A Simple and Efficient Algorithm for the Compression of MBROLA Segment Databases

Olivier van der Vrecken, Nicolas Pierret, Thierry Dutoit, Vincent Pagel, Fabrice Malfrere
Laboratoire de Théorie des Circuits et de Traitement du Signal (TCTS)
Faculté Polytechnique de Mons - Bd Dolez, 31, B-7000 Mons (BELGIUM)
Tel : +32 65 37 41 78 - E-Mail : vdv@tcts.fpms.ac.be

Abstract

Most state-of-the-art TTS synthesizers are based on a technique known as synthesis by concatenation, in which speech is produced by concatenating elementary speech units. The design of a high-quality TTS system implies the storage of a large number of segments. To facilitate the storage of these segments, this paper proposes a very low complexity coder to compress unit databases with a toll quality. A particular interest has been taken in the databases used by the MBROLA synthesizer, composed of fixed-length pitch periods with constrained harmonic phases. The coder developed here uses this special characteristic to reach compression rates from 7 to 9 without degrading the speech quality produced by the synthesizer, and with very limited computational cost.

1. Introduction

Most state-of-the-art TTS synthesizers are based on a technique known as synthesis by concatenation, in which speech is produced by concatenating elementary speech units, called segments. Segments are generally stored in the form of a stream of parameters related to a given speech model.

Well-chosen speech models allow data-size reduction, an advantage that is hardly negligible in the context of concatenation-based synthesis, given the high amount of segments to be stored. This feature has been rarely addressed in the scientific literature ; it is, however, an important criterion for the choice of a given speech model in end-user, low cost TTS systems.

The efficiency of a TTS system to produce high quality speech is therefore partly subordinated to the amount of degradation introduced by the speech coding phase.

Regarding the quality of synthetic speech, segments should obviously exhibit some basic properties:

- They should account for as many coarticulatory effects as possible. An obvious way of accounting for articulatory phenomena is to provide many variants for each phoneme. For instance, Hunt and Black [15] propose to concatenate non-uniform units that take into account both the prosodic and phonetic appropriateness of the units. Their method requires large databases (from 10 to 150 minutes).

- Given the restricted smoothing capabilities of the concatenation algorithm used in concatenative TTS systems, segments should be easily connectable. To obviate this difficulty, the use of longer units decrease the density of concatenation points, therefore providing better speech quality.

These basic properties involves large databases as the number of segments, and their length, are not kept as small as possible.

Whatever the synthesis method used, some trade-off is therefore necessary between the quality of synthetic speech and the size of the databases.

Accordingly, once a set of segments has been chosen, the data compression capabilities associated with speech models becomes a determinant factor for the ultimate size of the database. In Dutoit [1][7], the database compression capabilities of three leading speech models were studied in the context of TTS synthesis : the autoregressive model, the Hybrid Harmonic/Stochastic model and the « null model » implemented by TD-PSOLA. A sampling frequency of 16 kHz was assumed, which is a typical requirements for wide-band, high-quality synthesis. This study can be summarized as follows :

- ***Linear prediction synthesis of order 18 :*** Roy and Kabal [4] announce 2400 bit/s for coding 16 LSPs, so that a nearly transparent coding for p=18 could reasonably be achieved with about 3800 bit/s. White [5] reports on an average rate of 4500 bit/s for 18 LSP coefficients in the context of a variable rate CELP coder.

- A few years ago, most coding research teams involved in sinusoidal***, harmonic, or hybrid synthesis*** have struggled to tally the parameters of their coders into the 4800 bit/s required by the new federal standard of the U.S. Department of Defense ([6], [8], [9], [10]). Deketelaere [11] has further reduced bit rate to 2300 bit/s, and even to 900 bit/s thanks to vector quantization techniques. Regarding sinusoid amplitudes, important data reduction can be achieved by storing the envelope spectrum in the form of LPC parameters (a residual error variance and a set of LSP coefficients) [13]. The idea is readily extended to stochastic component variances, with the difference that a reduced prediction order can be used. Abrantes and Marques [10] have chosen 10 and 6 for the voiced and unvoiced prediction orders, respectively. Phase coding has proven to be a more difficult task, since coarse quantization results in systems that are reverberant. Marques [9] has proposed to predict harmonic phases from frame to frame and encode the prediction error. In contrast, an analytical model of phase as a function of frequency has been developed by Mac Aulay and Quatieri [6]. To our knowledge, there has still been no attempt to optimize the compression performance of hybrid models in wide-band conditions. One can therefore only put forward the coarse approximation of 10 kbit/s, obtained by extrapolation of the compression ratios reached for narrow-band coding.

- ***The TD-PSOLA "model"*** does not require any parameter estimation stage (except for pitch marking). It does not lend itself to any specific data reduction technique and can a priori be associated to any existing speech coding compression technique. In practice, the range of profitable coding algorithms is drastically constrained by the intrinsically high quality and simplicity of PSOLA. A zero-tap DPCM coder seems to be best suited, since it results in only one extra addition per sample at synthesis time. A transmission ratio of 40 kbit/s has been reported by Le Faucheur [12] for a sampling frequency of 8 kHz. A little bit less than 80 kbit/s seems to be a reasonable value for 16 kHz, since additional redundancy can be turned into account in the high frequency spectrum. This, for instance, reduces the size of a complete French diphone database recorded at 16 kHz from more than 5 Mbytes down to approximately 1.7 Mbytes.

2. Compression of MBROLA synthesizer Databases

Although Dutoit [14] also introduced a new model for concatenative speech synthesis, termed as MBROLA and recently made available for scientific use in the context of the MBROLA internet project [2], and predicted high compression ratios given the particular format of MBROLA databases, no particular coding scheme was proposed for this model. We now bridge this gap.

2.1. Structure of the MBROLA synthesizer database

In order to be effective, a coding system must be adapted to the particular characteristics of the database to compress.

The French diphone database we used for the experiments reported here is the one supplied to internet users in the context of the MBROLA project. Each diphone is composed of a sequence of constant length frames. Each frame comprises 120 samples and is tagged with a voiced/unvoiced, stationary/transient flag (2 bits). Voiced frames have a constant pitch period of 120 samples and identical low-frequency harmonic phases. These features have precisely been used produce a particularly efficient coder.

2.2. Requirements attached to the implementation of the coding algorithm

The encoding operation can be performed off-line and no constraint is composed on the computational complexity of the coder. On the other hand, the complexity of the decoder must be kept small in comparison with the average number of operations required by the MBROLA algorithm (3.3 operations per sample on the average).

2.3. Description of the algorithm

2.3.1. Principle

An analysis-by-synthesis coder with forward LP analysis has been chosen to encode diphones from the database [3]. The diphones are reconstructed by filtering an excitation vector, which is a combination of vectors derived from a single adaptive codebook and a variable number of stochastic codebooks, through a short-term predictor synthesis filter,

$$\frac{1}{A(z)} = \frac{1}{1 - \sum_{k=1}^{p} a_k z^{-k}}$$

where a_k are the short-term prediction coefficients and p is the order of the filter.

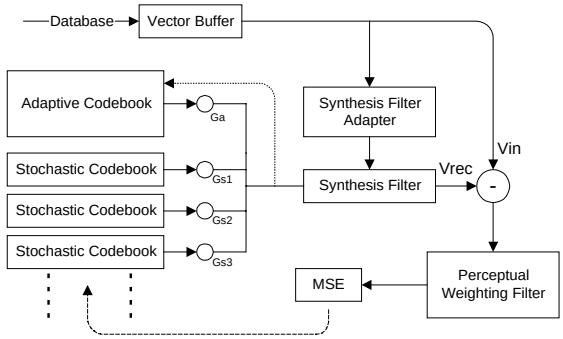


Fig 1. The database encoder.

The synthesis filter coefficients are adapted every frame. They result from a 2nd order LPC analysis on the original signal. The particularly small order has been chosen to limit the complexity of the decoding section. These coefficients are quantized on 8 bits.

We also adapt the coefficients of a perceptual filter every frame but they derive from a 10th order LPC analysis on the original signal. This perceptual filter has the standard transfer function,

$$W(z) = \frac{H\left(\frac{z}{G_1}\right)}{H\left(\frac{z}{G_2}\right)} \quad with\ G_1 = 0.9,\ G_2 = 0.6.$$

The excitation signal is found by minimizing the weighted mean-squared error over several samples, where the error signal is obtained by filtering the difference between the original and the reconstructed diphones through the perceptual weighting filter W(z).

The quality of the reconstructed diphones is fixed by choosing, once for all the frames, a minimum signal to noise ratio (SNR) defined as,

$$SNR = 20 \log \left(\frac{\sum_i (V_i in)^2}{\sum_i (V_i in - V_i rec)^2} \right)$$

where Vin and $Vrec$ are respectively an input frame and a reconstructed frame of the diphone to process.

Not only the diphone databases used by the MBROLA synthetizer are composed of pitch periods with constant length, but also the initial phases of the first harmonics of voiced segments are fixed to a constant value throughout the databases (i.e. the value of the initial phase only depends on

the order of the harmonic). With this characteristic, successive the frames in MBROLA databases are maximally similar. Therefore, the last excitation will be a perfect model for the next excitation and can be implemented by an adaptive codebook in the coder architecture.
This adaptive codebook gives its contribution for each frame and represents the long-term (periodic) component of the signal to encode.

Since the database is composed of frames with constant length, the task of this codebook simply consists of storing the excitation of the previous frame to form the basis of the next excitation. The codebook is implemented as a buffer with a capacity of 120 samples and an adaptative gain accounting for the evolution of the amplitude of the signal. This gain, proportional to the correlation between the reconstructed frame produced by the adaptive codebook and the original frame, is quantized on 4 bits. If the adaptive codebook is sufficient to reach the required signal to noise ratio, the excitation vector reduces to this adaptative component. If it is not the case, the excitation vector receives further contributions from *stochastic codebooks*.

If stochastic codebooks are used, the frames are decomposed into a set of subframes (e.g., 10 samples). For each subframe, the best complementary excitation vectors are searched in the stochastic codebooks. A parameter, denoted as *Nbr_Max*, fixes the maximum number of stochastic codebooks used to model each frame. The excitation vectors from the stochastic codebooks are then simply added to the long-term excitation from the adaptive codebook to obtain the expected signal to noise ratio.

Since the analysis involves synthesis, the description of the analysis procedure completely describes the decoder.

2.3.2. Initialization of the coder

An initialization problem appears for the first frame of each diphone to compress, i.e., 1244 times for the French database used in these experiments. This problem therefore has to be considered seriously.

Coding introduces bigger distortions at the beginning of each diphone as illustrated on Figure 2, which shows the evolution of the error energy as a function of the index of frames in a diphone. Moreover, this effect influences several frames.

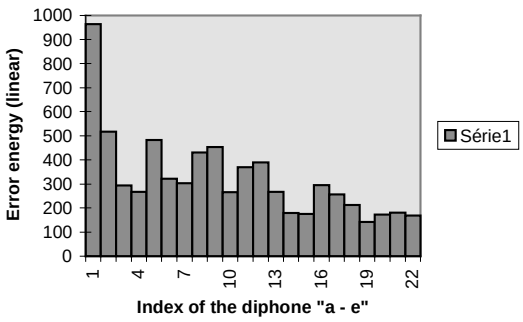


Figure 2 : Evolution of the energy of the error (on a linear scale) as a function of the index of frames in diphone [a-e].

To solve this problem, a second parameter, denoted as *Nbr_First*, defines the maximum number of stochastic codebooks allowed for the first frame of each diphone. This parameter is set to a value higher then *Nbr_Max*.

2.3.3. Complexity of the decoding algorithm

In order to evaluate the complexity of the decoding algorithm, we have synthetized several hundreds of synthetic speech signals (about 10 minutes on the whole) with the French coded database. The resulting average number of operations required to synthesize diphones with uncompressed and compressed databases is equal, respectively, to 3.3 and 8.2, which gives a contribution of 4.9 operations/sample for the coder.

2.4. Construction of the compressed database

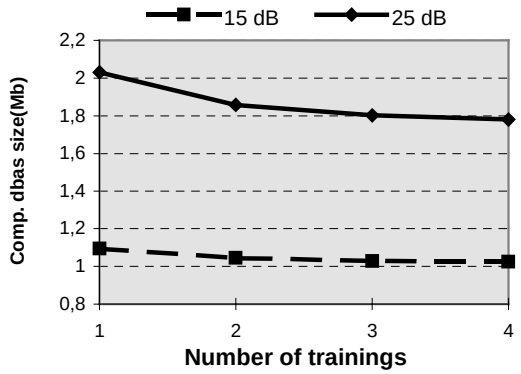
The compressed databases are composed of two parts : the tables for the quantification of the variables of the coder and compressed data themselves, (indexes in the quantization tables, number of codebooks used for each frames).

2.4.1. Storage of the information related to the quantization tables

The gain associated to the adaptive codebook is quantized on 4 bits. Since the adaptive codebook is supplied by the previous excitations, it does not take part in the elaboration of the compressed database.

In the current implementation of the coder, only one codebook is used and stored in the compressed database. This codebook contains 1024 vectors of 10 components. The associated gain is quantized on 64 levels (6 bits). The stochastic codebook is trained in a recursive way on one particular database so as to optimize this codebook to the compression of this database.

We illustrate hereafter the training of an English database with a size of 8.347 Mbytes. Two series of recursive trainings have been performed with signal to noise ratios equal to 25 dB and 15 dB, respectively. After 4 training passes, we obtain compressed databases with sizes equal respectively to 1.78 Mbytes and 1.025Mbytes.



2.4.2. *Storage of the information related to the speech vectors from the database*

For each frame, the following information is stored sequentially in the compressed database : the index of the

adaptive gain, the number of stochastic codebooks used (Nbr_Cod: 4 bits), an index for the quantization of the synthesis filter coefficients (8bits), the indexes corresponding to the stochastic codebooks (10 bits) and the associated gain (6 bits). In practice, typical Nbr_Max values are 0, 1 or 2. The following table summarizes the number of bits required to code each frame of the database :

Number of codebooks	4 bits
Adaptive gain	4 bits
Stochastic	Nbr_cod * Nt * 10 bits
Stochastic gains	Nbr_cod * Nt * 6 bits
VQ LPC	8 bits

where *Nt* is the number of subframes for the 120 samples (12 in the case of the French database).

3. Results

In this section, we take the quality of the synthetic speech produced by MBROLA with uncompressed databases as the reference.

The quality of a coded database can be adjusted by fine-tuning three parameters: the signal to noise ratio for each frame, the maximum number of stochastic codebooks (Nbr_Max) and the maximum number of stochastic codebooks for the first frame of each diphone (Nbr_First).

A MOS test has been realized with a group of 12 persons familiarized with the listening conditions to evaluate the distortions due to the integration of coding algorithm in the MBROLA synthesizer. These listeners have rated phonetically balanced records according to a five-level quality scale. Ratings are obtained by averaging numerical scores over several hundreds of speech records. The next table gives the MOS scores for our 4.6 Mbytes French diphone database.

	A	b	c	d
Nbr_Max	1	1	2	2
Nbr_First	2	3	5	5
SNR	10	12	20	25
Size (Kb)	334	429	619	759
Ratio	14	10.9	7.6	6.2
Score	2.8	3.1	3.7	4.2

For a compression rate equal to 7.6, it can be seen that the French database presents a toll quality. The last trial (d) offers a very good quality for a compression ratio of 6.2. It was found to be transparent by listeners.

As of today, the MBROLA project has made available diphone databases for French (4.57 Mb), German (10.77 Mb), Spanish (2.62 Mb), Romanian (3.53 Mb) and Dutch. Since downloading such large databases through the internet network is not a straightforward tasks, all these freely available databases will be also provided in the compressed format as described in this paper. Given the high compression ratios (for the reduced additional computational load) we have obtained, it is now even possible to store high-quality, wide-band MBROLA databases on floppy disks.
New databases in Brazilian Portuguese, Swedish, in Korean, Breton, will be soon available.

4. Conclusion

It is generally believed that high quality time domain synthesis techniques requires big databases, unless they are combined with a complex coder, which tremendously increases their computational cost. This papers shows the contrary ; the MBROLA synthesis technique exhibits virtually unequaled data compression capabilities (as compared with other time domain synthesis techniques) with very limited additional computational cost (about 5 operations per sample).

Consequently, the MBROLA synthesizer still runs in real time on an Intel386 processor, even with the compression scheme mentioned above. More information and examples of synthesis speech files are available at the URL address : http://tcts.fpms.ac.be/synthesis/.

5. References

[1] T. Dutoit ,"High Quality Text-To-Speech Synthesis of the French Language", Ph. D. dissertation, Faculté Polytechnique de Mons. October 1993.

[2] T. Dutoit, V. Pagel, N. Pierret, F. Bataille and Olivier van der Vrecken, "The MBROLA Project: Towards a Set of High Quality Speech Synthesizer Free of Use for non commercial purposes", International Conference on Speech and Language Processing, Philadelphia, 1996.

[3] M.R. Schroeder and B. Atal, "Code-Excited Linear Prediction (CELP): High-Quality Speech at Very Low Bit Rates", ICASSP 1985, pp. 937-940.

[4] G. Roy and P. Kabal, "Wideband CELP Speech Coding 16 kbit/s", International Conference on Acoustics, Speech, and Signal Processing (ICASSP), vol. 1, pp. 17-20, 1991.

[5] S. White, "Codeur CELP à débit variable: application au codage de diphones", Proc. 19ème Journée d'Etudes sur la Parole, Montreal, 1990.

[6] R.J. Mac Aulay and T.F. Quatieri, "Sine-wave phase coding at low data rates", Proc. ICASSP 91, S9.1,pp. 577-580.

[7] T. Dutoit," *An Introduction to Text-to-Speech Synthesis*", Kluwer Academic Publishers, April 1997.

[8] J.C. Hardwick and J.S. Lim, "A 4.8 kbit/s Multi-Band Excitation Speech Coder", Proc. ICASSP 88, pp. 374-377.

[9] J. Marques, L. Almeida, J. Tribolet, "Harmonic coding at 4.8kbit/s", Proc. ICASSP 90, vol. 1, pp. 17-20.

[10] A.J. Abrantes and J.S. Marques, "Hybrid Harmonic Coding of Speech", EUSIPCO 92, 25-28 August 92, Brussels, pp. 487-491.

[11] S. Deketelaere, H. Leich, B. Wery, I. Deman, M. Dothey, "A New Quantization Technique for LSP Parameters and its Application to Low Bit Rates Multi-Band Excited Vocoders", EUSIPCO 92, Brussels, pp. 475-478.

[12] L. Le Faucheur, O. Boeffard, B. Cherbonnel, S. White, "Un algorithme de synthèse de parole de haute qualité", Proc. Séminaire SFA/GCP, 1991, pp. 104-107.

[13] D. Rowe, W. Cowley, A. Perkis, "A Multi-Band Excitation Linear Predictive Speech Coder", Proc. Eurospeech 91, Genova, pp. 239-243.

[14] T. Dutoit, B. Gosselin, " On the use of a hybrid harmonic/stochastic model for TTS synthesis-by-concatenation", Speech Communication 19, 1996, 119-143.

[15] A. Hunt and A. Black, « Unit Selection in a Concatenative Speech Synthesis System using a Large Speech Database », ICASSP 96, Vol 1, 373-376.