Low Bit-Rate Speech Codec Based on a Long-Term Harmonic Plus Noise Model

Faten Ben Ali, Sonia Djaziri-Larbi, Laurent Girin

Low bit-rate Speech Codec based on a Long Term Harmonic plus Noise Model

F. Ben Ali*, S. Djaziri-Larbi, L. Girin,

Abstract

The long-term harmonic plus noise model (LT-HNM) for speech shows an interesting data compression, since it exploits the smooth evolution of the time trajectories of the short-term harmonic plus noise model parameters, by applying a discrete cosine model (DCM). In this paper, we extend the LT-HNM to a complete low bit-rate speech coder. A Normalized Split Vector Quantization (NSVQ) is proposed to quantize the variable dimension LT-DCM vectors. The NSVQ is designed according to the properties of the DCM vectors obtained from a standard speech database. The obtained LT-HNM coder reaches an average bit-rate of 2.7kbps for wideband speech. The proposed coder is evaluated in terms of modeling and coding errors, bit-rate, listening quality and intelligibility.

Index Terms

Low bit-rate, speech coding, long term modeling, harmonic plus noise model, variable dimension vector quantization.

I. Introduction

In speech/music coders and analysis/synthesis systems, spectral modeling is generally made on a short-term (ST) frame-by-frame basis: every 20ms or so. This is the case for most spectral models, including the linear prediction (LP) model [1] and the sinusoidal model [2], [3]. The main justification of the ST processing is that the signal is only locally (quasi-) stationary and in interactive applications, the segmentation is necessary for quasi-real-time processing.

For speech signals, the evolution of the vocal tract configuration and glottal source activity is quite smooth and regular for many speech sequences. Therefore, high correlation between successive ST spectral parameters has been evidenced and can be exploited, especially in coding applications. For example, inter-frame LSF correlation is exploited in the LP coding schemes of [4] and in matrix quantization [5]. In parallel, some studies have attempted to explicitly take into account the smoothness of LP spectral parameters evolution in speech coding techniques [6].

In all those studies, the interframe correlation has been considered “locally”, that is, between only two (or three for matrix quantization) consecutive frames. This is mainly because full-duplex telecommunications require limiting the coding delay. This constraint can be relaxed in many other applications in half-duplex communication, storage, or transformation and synthesis. These applications include archiving, Text-to-Speech modificatin/synthesis, telephony surveillance data, digital answering machines, electronic voice mail, digital voice logging, electronic toys and video games [7]–[9].

In particular, transformation and synthesis of speech in the decoder is an important application with relaxed delay constraints. Transformation systems need an efficient and flexible representation of signals and a flexible access to the parameters for easy manipulation of the signal in the decoder. In MPEG-4 Parametric Audio Coding, audio signals (speech and music) are represented by object-based models (harmonic tones, individual tones and noise). This representation of signals by frequency and amplitude parameters permits simple and independent pitch and playback speed modifications at the decoding stage [10]–[12].

In such applications, the analysis-modeling-coding-synthesis process can be considered on larger signal windows, i.e. on what is referred to as a long-term (LT) section of signal in the following. In that vein, the Temporal Decomposition technique [13], [14] consists of decomposing the trajectory of spectral (LP) parameters into “target vectors” which are sparsely distributed in time and linked by interpolative functions. This method has been applied to speech coding [15], and it remains a powerful tool for modeling the temporal structure of speech signals. Following another idea, the authors of [16] proposed to compress matrices of LSF parameters using a two-dimension (2D) transform, e.g. a 2D Discrete Cosine Transform (DCT), similarly to block-based image compression. They provided interesting results for different temporal sizes, up to 20 (10ms-spaced) LSF vectors. A major point of this kind of method is that it jointly exploits the time and frequency correlation of LSF values.

More recently, Dusan et *al.* have proposed in [17] to model the trajectories of ten consecutive LSF parameters by a fourth-order polynomial model. In addition, they implemented a very low bit rate speech coder exploiting this idea. At the same time, it was proposed in [18] to model the LT trajectory of sinusoidal speech parameters (the phase and the amplitude of each harmonic) with a Discrete Cosine Model (DCM). In contrast to [17], where the length of parameter trajectories and the order of the model were fixed, in [18] the long-term frames are continuously voiced sections of speech, which exhibit very variable size and “shape”: such a section can contain several phonemes or syllables. Therefore, the LT-DCM is adjusted to

F. Ben Ali and S. Djaziri-Larbi are with Université Tunis El Manar, Ecole Nationale d’Ingénieurs de Tunis, Signals and Systems Lab, Tunis, Tunisia. e-mail: ben_ali_faten@yahoo.fr, sonia.larbi@enit.rnu.tn.

L. Girin is with GIPSA-lab, Univ. Grenoble Alpes, and INRIA Grenoble Rhône-Alpes, Grenoble, France. e-mail: laurent.girin@gipsa-lab.grenoble-inp.fr.

the characteristics of the modeled speech section, resulting in a variable trajectory size and model order, compared to the ten-to-four conversion of [17]. In [19], this adaptive scheme was extended to the LT-modeling of spectral envelope parameters, leading to a so-called 2D-cepstrum. Again, only voiced speech sections were processed, and they were considered as purely harmonic. The LT-DCM modeling has also been extended to LSF parameters in [20], including quantization aspects and the processing of both voiced and unvoiced sections.

An important extension of the LT-modeling within the sinusoidal framework has been proposed in [21], [22] based on the two-band Harmonic+Noise model (HNM) of [23]. Such HNM is particularly appropriate for modeling mixed voiced/unvoiced speech signals. In [21], [22], the DCM has been applied to the trajectories of the two-band HNM model parameters: the spectral envelope that here encompasses both harmonic and noise amplitude parameters, the fundamental frequency F_0, and the voicing cut-off frequency F_V that separates the low-frequency harmonic band and the high-frequency noise band. The results of [21], [22] have thus generalized the modeling of the spectral envelope to any harmonic/noise combination, and has introduced the LT modeling of the F_V parameter.

In the present paper, we extend the LT-HNM presented in [21], [22] to a complete low bit-rate LT-HNM speech coder by addressing quantization issues. Before entering into technical choices and details, it can be noted that, although the sinusoidal model and its different variants (including the HNM) have shown good performance in various speech processing applications such as speech transformation and synthesis [23]–[26], only a few works have attempted to implement a speech codec based on the ST sinusoidal model. This can be due to the difficulty of coding variable-size sets of amplitudes, and possibly frequencies and phases, especially if no harmonicity is assumed.

In [27], spectral amplitudes and corresponding frequency positions are gathered in pair-vectors and coded using a vector quantization, while phases are scalar quantized. The obtained speech codec provides bit-rates in the range of 3.75–7.75 kbps for narrowband speech. A low bit-rate narrowband 2.4/4.8kbps speech codec based on the ST sinusoidal model is presented in [28]. To reduce the parameter set, the sinusoidal components are forced to fit a harmonic model for voiced speech as well as for unvoiced speech (a low fundamental frequency is chosen for noise representation). Harmonic amplitudes are then transformed to a fixed length cepstral parameters set and transformed back to frequency domain for DPCM (Differential Pulse Code Modulation) quantization.

The objective of this paper is to present a methodology for the design of a (very) low-bitrate long-term speech coder based on the Harmonic + Noise Model, and using existing ST-HNM analysis-synthesis methods and our previous work on long-term spectral modeling. In the present paper we thus focus on quantization aspects.

More specifically, the novelty lies in the vector quantization of the LT-DCM vectors that model the time trajectories of the ST-HNM parameters. A main challenge is to cope with the dimension variability of the LT-DCM vectors across LT-sections (in addition to the dynamic variability). Therefore, the proposed LT-HNM coder focuses on the design of a vector quantization stage directly fitted to the properties of the LT-DCM coefficients, especially their dimension variability and their dynamics. In the literature, different quantization methods are proposed, taking into consideration these two properties: i) a mean-gain-shape approach [29] is used when the coefficient values have a large dynamic, and ii) a split vector quantization technique is proposed to face the variable vector dimension [30]. We follow this general line, and we propose to apply a normalized split vector quantization (NSVQ) technique to quantize the LT-DCM vectors corresponding to the LT time-trajectories of spectral amplitudes, fundamental frequency and voicing cut-off frequency. In the core of the paper, we motivate the choice for this technique, w.r.t. other possible solutions.

Importantly, it must be made clear that the objective of this paper is not to design and thoroughly evaluate the best possible long-term coder, nor it is even to show that the HNM is the best short-term model candidate to be integrated in the LT framework for such a task. Rather, it is to show the feasibility and potential efficiency of the long-term approach to speech coding in the HNM framework, i.e. we want to show that the long-term approach can lead to a LT-HNM coder that is more efficient than the ST-HNM (with similar ST parameterization) in terms of quality/bitrate ratio (postulating that the delay is not an issue in the targeted applications).

The paper is organized as follows. In Section II, a summary of the ST-HNM is given to introduce the parameters to be LT-modeled. An overview of the LT-HNM, relying on previous work, is presented in Section III. In Section IV, we present the proposed NSVQ approach for the LT-DCM vectors. Statistics of LT-DCM vectors properties and the design of the quantization stage are presented and discussed in Section V. Experimental results related to coding errors, listening quality, intelligibility measure and obtained bit-rates are presented and discussed in Section VI.

II. Short Term Harmonic plus Noise Model (ST-HNM)

The HNM concept has been first proposed in [31] as the multi-band excitation model: it splits the frequency band into voiced and unvoiced sub-bands, where voiced sub-bands are modeled by harmonic components, whereas unvoiced bands are modeled by (colored) noise. This model is dedicated to represent sounds with a mixed harmonic/noise structure, such as mixed voiced/unvoiced sounds of speech. This model inspired the two-band HNM and the residual error HNM, both proposed by Stylianou *et al.* in [23], [32].

In this study, we used the two-band ST-HNM presented in [21], [22], based on the generic two-band HNM of [23]. The frequency band is split into two sub-bands, as illustrated by Fig. 1: a harmonic sub-band containing harmonics of the speech

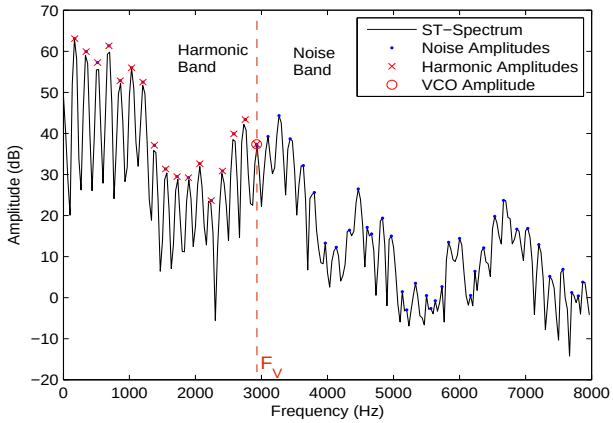

Fig. 1: Two-band HNM: harmonic lower sub-band containing harmonics of F_0 and noise upper sub-band.

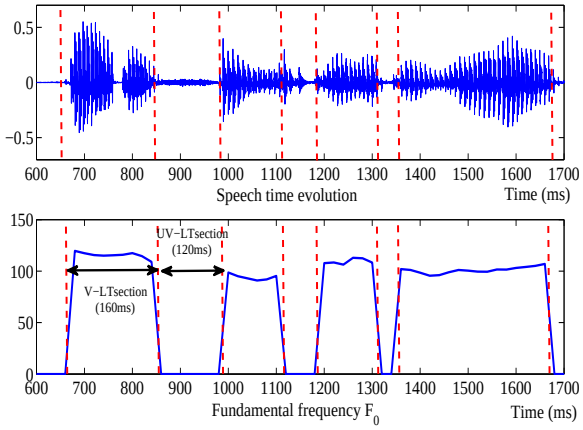

Fig. 2: Example of temporal segmentation of speech into voiced (V) and unvoiced (UV) LT-sections.

signal (low frequencies), and a noise sub-band containing high frequency noise components. These two bands are separated by a time-varying voicing cut-off frequency F_V, which is the last harmonic frequency in the harmonic band. In this model, the speech signal is segmented into short-term frames with a duration of 30ms and a fixed hop-size of 20ms, as in [22]. The ST-HNM parameters are extracted from each ST-frame, as detailed in [22]. For each ST-frame, these parameters are:

- **Fundamental frequency F_0:** F_0 is obtained for each ST-frame using the autocorrelation-based method of [33].
- **Voicing cut-off (VCO) frequency F_V:** F_V is computed only for voiced ST-frames using the technique given in [34]. For unvoiced frames, F_V is set to zero. Since F_V is the frequency of the last harmonic in a ST-frame, we have: $F_V(k) = I_k F_0(k)$, where I_k and $F_0(k)$ are respectively the number of harmonics and the fundamental frequency in the k^{th} ST-frame.
- **Harmonic parameters:** For each k^{th} ST-frame, a harmonic amplitude vector with size I_k is obtained by the iterative analysis-by-synthesis method described in [3]. The corresponding harmonic frequencies are obtained by multiplying $F_0(k)$ by the harmonic order.
- **Noise parameters:** The noise band is modeled by the sum of sinusoids at different noise frequencies. For each ST-frame, noise amplitudes and frequencies are obtained by a peak-picking technique, similar to that used in [2].

III. Long Term Harmonic plus Noise Model (LT-HNM)

The aim of the LT-modeling of ST-parameter time-trajectories is to capture the temporal correlation between successive ST-parameters. This has the advantage to reduce significantly the size of the model data. The implementation of the LT-HNM is detailed in [21], [22]. We summarize in the following the LT modeling steps.

A. Segmentation of the speech signal into LT-sections

The speech signal is first segmented, according to F_0 values, into LT-sections, i.e. blocks of ST-frames of variable duration, based on voiced/unvoiced decision. Each LT-section is either entirely voiced ($F_0 \neq 0$ for all successive frames) either entirely

unvoiced ($F_0 = 0$ for all successive frames). Typically, the duration of a long-term section can be several hundreds of milliseconds, and may contain up to ca. 60 ST-frames. This temporal segmentation is illustrated in Fig.2. The LT-model is then applied to the trajectories of ST-parameters along each LT-section.

B. Discrete Cosine Model (DCM) for the LT-modeling of the ST-HNM parameters trajectories

This study is based on the DCM to model the time-trajectories of the ST-HNM parameters within a LT section. The DCM approaches the data by a discrete sum of cosine functions. This model was first used for cepstral modeling in [28], [35]. Then it was applied to the LT modeling of harmonic parameters in [18], [19] and LP parameters in [20]. The DCM is defined as follows:

$$\tilde{X}(n) = \sum_{p=0}^{P} C(p) \cos(p\pi \frac{n}{N}),\ n = 1, \cdots, N, \tag{1}$$

where the vector $\tilde{\mathbf{X}} = [\tilde{X}(1), \cdots, \tilde{X}(N)]^T$ is the DCM model of the data vector $\mathbf{X}$, both of length N and indexed by n. $\mathbf{C} = [C(0), \cdots, C(P)]^T$ is the DCM vector of $P+1$ coefficients, where P is the DCM order. In cepstral modeling, $\mathbf{X}$ represents the log-spectrum amplitudes and n is a frequency index [35]. In LT-modeling, $\mathbf{X}$ contains the time-trajectory of a parameter and n is a time index. In a general manner, the DCM exhibits a good numerical stability compared to other models, especially the polynomial model when P becomes large.

In [21], [22], a detailed description of the application of this model to the trajectories of the ST-HNM parameters is given. The LT-DCM coefficients $\mathbf{C}$ are computed by minimizing a Weighted Mean Square Error between model and data. Two iterative algorithms are proposed in [22] to determine the optimal model order. A first "1D" iterative algorithm is applied to the trajectory of F_0 on each LT (voiced) section, to provide the optimal LT-DCM coefficient vector $\mathbf{C_{F_0}}$. This algorithm is also applied to the trajectory of the voicing cut-off (VCO) frequency F_V to provide $\mathbf{C_{F_V}}$. For the LT modeling of the spectral amplitudes, harmonic and noise amplitudes in a ST-frame are first gathered in a unique vector. Then a two-dimension DCM is applied. The first DCM is applied within each ST-frame along the frequency axis (the same model order is used for all ST-frames in a LT-section). The second DCM is a time-dimension DCM along a LT-section, applied to the time-trajectory of each coefficient obtained from the first frequency-domain DCM. For each LT section, we obtain a LT-DCM coefficient matrix denoted $\mathbf{C_A}$. The first dimension of the matrix, is the frequency DCM order plus 1, and the second dimension, is the temporal DCM order plus 1. Both orders are determined by the iterative algorithm presented in [22]. This 2D-DCM can be seen as an extension of the 2D-cepstrum of [19] to the HNM model.

C. LT-HNM speech synthesis

The time-trajectories of the LT-modeled ST-HNM parameters are obtained from the LT-DCM coefficients $\mathbf{C_{F_0}}$, $\mathbf{C_{F_V}}$ and $\mathbf{C_A}$ by applying (1).[1] The mathematical details are given in [21]. The HNM synthesized speech signal is the summation of a purely harmonic signal and a noise-like signal as detailed in [21]. Harmonic amplitudes are obtained by sampling the modeled spectrum at harmonic frequencies (multiples of the modeled F_0), while a regular sampling of the noise sub-band is used to obtain the noise amplitudes and noise frequencies. Harmonic amplitudes are linearly interpolated across frames, and cumulative instantaneous phases are approached by a continuous summation of harmonic frequencies (multiplied by 2π) with null initial phases for each harmonic trajectory. The noise-like signal is synthesized using an overlap-add technique, with random phases, similar to [3].

IV. LT-DCM COEFFICIENTS CODING

In this section, we present the core contribution of the present paper, i.e. the coding techniques that we applied to our LT-HNM in order to derive a complete LT-speech coder. The parameters to be coded and sent to the receiver for each LT-section are: i) the LT-DCM coefficients of the HNM parameters trajectories ($\mathbf{C_{F_0}}$, $\mathbf{C_{F_V}}$ and $\mathbf{C_A}$), and ii) the LT-section length K (the numer of ST-frames in a LT-section), which is required for synthesis. For simplicity, and when appropriate, we use in the following a common and simplified notation $\mathbf{C}$ for all DCM vectors, i.e. $\mathbf{C_{F_0}}$, $\mathbf{C_{F_V}}$ and the rows of $\mathbf{C_A}$. We propose to apply a mean-gain-shape vector quantization (VQ) to the LT-DCM coefficient vectors $\mathbf{C}$, while a binary representation is used for the LT-section length K. Note that the Discrete Cosine Transform (DCT), which is close to DCM, has been widely used in image and video coding [36] and a modified DCT (MDCT) is used in some high quality audio coders as the MPEG-2 AAC standard [37]. However, to our knowledge, no previous studies dealt with the quantization of DCM coefficients for speech applications.

A. Scalar quantization of mean LT-DCM coefficient

To guide our choices for the design of the LT-DCM quantizers, we first observed the distribution of the LT-DCM vector coefficients. For this aim, we applied the LT-HNM on the training speech material described in Section V-A. This resulted in

[1] For spectral amplitudes, (1) is first applied on the time axis, and then on the frequency axis for each ST-frame.

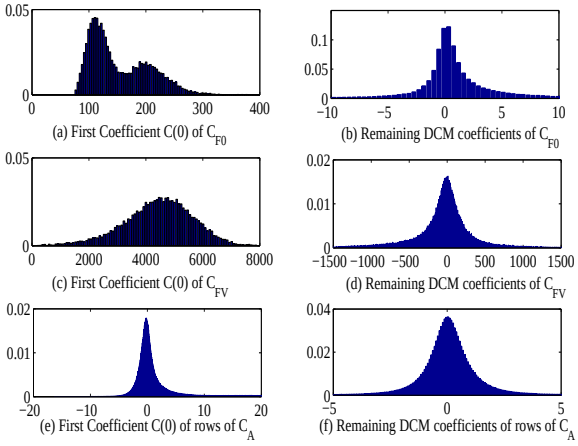

Fig. 3: Normalized Histograms of $C(0)$ and of the remaining coefficients of $\mathbf{C}$, for F_0, F_V and rows of $\mathbf{A}$. $C(0)$ values are higher than the remaining coefficients of $\mathbf{C}$ for F_0 and F_V.

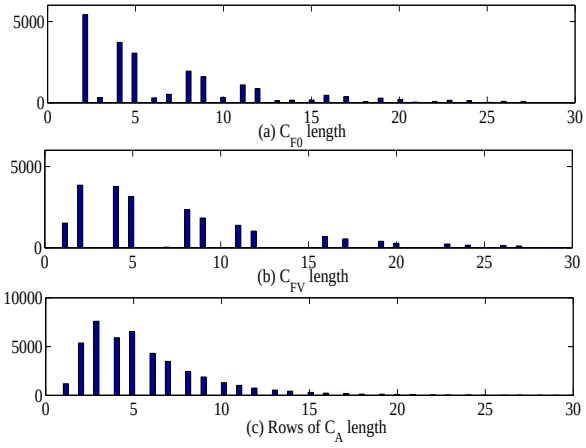

Fig. 4: Length variability of the LT-DCM vectors $\mathbf{C_{F_0}}$, $\mathbf{C_{F_V}}$ and rows of $\mathbf{C_A}$. The vector lengths lie in [1, 20] with some few vectors reaching 30 coefficients.

a database of LT-DCM coefficients composed, for each training LT-section, of two LT-DCM vectors, $\mathbf{C_{F_0}}$ and $\mathbf{C_{F_V}}$, and one LT-DCM matrix $\mathbf{C_A}$. Fig. 3 shows the histograms of the first LT-DCM coefficient of $\mathbf{C_{F_0}}$ (a), of $\mathbf{C_{F_V}}$ (c) and of the rows of $\mathbf{C_A}$ (e), compared with the histograms of the remaining coefficients of the LT-DCM vectors (Fig. 3 (b), (d) and (f)). The first coefficient of each LT-DCM vector $\mathbf{C_{F_0}}$ and $\mathbf{C_{F_V}}$, denoted $C(0)$, is significantly higher than the other values of the vector, since it represents the mean value of the modeled data trajectory. We note that this property is not noticeable in the case of $\mathbf{C_A}$ coefficients. We can see for example that the first coefficient $C_{F_0}(0)$ of $\mathbf{C_{F_0}}$ exhibits a bimodal distribution with modes at typical average F_0 values for male and female speech. Consequently, the first coefficients $C(0)$ are discarded from the vector quantization in order to increase its efficiency. Let us denote the new coefficient vectors and matrix rows (without the first coefficient $C(0)$) by $\dot{\mathbf{C}}_{\mathbf{F_0}}$, $\dot{\mathbf{C}}_{\mathbf{F_V}}$ and $\dot{\mathbf{C}}_{\mathbf{A}}$ ($\dot{\mathbf{C}}$ in generic form). Applying the mean-shape principle of vector quantization, the first coefficient $C(0)$ of each LT-DCM vector is coded separately using scalar quantization (the "shape" coding of $\dot{\mathbf{C}}$ is presented in the next subsection). Optimal scalar quantizers adapted to the statistical properties of the $C(0)$ database are designed by applying the Lloyd-Max algorithm [30].

B. Dimension variability of the remaining LT-DCM vectors

The LT-DCM vectors $\dot{\mathbf{C}}$ (be it $\dot{\mathbf{C}}_{\mathbf{F_0}}$, $\dot{\mathbf{C}}_{\mathbf{F_V}}$ or a row of $\dot{\mathbf{C}}_{\mathbf{A}}$) have variable dimension, due to the variable duration of LT-sections and to the dynamics of the time trajectories of the HNM parameters. Therefore, variable LT-DCM orders are obtained to reach the target LT-modeling errors. Fig. 4 shows the length variability of the LT-DCM vectors: The LT-model order is very scattered within the range $[1, 30]$. We thus deal with a variable dimension vector quantization problem, with possibly long vectors.

In the literature, some studies address the quantization of variable dimension vectors and propose some solutions adapted to each case of study. A non square transform vector quantization (NSTVQ) is proposed in [38], [39] to code harmonic

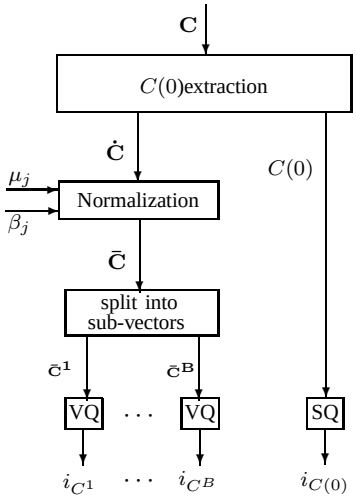

Fig. 5: Diagram of the proposed NSVQ.

amplitudes of the excitation in a LP codec: a non square linear transform is applied to the variable dimension vectors in order to obtain fixed length vectors which are then submitted to VQ. Another solution used for coding variable dimension harmonic amplitude vectors, and called Variable Dimension Vector Quantization (VDVQ), consists in designing a single universal fixed length codebook and using a binary selector vector that points on the non zero components of the harmonic amplitude vectors [40], [41]. In [42], the frequency scale is transformed from Hertz to Mel scale to obtain fixed-length harmonic amplitude vectors. The latter solution cannot be used in the case of the LT-HNM as it solves the dimension variability problem only on the frequency scale. The limit of the VDVQ is that the maximum vector length must be fixed, while in our case the maximum discrete cosine model order is controlled by the analysis-synthesis fitting of the model to the data. Concerning the NSTVQ, the proposed LT-HNM incorporates already one (two in case of spectral amplitudes) non square transform (the DCM) applied to each parameter leading to "decorrelated" and "energy-concentrated" coefficients: adding an additional non square transform prior to quantization may dangerously increase the information loss. In the following section, we develop a variable dimension vector quantization fitted to the particular constraints of the LT-HNM and to the particular characteristics of the DCM coefficients, referred to by the Normalized Split Vector Quantization (NSVQ).

C. Proposed Normalized Split Vector Quantization

The proposed NSVQ quantizer for the remaining DCM vectors $\dot{\mathbf{C}}$ is summarized in Fig. 5. As the LT-DCM vectors corresponding to F_0, F_V and spectral amplitudes $\mathbf{A}$ have similar characteristics, the same type of quantizer is applied to all of them, although a code-book is designed for each of them. Due to the shape and length variability of the DCM vectors, the proposed quantization technique is based on mean-gain-shape quantization and split vector quantization. The mean-gain-shape technique implies that we work with normalized coefficients, and the splitting technique consists in splitting a long vector into several sub-vectors [43], as shown on Fig. 5.

1) Normalization of the LT-DCM coefficients: The amplitude envelope of the coefficients within a given LT-DCM vector $\dot{\mathbf{C}}$ typically decreases with the coefficient rank. This results in an important variation of the DCM coefficient values across successive sub-vectors when splitting a DCM vector for quantization. In order to optimize the efficiency of the quantization codebook, we propose to normalize the LT-DCM vectors, such that all DCM coefficients vary in the same range, here in $[-1, 1]$. The purpose of the shape normalization is to facilitate the coding of all sub-vectors with the same codebook. In other words, the normalization enables to reduce the size of the codebook for a similar coding efficiency. We propose to apply the following vector normalization:

$$\bar{C}_i(j) = \frac{\dot{C}_i(j) - \mu_j}{\beta_j},\ j = 1, \ldots,\ \max_i\{P_i\}, \tag{2}$$

where $\dot{C}_i$ and P_i refer respectively to the LT-DCM vector indexed by i and the corresponding model order. μ_j and β_j are respectively the mean value and the maximum (absolute centered) value of all DCM coefficients of rank j in the training database, given for $j = 1$ to $\max_i\{P_i\}$ by:

$$\mu_j = \frac{1}{\text{card}\{\mathcal{M}_j\}} \sum_{i \in \mathcal{M}_j} C_i(j), \tag{3}$$

$$\beta_j = \max_{i \in \mathcal{M}_j} \{|C_i(j) - \mu_j|\}, \tag{4}$$

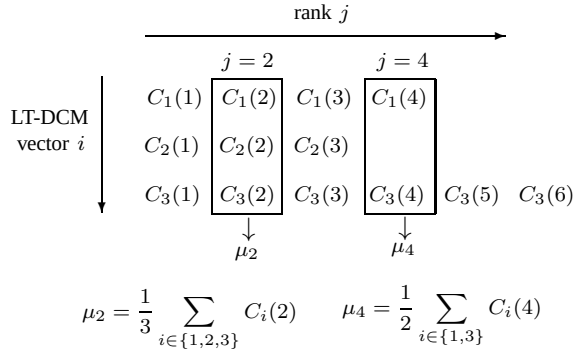

Fig. 6: Example of the mean values calculation used for vector normalization.

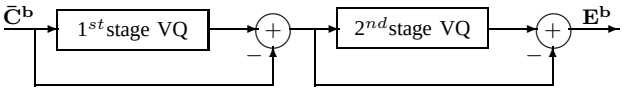

Fig. 7: Two-stage cascaded vector quantization.

where $\mathcal{M}_j$ is the set of LT-DCM vectors indices i with $C_i(j) \neq 0$ and card $\{\mathcal{M}_j\}$ is the cardinality of $\mathcal{M}_j$. Fig. 6 gives an example to explain the calculation of μ_j (and β_j): for a given rank j, μ_j is the mean of all coefficients of rank j in the LT-DCM vectors indexed in $\mathcal{M}_j$. Note that μ_j and β_j are calculated from a training database and then saved in the coder and the decoder, i.e. they are not concerned with quantization. Remember that the first coefficient $C(0)$ is not concerned by this normalization, since it is quantized separately (cf. IV-A). Note that (2) is inspired from the mean-gain-shape VQ in [29], except that in our case the normalization is carried out across all vectors of the training database, while in [29] it is a local normalization of each vector.

2) Splitting the normalized DCM vectors into equal-length sub-vectors: The LT-DCM vectors have a variable and possibly large dimension, as shown on Fig. 4. To avoid the use of a large training database and to reduce the size of the codebook, we propose to split the normalized vector $\bar{\mathbf{C}}$ into B smaller equal-length sub-vectors, denoted $\bar{\mathbf{C}}^{\mathbf{b}}$, $b = 1, \cdots, B$. Since the size of $\bar{\mathbf{C}}$ is not necessarily a multiple of the fixed sub-vector size, the last sub-vector of each vector is zero padded. Note that B is variable: it depends on the length of the corresponding LT-DCM vector.

3) Two-stage Vector Quantization: A two-stage vector quantization is applied to the fixed-length LT-DCM sub-vectors. The two cascaded vector quantizers, illustrated on Fig. 7, provide a higher quantization accuracy when using a training database with limited size, and much lower computational complexity compared to single-stage VQ [43]. The 1st-stage quantizer is applied to $\bar{\mathbf{C}}^{\mathbf{b}}$ while the resulting error vector is quantized by the 2nd-stage quantizer. The total quantization error corresponding to sub-vector $\bar{\mathbf{C}}^{\mathbf{b}}$ is given in the sub-vector $\mathbf{E}^{\mathbf{b}}$.

4) Coded stream: For each LT-section, the parameters sent to the receiver are the LT-section length K and the quantization indices of $C(0)$ and $\bar{\mathbf{C}}^{\mathbf{b}}$ for each HNM parameter. The number of sub-vectors B for each DCM vector must also be sent for each HNM parameter. The order P of the DCM applied to the spectral amplitudes on the frequency axis is also needed to determine the first dimension of the matrix $\mathbf{A}$.

D. The LT-DCM decoding

The decoding of the LT-DCM vectors is carried out by inverting the quantization and normalization steps given in IV-C. The decoded sub-vectors are represented by the codewords indexed by i_{C^b} in the codebook. We first concatenate the coded sub-vectors of each LT-DCM vector. Then, we apply the denormalization corresponding to (2):

$$C^q(j) = \beta_j \bar{C}^q(j) + \mu_j, \quad j = 1, \cdots, P_i, \tag{5}$$

where P_i is the order of the LT-DCM vector being decoded. Remind that the normalization coefficients μ_j and β_j are stored in the receiver and the exponent q refers to the coded data. The obtained DCM vector is finally concatenated to the decoded first coefficient $C^q(0)$ leading to the final coded LT-DCM vector $\mathbf{C}^{\mathbf{q}}$.

V. Codebooks design, bit allocations and bit-rates

In this section, we describe the experimental procedure for the design and the test of the proposed LT-HNM speech codec. We first describe the speech databases that we used for training and testing the codec. We then detail the design of the vector and scalar quantizers codebooks and we discuss different bit allocation configurations and the resulting bit-rates.

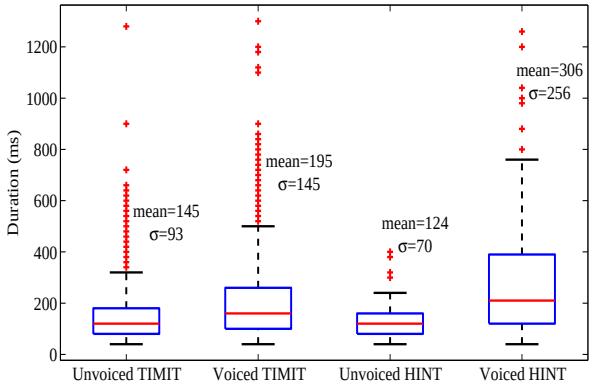

Fig. 8: Duration (ms) of voiced and unvoiced LT-sections for TIMIT and HINT test databases. On each box, the central (red) mark is the median, the edges of the box are the 1^{st} and 3^{rd} quartiles, the whiskers extend to the most extreme data points (outliers are plotted individually by the '+' signs).

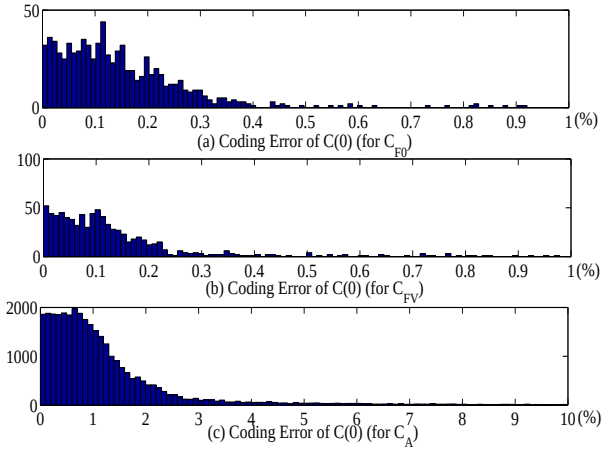

Fig. 9: Histograms of the coding rRMSE of $C(0)$ with an 8-bit optimal scalar quantization.

A. Speech material

In this study, we used the American TIMIT database [44], sampled at 16kHz. $2{,}720$ items of this database, each consisting of a complete sentence, were used for the training of the quantizers (38% female and 62% male speakers, with a total duration of about 122min). The segmentation of the training speech samples into voiced and unvoiced LT-sections yielded $44{,}122$ LT-sections: 49% voiced (ca. 69min) and 51% unvoiced (ca. 53min). The mean duration of a voiced LT-section is about 195ms and about 145ms for unvoiced LT-sections. The test database is composed of 300 items (150 female and 150 male speakers) with a total duration of about 14min. It is composed of $4{,}969$ LT-sections, 49% voiced and 51% unvoiced. Statistics about the duration of LT-sections are given in Fig. 8.

In parallel, a French speech database was used for subjective listening quality and intelligibility assessment with French speaking subjects. This database was developed for vocal audiometry for the Hearing in Noise Test (HINT) [45] and is composed of 20 phonetically balanced sentences (only male speakers) sampled at 16kHz, with a total duration of 63sec. The segmentation in LT-sections yielded 136 voiced LT-sections (66% of the total duration) and 174 unvoiced LT-sections (34% of the total duration). The statistics of the LT-section durations for HINT are displayed on Fig. 8.

B. Design of the VQ Codebooks

A VQ codebook is designed for each type of LT-DCM parameter vector, i.e. for F_0, F_V and the rows of $\mathbf{A}$.

A two-stage VQ is used, as detailed in Section IV-C3. Each codebook is optimized using the Linde-Buzo-Gray (LBG) algorithm [30], [43]. 50 iterations are run to obtain the codebook for each stage. This iteration number shows a good convergence of the optimization algorithm for each codebook.

C. Design of the scalar quantizers for $C(0)$

An optimal scalar quantization codebook is designed for the first coefficient $C(0)$ of each HNM parameter (F_0, F_V and $\mathbf{A}$). The scalar quantizers are optimized according to the distribution of these coefficients on Fig. 3. The histograms of the relative Root Mean Squared Errors (rRMSE) for each LT-DCM vector, given by equation (6), resulting from an 8-bit optimal scalar quantization, are shown in Fig.9: The results show that the coding errors lie around 0.1% for F_0 and F_V, and around 1% for spectral amplitudes A.

$$E_{C(0)} = \sqrt{\frac{(C(0) - C^q(0))^2}{C(0)^2}}, \tag{6}$$

where $C^q(0)$ is the coded value of $C(0)$.

D. Bit allocation and bit-rates

A different bit allocation is assigned to each codebook. We denote by N_0 the bit allocation of the coefficients $C(0)$ and by N_1 and N_2 the bit allocations of the first and the second stage VQ respectively. A different bit allocation (N_0, N_1, N_2) is assigned for each HNM parameter (F_0, F_V and $\mathbf{A}$). We discuss in the following the results for two configurations of the bit allocation, given in Table I. The first configuration corresponds to the largest codebook size we could generate, when taking into consideration database size, complexity and computing time limits, while the second configuration is a trade-off between low bit-rate and listening quality.

For each bit-allocation, the obtained average bit-rate R_T is the summation of four basic average bit-rates for the HNM parameters: R_{F_0}, R_{F_V} and R_A and R_K for the LT-section length K:

$$R_T = R_{F_0} + R_{F_V} + R_A + R_K. \tag{7}$$

For each LT-section, let B and P be the number of sub-vectors in a DCM-vector $\dot{\mathbf{C}}$ and the first dimension of $\mathbf{A}$ respectively. The obtained average bit-rates for each LT-section are given by:

$$R_{F_{[0,V]}} = \frac{1}{T}[N_0 + N_B + B(N_1 + N_2)], \tag{8}$$

$$R_A = \frac{1}{T}[N_P + N_B + (P+1)[N_0 + B(N_1 + N_2)]], \tag{9}$$

$$R_K = \frac{1}{T}N_K, \tag{10}$$

where $R_{F_{[0,V]}}$ can be R_{F_0} or R_{F_V}, N_B, N_K, N_P represent the number of bits used for the binary representation respectively for the number of sub-vectors B for a DCM-vector, the number of ST-frames K in a LT-section and the frequency-dimension DCM order P (first dimension of $\mathbf{A}$) and T is the duration of the LT-section. Note that the number of sub-vectors B is the same for all rows of the matrix $\mathbf{A}$ in a LT-section, since the same DCM order is used for the temporal dimension.

In Table I, we show the obtained average bit-rates over all LT-sections of the test database. Here a sub-vector length was set to 5 coefficients and the bit allocation was fixed to: $N_K = 7$, $N_P = 6$ and $N_B = 2$. The first bit allocation configuration yields an average bit-rate of 3,685 bps, while the second bit allocation configuration yields an average bit-rate of 2,721bps. Note that an important part of the bit-rate (ca. 88%) is dedicated to the coding of spectral amplitudes.
The coding errors corresponding to both considered bit allocations are evaluated in the following section.

VI. Evaluation of the complete LT-HNM speech codec

The evaluation of the LT-HNM speech codec is carried out using the test speech database described in Section V-A. We first provide illustrative examples of LT-modeled and quantized parameter trajectories. Then we present quantitative measures of LT-modeling/coding errors for each HNM parameter. Finally, perceptual listening quality of the coded speech is evaluated with the objective quality assessment algorithm PESQ [46], [47] (we used here WB-PESQ for wide-band speech) and with subjective mean opinion score (MOS) tests. Additional subjective tests are processed to assess the intelligibility of the coded speech.

	First Allocation				Second Allocation			
	N_0	N_1	N_2	Bit-rate (bps)	N_0	N_1	N_2	Bit-rate (bps)
F_0	8	9	6	179	6	7	5	143
F_V	8	9	6	181	6	7	5	144
$\mathbf{A}$	8	7	7	3284	6	5	5	2721
K	-	-	-	41	-	-	-	41
Total	-	-	-	**3685**	-	-	-	**2721**

TABLE I: Two configurations of bit allocation and corresponding average bit-rates for the quantization of the LT-DCM coefficients of each HNM parameter.

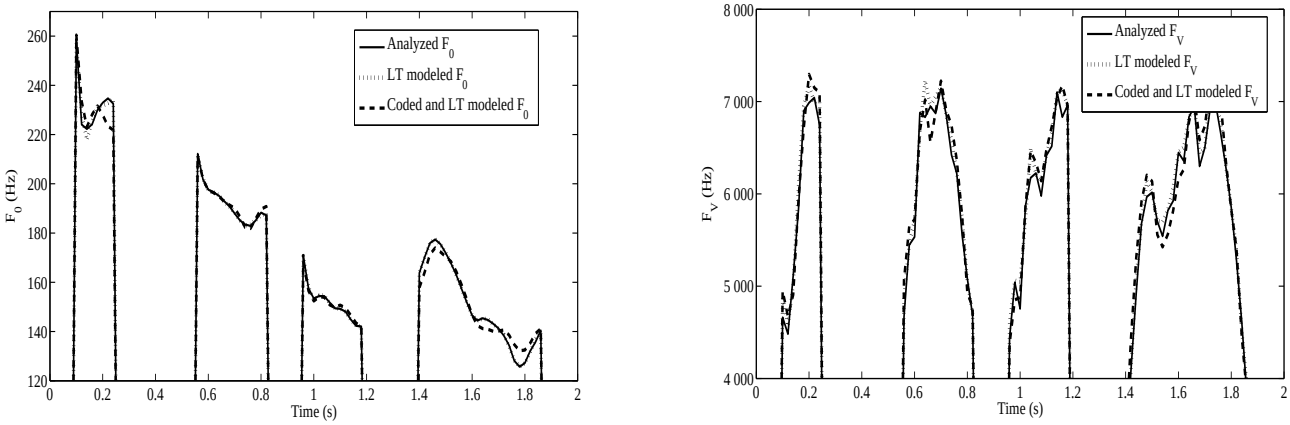

Fig. 10: Example of time trajectory of analyzed, LT modeled and both LT modeled and coded F_0 (top) and F_V (bottom) ($R_T \approx 3.6$ kbps). The LT-modeled trajectories fit better the analyzed values than the coded ones.

A. Examples of LT-modeled/quantized parameter trajectories

Fig. 10 illustrates an example of the reconstruction of the HNM parameters, after modeling with the LT-DCM and after LT-coding (LT-DCM + quantization) at $R_T \approx 3.6$ kbps. The time trajectories of F_0 and F_V are displayed in the left and right figure, respectively. Globally the trajectories of the LT-modeled parameters and of the LT-modeled and quantized parameters follow well the original (i.e. ST) trajectories. We note on this example that the reconstruction of F_0 is more accurate than that of F_V, i.e. closer to the ST parameter trajectories.

Fig. 11 displays an example of reconstructed spectral amplitudes vector in a voiced ST-frame, after LT-modeling and after LT-modeling + quantization. We see in this figure that globally, the spectral shape is well modeled and coded by the proposed technique. In this example, the effect of the quantization is moderate compared to the effect of the LT-modeling. In addition, the LT-modeling is less accurate in the noise-band compared to the harmonic band.

B. Measure of the coding and modeling errors

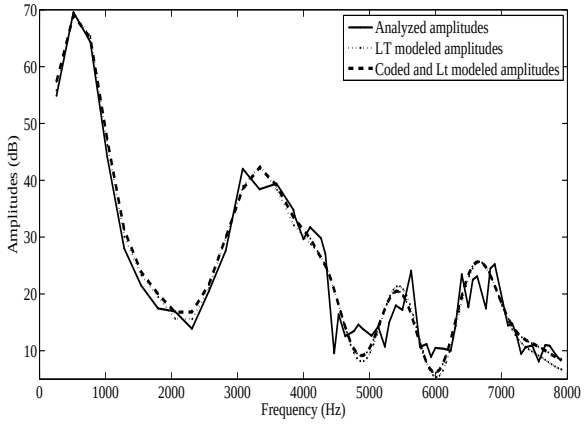

Fig. 11: Example of analyzed, LT modeled and both LT modeled and coded spectral amplitudes in a voiced ST-frame. ($R_T \approx 3.6$ kbps)

Three errors are considered for each HNM parameter, as depicted on Fig. 12: i) e^{LT}, the LT-modeling error, ii) e_q, the quantization error and iii) e_q^{LT}, the total coding error resulting from both LT-modeling and quantization. These errors are evaluated for each LT-section indexed by m. For the frequencies F_0 and F_V, we compute the error rate in $\%$ (rRMSE) as:

$$rRMSE = \sqrt{\frac{1}{K}\sum_{k=1}^{K}\frac{(F(k)-\widetilde{F}(k))^2}{F(k)^2}}, \tag{11}$$

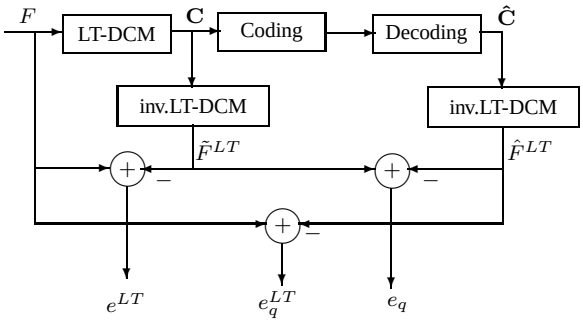

Fig. 12: The considered errors: LT-modeling error e^{LT}, quantization error e_q and total error (or LT+coding error) e_q^{LT}.

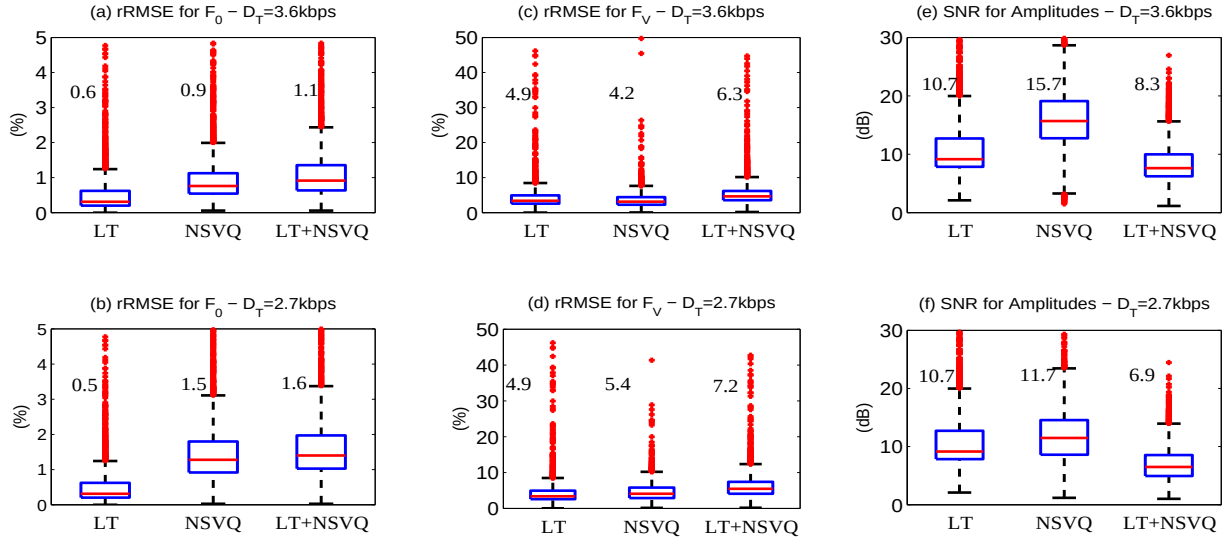

Fig. 13: The obtained rRMSE for F_0 (a and b) and F_V (c and d), and SNR for amplitudes (e and f); (a), (c) and (e): $R_T \approx 3.6$ kbps; (b), (d) and (f): $R_T \approx 2.7$ kbps. Scores on the plots indicate the mean values of the data box.

where F refers to F_0 or F_V and $\widetilde{F}$ is the modeled and/or coded version of F, and k and K are respectively the index and the number of ST-frames in the LT-section. For the spectral amplitudes $\mathbf{A}$, a signal-to-noise ratio (SNR) is evaluated in dB for each LT section according to:

$$\text{SNR} = 10\log_{10}\left[\frac{1}{K}\sum_{k=1}^{K}\frac{\sum_{f=1}^{N_k} A_f(k)^2}{\sum_{f=1}^{N_k}(A_f(k)-\widetilde{A}_f(k))^2}\right], \tag{12}$$

where N_k is the number of frequency components in the k^{th} ST-frame, and $\widetilde{A}_f$ is the modeled and/or coded version of A_f.

Fig. 13 displays the statistics of the errors of the three parameters F_0, F_V and $\mathbf{A}$. Both bit-allocations of Table I are considered. Comparing the results of Fig. 13 (a) and (b) to (c) and (d), we may note that the errors on F_0 are smaller than those on F_V. This is in part due to the dynamic behavior of the time trajectories of F_V compared to the smoother time trajectory of F_0, as illustrated in Fig. 10. Another reason is the rounding of the modeled F_V values to a multiple of the modeled F_0, which induces cumulative errors. In a general manner, modeling error and quantization error cumulate to yield the total error (see the related discussion in [48]). For F_0, the LT-modeling error is significantly lower than the quantization error, hence the quantization error is much closer to the total error. In other words the total error is mostly due to the quantization. This confirms the observation made in Fig. 10. For F_V, the contributions of the LT-modeling and of the quantization to the total error are more balanced. In contrast, Fig. 13 (e) and (f) show that, at $R_T \approx 3.6$kbps, the distortion due to the LT-modeling of the amplitudes is higher than that caused by the quantization. Indeed, the mean SNR due to LT-modeling is around 10.7dB, while it reaches 15.7dB for the quantization. The resulting average SNR for the complete LT-coding process is about 8.3dB. This confirms the observation made in Fig. 11.

$R_T \approx 2.7$kbps is a configuration with a better balance between LT-modeling and quantization: the corresponding average SNRs are closer (about 10.7dB and 11.7dB respectively). The total average error is 6.9dB. As expected, the overall results of Fig. 13 confirm that e_q, and thus e_q^{LT}, are higher at the lower bit-rate.

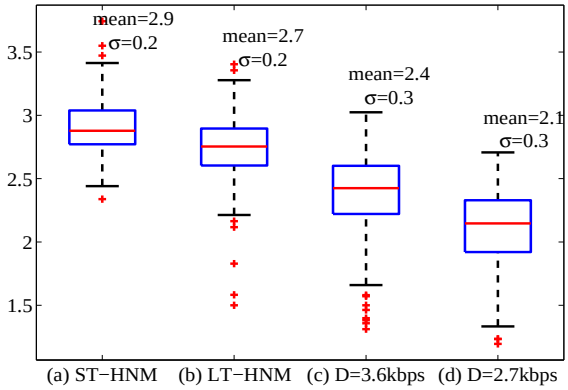

Fig. 14: Listening quality: PESQ scores of TIMIT test database.

The way the modeling error and the quantization error contribute to the total error is not easy to characterize and is not expected to be linear. The optimal control of the total error by an automatic "weighting" of the LT-modeling and quantization processes remains out of the scope of the present study, but it is thus a very interesting perspective to improve the proposed LT-HNM codec.

C. Listening quality assessment

We first assess the perceptual listening quality of the synthesized speech using the ITU-T standard Perceptual Evaluation of Speech Quality (PESQ) algorithm [47]. PESQ is an intrusive measure, i.e. it compares the degraded speech to the original sample and delivers a score in the range from -0.5 (*bad quality - very annoying*) to 4.5 (*excellent quality-imperceptible*). Fig. 14 shows the PESQ scores obtained for all the test database and at different steps of the LT-HNM coder: Fig. 14(a) corresponds to the PESQ scores of the ST-HNM modeled speech, while Fig. 14(b) shows the scores of the LT-HNM speech (LT-DCM modeling of HNM parameters without quantization). From these results, it is clear that the main quality degradation is due to the first step, i.e. the ST-HNM, where the mean PESQ score is 2.9, which indicates a *slightly annoying impairment*, whereas the LT-HNM speech displays a score of 2.7, which is in the same quality range (*slightly annoying*).

Fig. 14(c) and Fig.14(d) show the mean and standard deviation of the PESQ scores of the coded LT-HNM speech for both considered bit-rates. The PESQ scores corresponding to $R_T \approx 3.6$kbps indicate a mean score degradation of about 0.3 compared with the LT-HNM results, which seems reasonable. And, as expected, Fig. 14(d) shows that the speech quality decreases with the bit-rate: at $R_T \approx 2.7$kbps, the mean PESQ score reaches 2.1, which corresponds to *annoying* quality.

Note that the overall average PESQ scores degradation from ST-HNM to coded LT-HNM speech ($R_T = 2.7$kbps) is about 0.8 (from 2.9 to 2.1), which emphasizes again that the overall listening quality loss is to a large extent due to the initial ST-HNM representation of the speech signal, and not only to LT-modeling and quantization. We believe that a series of improvement can be conducted, not only on the proposed LT-coding techniques, but also on the initial ST-HNM on which these LT-coding techniques were applied.

To confirm the objective ratings, subjective listening tests were also carried out in-lab with 12 nave male and female french speaking listeners, aged within 23-30 years, using the HINT database (in French) [45]. Subjects listened (with high-quality headphones) to randomly played speech samples, composed of original, ST- and LT-HNM synthesized samples without and with coding (at $R_T \approx 2.7$ kbps). Listeners were asked to rate the listening quality of the heard sentences according to the ITU-T P.800 recommendation [49], using Absolute Category Rating (see Table II). For comparison, the PESQ scores for the French HINT database were also computed.

The obtained MOS and PESQ scores for the French test database are shown on Fig. 15. We first note that, in the case of the LT-HNM (with and without coding), the average PESQ and MOS scores are similar (about 2.5 for the LT-modeled speech and 1.9 for the coded speech at $R_T = 2.7$kbps), which proves a high correlation between objective (PESQ) and subjective

MOS	Quality	Impairment
5	Excellent	Imperceptible
4	Good	Perceptible but not annoying
3	Fair	Slightly annoying
2	Poor	Annoying
1	Bad	Very annoying

TABLE II: Mean Opinion Score (MOS).

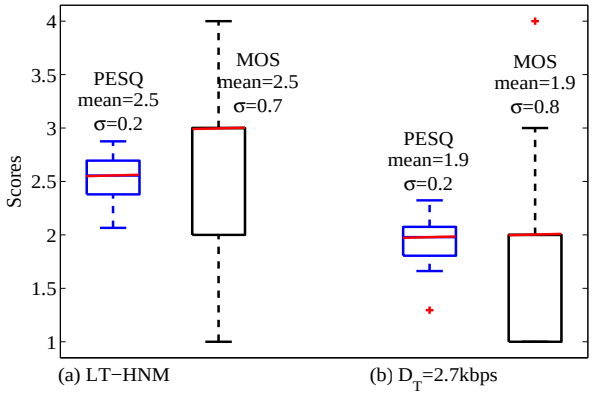

Fig. 15: PESQ and MOS scores for HINT speech samples: (a) ST-HNM, (b) LT-HNM only, (c) Coded speech at $R_T \approx 2.7kbps$.

(MOS) quality scores when applying the LT-model. However, this is not true for the ST-HNM synthesized samples, where the mean PESQ score is equal to 3.1, while the mean MOS score reaches 3.6. We also note that the average PESQ and MOS scores of the LT-modeled and the coded HINT samples (2.5 and 1.9 respectively), are lower than the average scores of TIMIT samples (2.7 and 2.1 respectively) (cf. Fig. 14). This may be due to the presence of longer LT-sections in French, as observed on Fig. 8, where the average duration of TIMIT voiced LT-section is 195ms, while it reaches 306ms for the French database. In addition, the quantizers were designed using training samples from TIMIT and not from HINT (since this latter database is not large enough). Also, note that some studies have reported a language dependency of PESQ assessment tool [50] [51] [52].

D. Intelligibility measure

Subjective intelligibility tests have also been conducted to assess the intelligibility of the LT-HNM modeled/coded speech. The Hearing in Noise Test measures a person's ability to hear speech in quiet and in noise, it has been developed for medical use to measure the sentence Speech Reception Threshold (sSRT[2]) [53], but this test is nowadays widely used to evaluate the speech intelligibility of enhanced and coded speech [54]. We carried out the HINT test with 12 French speaking subjects who listened (with high-quality headphones) to 12 different French speech samples from the French database: 6 LT-HNM and 6 coded LT-HNM at 2.7 kbps. They were asked to repeat each sample after listening to it. The intelligibility is measured by the rate of correct words from all listened words over all test samples [54]. We obtained an intelligibility rate of 99.7% for the LT-HNM synthesized speech, and 94.5% for the coded LT-HNM speech, which indicates that the coded LT-HNM speech provides a good intelligibility even if the listening quality was rated as annoying.

E. Discussion

Although the results presented above show that the proposed coder provides a good intelligibility at low bit-rates, the enhancement of the global listening quality remains an important issue for the comfort of the user.

It seems too early to compare the performance of the proposed coder with thoroughly optimized commercial coders, as the NB-HVXC or the WB-EVS (wide-band enhanced voice services codec) for example, which provides a good quality (MOS$\approx$3.5) at 5.9kbps [55]. We emphasize that the results of section VI are related to the coding of wide-band speech at such low bit-rates as 2.7kbps. However, it is worth to note that the MPEG-4 parametric audio coders HVXC (Harmonic Vector Excitation Coder) [56] and HILN (Harmonic and Individual Lines plus Noise) [57] provide listening quality of the coded narrow-band signals at 2 and 6kbps, respectively, which lies in the same range (MOS$<$3) as the results of Fig. 14c).

According to the quality ratings of Fig. 14 and Fig. 15, it is clear that the listening quality degradation is mainly due to the modeling part of the coder (i.e. ST-HNM and LT-HNM) rather than to the quantization part. To reduce the speech distortion, it would be interesting to strengthen the modeling constraints on the ST- and LT-HNM (higher modeling order, lower modeling errors, etc.) to reach higher quality ratings prior to quantization. In addition, the impact of each parameter (frequencies $F_0,\ F_V$ and amplitudes $\mathbf{A}$) on the listening quality needs to be analyzed separately in order to recognize which of them has to be modeled more accurately. The quantization stage can then be evaluated at lower (and different) bit allocations (N_0, N_1, N_2) to achieve a trade-off between the target bit-rate and the listening quality.

[2]sSRT: in speech audiometry, it is the decibel level at which 50% of heard words can be repeated correctly by the subject.

VII. Conclusion

The objective of this paper is to evaluate the feasibility and efficiency of the LT approach to speech coding in the HNM framework. We thus presented the design of a complete low bit-rate speech coder based on the long-term harmonic plus noise model (LT-HNM) [22] by adding a variable-dimension vector quantization stage. To our knowledge, no previous studies addressed the quantization of DCM coefficients obtained from the LT-modeling of speech signals. Hence we carried out a statistical study of these coefficients to design an appropriate quantization technique. The proposed Normalized Split Vector Quantization (NSVQ) is adjusted to the properties of these DCM coefficients. We presented first experiments to evaluate the proposed LT-HNM speech coder with two bit allocations, achieving the average bit-rates 3.6kbps and 2.7kbps for wide-band speech. Although the proposed coder achieved good intelligibility at both tested bit-rates, the global signal quality can still be improved. The results of section V indicate that the modest listening quality is mainly due to the ST- and LT-modeling part of the coder, with mean PESQ scores of 2.9 and 2.7 respectively. Indeed, the quantization stage reduces the mean listening quality score by 0.3 and 0.6 respectively at 3.6kbps and 2.7kbps.

The LT-HNM coder that we propose in this paper can still be improved to make it good candidate for commercial applications. These improvements will be addressed in future work. Particularly, the ST and LT target modeling errors can be adjusted to achieve a given quality score prior to quantization. Then, a compromise between target bit-rate and global quality has to be achieved, for example by optimizing the bit allocation to the different HNM parameters according to their impact on the achieved quality. Besides, in order to decrease the bit-rate, we think about introducing perceptual criteria to reduce the short-term data-rate prior to quantization, as proposed in [58], where the auditory masking is exploited to discard inaudible frequency components from coding.

References

[1] B. Atal and S. Hanauer, "Speech analysis and synthesis by linear prediction of the speech wave," *The Journal of the Acoustical Society of America*, vol. 50, 1971.

[2] R. McAulay and T.F. Quatieri, "Speech analysis/synthesis based on a sinusoidal representation," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 34, no. 4, 1986.

[3] E.B. George and M.J.T. Smith, "Speech analysis/synthesis and modification using an analysis-by-synthesis/overlap-add sinusoidal model," *IEEE Transactions on Speech and Audio Processing*, vol. 5, no. 5, 1997.

[4] F.-R. Jean and H.-C. Wang, "Transparent quantization of speech LSP parameters based on KLT and 2-D-prediction," *IEEE Transactions on Speech and Audio Processing*, vol. 4, no. 1, 1996.

[5] C.S. Xydeas and C. Papanastasiou, "Split matrix quantization of LPC parameters," *IEEE Transactions on Speech and Audio Processing*, vol. 7, no. 2, 1999.

[6] F. Norden and T. Eriksson, "Time evolution in LPC spectrum coding," *IEEE Transactions on Speech and Audio Processing*, vol. 12, no. 3, 2004.

[7] V. Ramasubramanian and H. Doddala, *Ultra Low Bit-Rate Speech Coding*, Springer, 2015.

[8] H. Hassanein, A. Brind'Amour, S. Dery, and K. Bryden, "Frequency selective harmonic coding at 2400 bps," in *37th Midwest Symposium on Circuits and Systems*, 1994.

[9] M. Hasegawa-Johnson and A. Alwan, "Speech coding: Fundamentals and applications," *Encyclopedia of Telecommunications*, 2003.

[10] B. Edler and H. Purnhagen, "Concepts for hybrid audio coding schemes based on parametric techniques," in *105^{th} Audio Engineering Society Convention*, 1998.

[11] H. Purnhagen, B. Edler, and C. Ferekidis, "Object-based analysis/synthesis audio coder for very low bit rates," in *104^{th} Audio Engineering Society Convention*, 1998.

[12] H. Purnhagen, "An overview of MPEG-4 audio version 2," in *Audio Engineering Society Conference: High-Quality Audio Coding*, 1999.

[13] B.S. Atal, "Efficient coding of LPC parameters by temporal decomposition," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 1983.

[14] A. M. L. Van Dijk-Kappers and S.M. Marcus, "Temporal decomposition of speech," *Speech Communication*, vol. 8, 1989.

[15] Y.M. Cheng and D. O'Shaughnessy, "On 450-600 b/s natural sounding speech coding," *IEEE Transactions on Speech and Audio Processing*, vol. 1, no. 2, 1993.

[16] N. Farvardin and R. Laroia, "Efficient encoding of speech LSP parameters using the discrete cosine transformation," in *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 1989.

[17] S. Dusan, J.L. Flanagan, A. Karve, and M. Balaraman, "Speech compression by polynomial approximation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 2, 2007.

[18] L. Girin, M. Firouzmand, and S. Marchand, "Perceptual long-term variable-rate sinusoidal modeling of speech," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 3, 2007.

[19] M. Firouzmand and L. Girin, "Long-term flexible 2D cepstral modeling of speech spectral amplitudes," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2008.

[20] L. Girin, "Adaptive long-term coding of LSF parameters trajectories for large-delay/very- to ultra-low bit-rate speech coding," *EURASIP Journal on Audio, Speech, and Music Processing*, , no. 1, 2010.

[21] F. Ben Ali, L. Girin, and S. Djaziri-Larbi, "Long-term modelling of parameters trajectories for the harmonic plus noise model of speech signals," in *International Congress on Acoustics (ICA)*, 2010.

[22] F. Ben Ali, L. Girin, and S. Djaziri-Larbi, "A long-term harmonic plus noise model for speech signals," in *Conference of the International Speech Communication Association (Interspeech)*, 2011.

[23] Y. Stylianou, "Applying the harmonic plus noise model in concatenative speech synthesis," *IEEE Transactions on Speech and Audio Processing*, vol. 9, no. 1, 2001.

[24] T. Quatieri and R.J. Mcaulay, "Speech transformations based on a sinusoidal representation," in *IEEE International Conference Acoustics, Speech, and Signal Processing (ICASSP)*, 1985.

[25] T.F. Quatieri and R. McAulay, "Shape invariant time-scale and pitch modification of speech," *IEEE Transactions on Signal Processing*, vol. 40, no. 3, 1992.

[26] Y. Stylianou, J. Laroche, and E. Moulines, "High-quality speech modification based on a harmonic+ noise model," in *European Conference on Speech Communication and Technology (EUROSPEECH)*, 1995.

[27] D. S. Likhachov and A. Petrovsky, "Parameters quantization in sinusoidal speech coder on basis of human auditory model," in *International Conference on Speech and Computer (SPECOM)*, 2004.
[28] R. McAulay and F. Quatieri, *Advances in speech signal processing*, chapter Low-rate speech coding based on the sinusoidal model, CRC Press, 1992.
[29] K.L. Oehler and R.M. Gray, "Mean-gain-shape vector quantization," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 1993.
[30] A. Kondoz, *Digital speech-coding for low bit rate communication systems*, John Wiley & Sons Ltd, 2004.
[31] D.W. Griffin and J.S. Lim, "Multiband excitation vocoder," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 36, no. 8, 1988.
[32] Y. Stylianou, "Decomposition of speech signals into a deterministic and a stochastic part," in *International Conference on Spoken Language (ICSLP)*, 1996.
[33] P. Boersma, "Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound," in *Proceedings of the institute of phonetic sciences*, 1993.
[34] K. Hermus, L. Girin, H. Van Hamme, and S. Irhimeh, "Estimation of the voicing cut-off frequency contour of natural speech based on harmonic and aperiodic energies," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2008.
[35] O. Cappe, J. Laroche, and E. Moulines, "Regularized estimation of cepstrum envelope from discrete frequency points," in *IEEE ASSP Workshop on Applications of Signal Processing to Audio and Acoustics*, 1995.
[36] J. Albert and J. Kari, *Handbook of Weighted Automata*, chapter Digital Image Compression, Springer, 2009.
[37] T. Painter and A. Spanias, "Perceptual coding of digital audio," *Proceedings of the IEEE*, vol. 88, no. 4, 2000.
[38] P. Lupini and V. Cuperman, "Vector quantization of harmonic magnitudes for low-rate speech coders," in *Global Telecommunications Conference (GLOBECOM)*, 1994.
[39] C. Li, P. Lupini, E. Shlomot, and V. Cuperman, "Coding of variable dimension speech spectral vectors using weighted nonsquare transform vector quantization," *IEEE Transactions on Speech and Audio Processing*, vol. 9, no. 6, 2001.
[40] A. Das, A.V. Rao, and A. Gersho, "Variable-dimension vector quantization of speech spectra for low-rate vocoders," in *IEEE Data Compression Conference (DCC)*, 1994.
[41] A. Das, A.V. Rao, and A. Gersho, "Variable-dimension vector quantization," *IEEE Signal Processing Letters*, vol. 3, no. 7, 1996.
[42] C. Li and V. Cuperman, "Analysis-by-synthesis multimode harmonic speech coding at 4 kb/s," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2000.
[43] A. Gersho and R. M. Gray, *Vector Quantization and Signal Compression*, Springer Science & Business Media, 2012.
[44] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, N. L. Dahlgren, and V. Zue, "TIMIT Acoustic-Phonetic Continuous Speech Corpus, Linguistic Data Consortium," 1993.
[45] "HINT speech database (in french)," Collège National d'Audioprothèse. CD for vocal audiometry.
[46] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, "Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2001.
[47] International Telecommunication Union, "Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs," Recommendation P.862, ITU-T, 2001.
[48] L. Girin, "Long-term quantization of speech LSF parameters," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2007.
[49] International Telecommunication Union, "Methods for subjective determination of transmission quality," Recommendation P.800, ITU-T, 1996.
[50] S. Djaziri-Larbi, F. Ben Ali, and M. Jaidane, "Stationarity assumption and frame segmentation in objective quality evaluation systems: A language dependency," in *International Audio Engineering Society Conference on Sound Quality Evaluation*, 2010.
[51] F. Ben Ali, S. Djaziri Larbi, M. Jaidane, and K. Ridane, "Experimental mappings and validation of the dependence on the language of objective speech quality scores in actual GSM network conditions," in *European Signal Processing Conference (EUSIPCO)*, 2009.
[52] Zhenyu Cai, N. Kitawaki, T. Yamada, and S. Makino, "Comparison of MOS evaluation characteristics for chinese, japanese, and english in IP telephony," in *International Universal Communication Symposium (IUCS)*, 2010.
[53] M. Nilsson, S. D. Soli, and J. A. Sullivan, "Development of the hearing in noise test for the measurement of speech reception thresholds in quiet and in noise," *The Journal of the Acoustical Society of America*, vol. 95, no. 2, 1994.
[54] P. Loizou, *Speech enhancement: theory and practice*, CRC press, 2013.
[55] 3GPP TR26.952, "Codec for Enhanced Voice Services (EVS); performance characterization," Tech. Rep., 3GPP, 2015.
[56] M. Nishiguchi, A. Inoue, Y. Maeda, and J. Matsumoto, "Parametric speech coding-HVXC at 2.0-4.0 kbps," in *IEEE Workshop on Speech Coding Proceedings*, 1999.
[57] H. Purnhagen and N. Meine, "HILN-the MPEG-4 parametric audio coding tools," in *IEEE International Symposium on Circuits and Systems*, 2000.
[58] F. Ben Ali and S. Djaziri-Larbi, "Perceptual long-term harmonic plus noise modeling for speech data compression," in *IEEE Global Conference on Signal and Information Processing (GlobalSIP)*, 2015.

LIST OF FIGURES