significant difference in the speech quality for that proposed phonetic vocoding scheme. The mere application of BPVC would decrease significantly the bit rate, once the Fourier magnitudes request at least 8 bit for VQ. In [8] mixed excitation is employed by an HMM-based text-to-speech synthesis where all the mixed excitation parameters were modeled by one single HMM framework jointly with the spectral parameters. Based on this possibility, the training of HMM models with BPVC and mel-cepstral coefficients in a single framework would take the advantage of building mixed excitation for phonetic vocoding with no increment of bit rate when compared with the binary excitation pulse train/random noise.

The BPVC modeling was carried out during the training part of the HMM speech recognizer for the encoder, where the observation features comprised two streams:

- stream 1: mel-cepstral coefficients, and their delta and delta-delta, i.e., $\{c_0, \ldots, c_M, \Delta c_0, \ldots, \Delta c_M, \Delta^2 c_0, \ldots, \Delta^2 c_M\}$;
- stream 2: last four BPVC, and their delta and delta-delta, i.e., $\{v_2, \ldots, v_5, \Delta v_2, \ldots, \Delta v_5, \Delta^2 v_2, \ldots, \Delta^2 v_5\}$.

The BPVC used to train the HMM models were extracted at every 5-ms from the train database. The first BPVC v_1 was not modeled because this coefficient is responsible for the overall voicing decision, and this information is already consistently encoded jointly

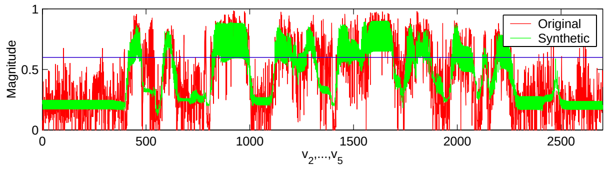


Fig. 2. Original and synthetic BPVC for one sentence from the test database. The horizontal line indicates the voicing threshold.

with pitch. Moreover, the HMM would be able to model with more precision four parameters than five. Despite BPVC modeling, only the mel-cepstral coefficients are considered to perform speech recognition, i.e., the output probability contribution for the stream containing the BPVC is set to zero. The reason for this lies on the fact that the BPVC consist on prosodic parameters, where their use could degrade the speech recognition process.

Figure 2 shows original BPVC, extracted from one sentence which was not used to train the models, and BPVC synthesized from HMM. One should perceive that only the last four BPVC $v_2, \ldots, v_5$ are drawn. It can be noticed that these parameters were modeled without loss of information, since the voicing threshold, represented by the horizontal line, almost conducts to the same voicing information for both cases, according to Section 2.2.2.

4. PERFORMANCE

A listening test was performed with a database composed of 40 phonetically balanced sentences which were not used to design the speech recognizer nor the pitch VQ codebooks. Ten listeners gave their respective intelligibility degrees for each sentence processed by the phonetic vocoder, and the average degree for each listener is shown in Table 3. It can be noticed that a good degree of intelligibility is achieved.

Table 3. Results for the test where 10 listeners gave their opinions about the intelligibility degree for 40 sentences.

List.	1	2	3	4	5	6	7	8	9	10	Mean
Int. (%)	71	80	75	82	85	87	77	83	79	80	80

In order to evaluate the effectiveness of mixed excitation, Figure 3 shows spectrograms of one sentence from the test set and its versions synthesized with mixed and binary excitations. The last one is obtained without considering the shaping filters, and by selecting pulse train for voiced and noise for unvoiced frames. It can be noticed that mixed excitation improves synthetic speech, specially in the high frequencies. A comparison test was also performed with 20 sentences from the test set with 10 listeners, where in average for 95% of the cases mixed excitation was preferred.

5. CONCLUSION

In this work, a phonetic vocoder which uses mixed excitation during the synthesis procedure was presented. The encoder carries out HMM-based speech recognition, whereas the decoder extracts from the HMM models mel-cepstral coefficients and bandpass voicing strengths. These last parameters are employed to build mixed excitation, while pitch information is vector quantized at every 80-ms super-frames. Experiments have shown that the vocoder at an average bit rate of 265 bps reaches good degree of intelligibility, and the use of mixed excitation significantly improves the quality when compared with the traditional excitation pulse train/random noise.

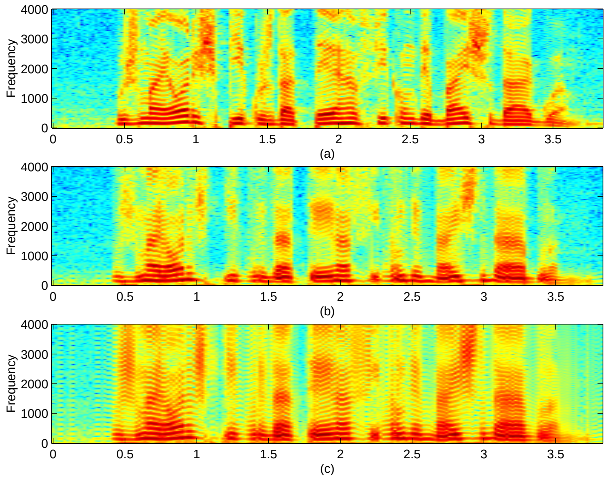


Fig. 3. Spectrograms for one sentence from the test set: (a) original; (b) processed by mixed excitation; (b) processed by binary excitation.

6. ACKNOWLEDGEMENTS

The authors would like to thank Professor Keiichi Tokuda for the useful discussions.

7. REFERENCES

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

[2] C. M. Ribeiro and I. M. Trancoso, "Phonetic vocoding with speaker adaptation," in *Proc. EUROSPEECH*, 1997.

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

[4] A. V. McCree and T. P. Barnwell III, "A mixed excitation LPC vocoder model for low bit rate speech coding," *IEEE Trans. Speech and Audio Processing*, vol. 3, no. 4, July 1995.

[5] R. da S. Maia, R. J. R. Cirigliano, D. Rojtenberg, and F. G. V. Resende Jr., "An HMM-based phonetic vocoder using mixed-excitation," in *Proc. ICOSYS*, Oct 2002.

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

[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.

[8] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "Mixed-excitation for HMM-based speech synthesis," in *Proc. EUROSPEECH*, 2001.