Proceedings of 20th International Congress on Acoustics, ICA 2010

23–27 August 2010, Sydney, Australia

Long-term modelling of parameters trajectories for the harmonic plus noise model of speech signals

Faten Ben Ali (1,2), Laurent Girin (2) and Sonia Djaziri Larbi (1)

(1) Unité de recherche Signaux et Systèmes, Ecole Nationale d'Ingénieurs de Tunis, Tunisia
(2) Grenoble Lab. of Images, Speech, Signal, and Automation, Grenoble, France

PACS: 43.72.*Ar* Speech analysis and analysis techniques; parametric representation of speech, 43.60.*Uv* Model-based signal processing

ABSTRACT

The harmonic plus noise model (HNM) is widely used for spectral modelling of sounds that combine harmonic and noise components, like speech signals and signals produced by a series of musical instruments. A simplified and efficient version of the HNM, developed by Stylianou et al., splits the frequency band of the signal into two bands: a harmonic part for low frequencies and a noise-like part for high frequencies, separated by a time-varying cut-off frequency. In this study, we propose to model the time trajectories of the parameters of this HNM model for non-stationary signals, especially focusing on speech signals. This is done for time intervals up to several hundreds of milliseconds, thus significantly longer than usual short-term time frames used in analysis/synthesis models and in speech coders. The goal is to capture and exploit the long-term correlation of spectral components, as can appear across spectral parameters extracted from consecutive short-term frames. Previous works by Firouzmand et al. dealt with long-term parametric modelling in the more general framework of the sinusoidal model (i.e. long-term modelling of amplitude and phase parameters). We propose to extend this work to the HNM framework in order to obtain a complete long-term HNM model. In this latter case, the parameters to be modelled on the long-term basis are the spectral envelope (that encompasses the harmonic and noise regions), the fundamental frequency (which characterizes the harmonic region) and the cut-off frequency (which separates the harmonic and noise bands). To do this, the speech signal is first segmented into voiced (actually mixed voiced/unvoiced) sections and unvoiced sections, and a discrete cosine model is used for representing the time-trajectory of HNM parameters over each entire section. The proposed long-term HNM model can be used for music and speech analysis/synthesis. It enables joint compact representation of signals (thus a promising potential for low bit-rate coding) and easy signal manipulation directly from the long-term parameters (e.g. time stretching by direct interpolation). We present several experimentations to prove the efficiency of this model. For instance, the proposed long-term HNM is compared to the short-term version in terms of listening quality and data rate.

INTRODUCTION

Two main parametric modelling techniques for speech signals have been used for years with great success. The first one is the classical LP[1] technique [1], which assumes that the speech signal is the result of a linear locally time-invariant filtering process between an excitation signal and the vocal tract filter. This LP model has been applied successfully to speech coding, hence the LPC[2] family of vocoders. The second model, also widely used, is the sinusoidal model, which represents speech signals by a sum of sinusoids [2, 3]. Those sinusoids can be harmonics of a fundamental frequency, leading to the harmonic model, that best models monophonic signals such as speech signals (single speaker). Alternately, they can remain generalized sinusoids to model polyphonic sounds. The sinusoidal model has been rather applied to speech transformations such as time stretching. Both LP and sinusoidal/harmonic models can be combined for further transformations involving separate modification (or on the contrary preservation) of the spectral envelope, such as frequency stretching / transposition.

A particular model based on the sinusoidal/harmonic model is the harmonic plus noise model (HNM) [4], which splits the frequency band into voiced and unvoiced sub-bands. Voiced sub-bands are modelled by harmonic components, whereas unvoiced bands are modelled by (coloured) noise. This model is dedicated to represent sounds with a mixed harmonic/noise structure, such as mixed voiced/unvoiced sounds of speech. A simplified version of the HNM model was developed in [5] and further works by Stylianou and collegues. This simplified version splits the frequency band into two sub-bands: a harmonic band in the low frequency region, and a noise band in the high frequency region (random components with spectral coloration though no clear temporal structure). Those two bands are separated by the voicing cut-off (VCO) frequency denoted F_V. According to this model, the speech signal can locally (i.e. on a limited time frame) be written as:

$$s(t) = \sum_{i=1}^{I} A_i cos[\phi_i(t)] + \nu(t), \tag{1}$$

where t denotes the discrete time index, i denotes the harmonic rank, ν is the high-frequency noise part of the signal, and ϕ_i is the instantaneous phase for each harmonic i given by (2)

$$\phi_i(t) = 2\pi i F_0 t + \varphi_i. \tag{2}$$

where F_0 is the fundamental frequency, A is the amplitude and φ_i the phase at the origin of harmonic i.

Speech signals are non-stationary. Therefore, in speech analysis/synthesis systems and speech coders based on parametric

[1] linear prediction
[2] linear prediction coding

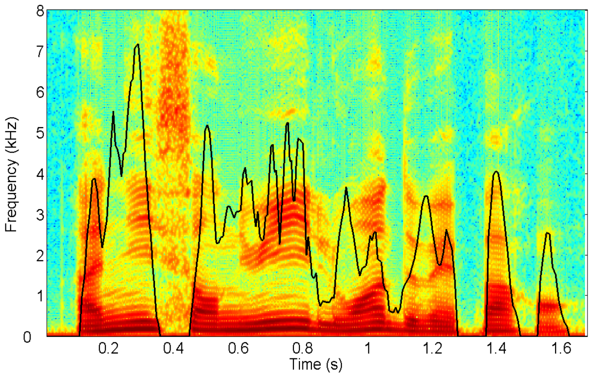


Figure 1: Voicing cut-off frequency variation of a speech signal containing voiced and unvoiced sections. The analysis of F_V is provided by the method presented in [15]. It can be seen that, globally, the spectrogram has a harmonic structure below F_V, and a rather random/noisy structure above F_V.

modelling, the speech parameters are generally processed on a short-term frame-by-frame basis, i.e. on successive analysis and synthesis frames that are approximately 20ms-long, to follow the speech time-dynamics. Since the evolution of the vocal tract is quite smooth and regular for many speech sequences, high correlation between successive parameters can be underlined. However, this long-run correlation has been relatively poorly exploited in speech processing systems. Quite few papers have considered the modelling and exploitation of this correlation beyond two or three successive frames. Let us however mention a few examples. For linear prediction, the trajectories of ten LSF[3] parameter vectors were modelled in [6] by a two-dimension discrete cosine transform (2D-DCT), similarly to what is done in block-based image compression. In [7], a fourth-order polynomial model was used for the same task, and this approach was applied to speech coding.

Recently, Firouzmand and collegues have presented a series of works dealing with adaptive long-term (LT) modelling of the time-trajectory of sinusoidal parameters [8, 9, 10, 11]. They modelled the time-trajectory of amplitude, phase, and spectral envelope parameters on quite long sections of speech (e.g. entirely voiced sections of several hundreds ms). For each parameter trajectory, a single so-called long-term model was fitted to the data using adaptive least-square optimization. Different kinds of long-term models were tested, and a discrete cosine model (DCM) similar to the classical DCT transform was favoured, as it provided good fitting to the data with few coefficients, together with good computational properties. This work was extended to the LPC framework in [12, 13] with the adaptive LT modelling of LSF vectors trajectories, and its application to LSF coding.

This paper is a further extension of this series of works. We propose to apply the long-term modelling approach to the 2-band HNM framework, i.e. we propose a long-term modelling of the time trajectories of the HNM model parameters. Those parameters are here the spectral envelope (that ensures intrinsic modelling of amplitude parameters), the fundamental frequency F_0, and the voicing cut-off frequency (an example of the time evolution of the voicing cut-off frequency is given in the spectrogram of Fig.1). As in the previous works mentioned above, the long-term frame boundaries are the voiced/unvoiced boundaries (or more exactly the limits between mixed voiced/unvoiced sections and completely unvoiced sections). For the fundamental frequency and the VCO frequency, we use the DCM. Of course, those latter parameters are defined and thus long-term modelled

[3] line spectral frequency

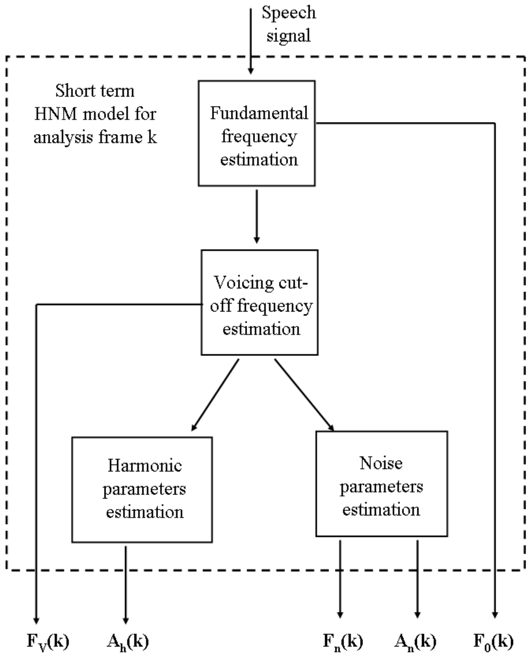


Figure 2: Analysis of the short-term HNM

only for the voiced sections. For spectral envelope modelling, we use a 2D-DCM model (2D is for two-dimension in frequency and in time) similar to the one that was used in [11] for the modelling of purely harmonic sections. However, we adapt this model along the frequency dimension to the coexistence of the two bands of the HNM model.

Note that the long-term models are not "directly" fitted to the speech signal at each sample, they rather are fitted to the short-term spectral parameters extracted for each analysis frame. Therefore, we first carry out short-term analysis of the parameters of interest, as described in the next section. Then we apply long-term modelling to those parameters, as described in the following section. We finally describe the synthesis of output signals from the LT modelled parameters. We conclude this paper with preliminary results that demonstrate the feasibility of the proposed approach, i.e. speech signal can be represented with a quite limited number of parameters using the proposed LT model, while preserving good quality.

ANALYSIS OF THE SHORT TERM HARMONIC PLUS NOISE MODEL

In this section, we describe the short-term analysis of the HNM parameters to be long-term modelled. The main steps of analysis are shown in Fig.2. They are detailed in the next subsections. We first estimate the fundamental frequency F_0, which is then used to compute the VCO frequency. Once the two bands of the model are separated, each one is analysed separately: harmonic analysis is carried out for the low frequency band, and noise analysis is carried out for the high frequency band.

Estimation of F_0 and short-term frame boundaries

The Praat software[4] is used to estimate the fundamental frequency, with the autocorrelation method described in [14]. We actually use the related pitch-mark analysis, i.e. the extraction of markers between successive periods of signal, and the fundamental frequency is provided by inversing the pitch period. This enables to provide further pitch-synchronous analysis, i.e.

[4] http://www.fon.hum.uva.nl/praat/

a set of HNM parameters for each pitch-mark. In the following, k denotes the short-term frame index, t_k denotes the pitch-mark value, which is the center of the analysis frame, and K denotes the number of short-term analysis frames (the number of periods) within a long-term section of speech signal. This F_0 estimation routine provides a data vector:

$$\mathbf{F_0} = [F_0(1), \cdots, F_0(K)]. \tag{3}$$

As detailed below, the size of the short-term analysis frames are different according to the type of further parameters under analysis.

Estimation of the voicing cut-off frequency $\mathbf{F_V}$

The method described in [15] is used to estimate the VCO frequency F_V. This method is based on the computation of a normalized spectrum for each frame and the maximisation of a cumulative energy divided into a cumulative periodic energy for the first band and a cumulative aperiodic energy for the second band. For this method, the size of the analysis frame must be equal to two pitch periods [15]. The VCO frequency is considered as the last harmonic in the frame signal. If I_k denotes the number of harmonics in the current analysis frame, then we have: $F_V(k) = I_k F_0(k)$. We denote by $\mathbf{F_V}$ the vector of measured F_V values:

$$\mathbf{F_V} = [F_V(1), \cdots, F_V(K)]. \tag{4}$$

Estimation of the harmonic parameters

The harmonic analysis consists in estimating the amplitude $A_h(i,k)$ of the i^{th} harmonic and the k^{th} analysis frame, for $i = 1$ to I_k. For each analysis frame k, we thus obtain a harmonic amplitude vector:

$$\mathbf{A_h}(k) = [A_h(1,k), \cdots, A_h(I_k,k)]^T, \tag{5}$$

where ^{T} denotes vector transposition.

The estimation is based on the iterative analysis-by-synthesis technique described in [3]. When the fundamental frequency is previously estimated, as is the case here, this process is equivalent to a basic least-square fitting between the harmonic model and the signal within a given frame. In order to obtain quite smooth amplitude parameters from one analysis frame to the next one, we also use here two successive pitch periods around the computing time index t_k as short-term frame.

Estimation of noise parameters

This analysis consists of estimating the frequencies and amplitudes of the spectral components of the upper noise-like band of the k^{th} analysis frame. This basically consists of detecting the main peaks of this frequency region. For this, we use a basic peak-picking algorithm, applied on FFT magnitude spectrum, similar to the one described in [2]. Here, the size of the analysis frame is fixed to 32ms (i.e. 512 points for 16KHz signals, hence classical short-term analysis frame) but the frame is still centered around time index t_k. Let us denote N_k the number of noise peaks in the spectrum. $\mathbf{F_n}(k)$ is the vector containing the frequencies of the detected peaks, and $\mathbf{A_n}(k)$ is the vector of corresponding amplitudes:

$$\mathbf{F_n}(k) = [F_n(1,k), \cdots, F_n(N_k,k)]^T, \tag{6}$$

$$\mathbf{A_n}(k) = [A_n(1,k), \cdots, A_n(N_k,k)]^T. \tag{7}$$

LONG TERM MODELLING OF THE HNM PARAMETERS

Once the short-term HNM parameters are estimated for the successive short-term frames of a long section of speech, their trajectory can be modelled on a long-term basis by applying an appropriate long-term model. The long-term modelling technique is first presented in a very general manner: we define the long-term model and we present the basic technique for calculating its coefficients from data. Then we present the application of this model to the HNM parameters. In this section, we assume that the speech signal has been first segmented into voiced and invoiced sections (using F_0 values), and the presented long-term modelling can be applied separately on each resulting section. The size of the data K depends on the length and characteristics (F_0 trajectory) of the section, but this dependence is omitted in the notations for simplicity.

The long-term model: DCM

The long-term model that is used in this study for each long-term frame of speech signal is the Discrete Cosine Model (DCM) that has been used in previous long-term studies within the sinusoidal and LPC frameworks [8, 10, 9, 11, 12, 13]. This model is defined as a linear combination of cosine functions (8):

$$\tilde{X}(t) = \sum_{p=0}^{P} c_p \cos(p\pi \frac{t}{N}). \tag{8}$$

where X denotes a general set of data to be modelled, and $\tilde{X}$ denotes the corresponding modelled data. $\mathbf{C} = [c_0\ c_1\ \cdots\ c_P]$ are the P+1 coefficients vector of the DCM model, and P is called the model order. The data index t runs arbitrary from 0 to N within the modelled data frame (hence N is the maximum value of t within the modelled data frame). In the present case of long-term trajectory modelling, t is a time index but it can represent any arbitrary physical quantity in other applications. For example, the DCM was used in [16] and [17] to model the short-term (log-scale) spectral envelope of speech/music signals, leading to cepstral coefficients. In those studies, the index corresponded to normalized frequency values, and N corresponded to the Nyquist frequency. We will use this frequency-modelling version in the 2D (two-dimension) version of the DCM model for amplitude modelling.

In [18], the DCM model was compared with a polynomial model and with a mixed cosine-sine model, within the sinusoidal modelling framework. Overall, the results were quite close, but the use of the polynomial model possibly led to numerical problems when the size of the modelled trajectory was large. Therefore, in the present study we consider only the DCM. Note finally that this model (or potential variants of it) is closely related to the discrete cosine transform (DCT) used for signal compression. Thus, this model has the ability to concentrate the most important part of information contained in the data set (say, its "global shape") into a limited number of coefficients. In other words, the goal of such modelling is to reduce the data dimension from K to $P+1$, with P significantly lower than K, while preserving data trajectory. It is interesting to note that, although the HNM parameters are initially defined frame-wise, the model provides a modelled value for each time index t. This property is expected to be very useful for straightforward synthesis of the modelled speech signal from modelled parameters, and also for potential transformations of this signal, as it provides a direct and simple way to proceed time interpolation for time-stretching/compression of speech: Interpolated HNM parameters can be calculated using (8) at any arbitrary instant, while the general shape of the parameter trajectory is preserved.

Estimation of the long-term model coefficients

Let us now consider the calculation of the vector of model coefficients $\mathbf{C}$, given that the order P is known (we will adjust P empirically in the experiments section). For this aim, we define the following vectors/matrices. The data set to be long-term modelled (i.e. a given set of HNM parameter values extracted at instants $t_1, \cdots, t_K$) is denoted $\mathbf{X} = [X(t_1), \cdots, X(t_K)]$. The corresponding modelled vector is denoted: $\tilde{\mathbf{X}} = [\tilde{X}(t_1), \cdots, \tilde{X}(t_K)]$.

The $(P+1) \times K$ DCM matrix that gathers the DCM terms evaluated at instants $t_1, \cdots, t_K$ is given by:

$$\mathbf{M} = \begin{pmatrix} 1 & 1 & \cdots & 1 \\ \cos(\pi \frac{t_1}{N}) & \cos(\pi \frac{t_2}{N}) & \cdots & \cos(\pi \frac{t_K}{N}) \\ \cos(2\pi \frac{t_1}{N}) & \cos(2\pi \frac{t_2}{N}) & \cdots & \cos(2\pi \frac{t_K}{N}) \\ \vdots & \vdots & \vdots & \vdots \\ \cos(P\pi \frac{t_1}{N}) & \cos(P\pi \frac{t_2}{N}) & \cdots & \cos(P\pi \frac{t_K}{N}) \end{pmatrix} \quad (9)$$

So that we have:

$$\tilde{\mathbf{X}} = \mathbf{C}\mathbf{M}. \quad (10)$$

$\mathbf{C}$ is estimated by minimizing the mean square error (MSE) between the modelled and original data. Since the modeling process aims at providing data dimension reduction for efficient signal representation, we assume that $P+1 < K$, and the optimal coefficient matrix is classically given by:

$$\mathbf{C} = \mathbf{X}\mathbf{M}^T((\mathbf{M}\mathbf{M}^T))^{-1}. \quad (11)$$

A weight matrix $\mathbf{W}$ (with a weight vector on its diagonal and zero elsewhere) can be introduced in this process to give more importance to some points in the computation of the model. In other words, such weighting constrains the model to be more accurate around specific points/regions of the data where they are granted large weight values. Finally note that in practice, we used the "regularized" version (13) of (11) proposed in [17]. Here, a diagonal "penalizing" term is added to the inverted matrix in (11) to fix possible ill-conditioning problems:

$$\mathbf{P} = \begin{pmatrix} 1 & 0 & & \cdots & & 0 \\ 0 & 8\pi^2 & & & & \\ & & \ddots & & & \vdots \\ \vdots & & & 8\pi^2 p^2 & & \\ & & & & \ddots & 0 \\ 0 & & \cdots & & 0 & 8\pi^2 P^2 \end{pmatrix} \quad (12)$$

and finally, (11) becomes:

$$\mathbf{C} = \mathbf{X}\mathbf{W}\mathbf{M}^T(\mathbf{M}\mathbf{W}\mathbf{M}^T + \lambda\mathbf{P})^{-1}, \quad (13)$$

where λ is a regularization term fixed empirically to a small value.

Application to the HNM parameters

The general guidelines of the application of the LT modelling presented above to HNM parameters is given in Fig.3. Let us remind that this modelling is applied independently for each successive voiced or unvoiced section of speech. The dependence of k on the section is omitted for simplicity. The fundamental frequency and the voicing cut-off frequency are modelled separately, and they are only modelled for voiced sections (for unvoiced sections, these parameters are set to zero for the whole section). The spectral envelope is modelled for both voiced and unvoiced sections.

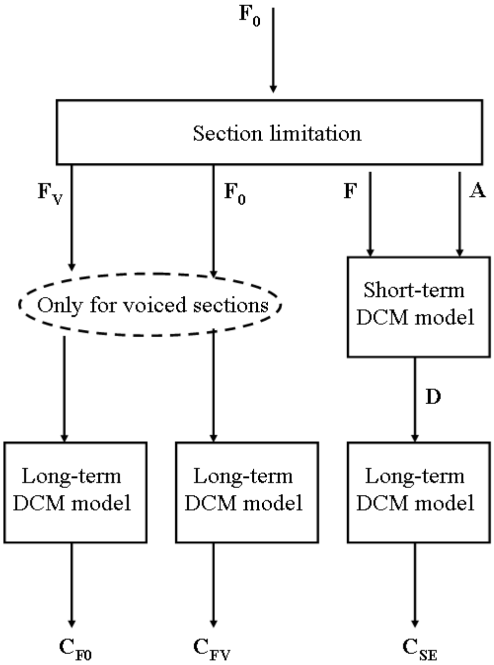


Figure 3: Complete long-term DCM modelling schema

LT modelling of $\mathbf{F_0}$

LT modelling of F_0 consists of applying equations (8) to (11) to the F_0 data vector (3). The resulting coefficients vector is denoted $\mathbf{C_{F0}}$. Note that when applying the DCM along the time axis (for F_0 and F_V), we do not use the weighting and regularizing terms. Indeed, the temporal trajectory of the fundamental frequency is generally quite smooth. Thus, the inverse matrix in (11) is generally well conditioned. Also, all time frames are assumed to have the same importance, hence no weight matrix W is necessary along the time axis.

In the experiments presented in this paper, the DCM order P is chosen empirically, respecting the two constraints of significant data compression ($P_{F_0} << K$) and good modelling accuracy.

LT modelling of $\mathbf{F_V}$

LT modelling of F_V consists of applying equations (8) to (11) to the VCO frequency data vector $\mathbf{F_V}$ (4). The resulting coefficients vector is denoted $\mathbf{C_{FV}}$. As for the fundamental frequency case, all frames are assumed to have the same contribution to the model, and the time-trajectories are "sufficiently" smooth so that the basic version of the DCM is used. The model order is also set empirically in order to provide high data compression and good modelling quality. Note that the VCO frequency F_V must be a multiple of the fundamental frequency. This constraint is not a relevant issue when modelling the F_V time trajectory, i.e. when calculating the model coefficients, but it will be considered at the synthesis stage, when exploiting the modelled trajectories.

2D-DCM modelling of the spectral amplitudes

The 2D-DCM modelling of amplitude parameters is similar to the one proposed in [11]. It is applied in two steps: a first DCM model is applied on the amplitude vector of each short-term frame in the frequency dimension[5], and then a second DCM is applied on the resulting coefficients along the time dimension. However, in [11] the spectral envelope was modelled only for voiced sections of speech where the frequency band was as-

[5] we remind that this is similar to the spectral envelope modelling of [17].

sumed to be totally harmonic, so that only harmonic amplitudes were (short-term) analyzed and 2D-modelled. In the present HNM framework, we analyse and model both harmonic and noise spectral amplitudes. Nevertheless, in the present study, we choose not to separate the two bands, and we apply a global spectral envelope model for the whole band. In other words, we use a single DCM to jointly model harmonic and noise-like components amplitudes. So, a preliminary step is to concatenate $\mathbf{A_h}(k)$ and $\mathbf{A_n}(k)$ values to have a single amplitude vector. We obtain for each short-term frame k a global vector $\mathbf{A}(k)$ containing all harmonic and noise amplitudes: $\mathbf{A}(k) = [\mathbf{A_h}(k)^T \ \mathbf{A_n}(k)^T]^T$. The model matrix (9) is here evaluated at frequency positions corresponding to the harmonics location and noise peaks location:

$$\mathbf{M} = \begin{pmatrix} 1 & \cdots & 1 & 1 & \cdots & 1 \\ \cos(\pi \frac{F_0(k)}{0.5}) & \cdots & \cos(\pi \frac{I_k F_0(k)}{0.5}) & \cos(\pi \frac{F_n(1,k)}{0.5}) & \cdots & \cos(\pi \frac{F_n(N_k,k)}{0.5}) \\ \vdots & & \vdots & \vdots & & \vdots \\ \cos(P_1 \pi \frac{F_0(k)}{0.5}) & \cdots & \cos(P_1 \pi \frac{I_k F_0(k)}{0.5}) & \cos(P_1 \pi \frac{F_n(1,k)}{0.5}) & \cdots & \cos(P_1 \pi \frac{F_n(N_k,k)}{0.5}) \end{pmatrix} \quad (14)$$

(Note that 0.5 represents the Nyquist frequency since (14) is defined for normalized frequencies).

The DCM coefficients for frame k, $\mathbf{D}_{\bullet,k} = [d_{0,k}, \ldots, d_{P1,k}]^T$, are obtained by applying (13) to the transposed amplitude vectors:

$$\mathbf{D}_{\bullet,k}^T = \mathbf{A}(k)^T \mathbf{W} \mathbf{M}^T (\mathbf{M} \mathbf{W} \mathbf{M}^T + \lambda \mathbf{P})^{-1}. \quad (15)$$

As mentioned before, P_1 must be lower than the total number of frequency points (harmonic and noise components) in each frame k. It is also assumed that P_1 is the same for each frame in a given long-term section (however it can be different across sections). So we set empirically P_1 to the minimum number of frequency points across all frames within the long-term section minus one:

$$P_1 = min_{k=1\mathrm{to}K}(I_k + N_k - 1). \quad (16)$$

In the present study, a weight diagonal matrix is used to give more importance (hence modelling accuracy) to the harmonic amplitudes than to the noise peaks during the short-term DCM modelling of the spectral envelope, since this was shown to ensure higher global quality for synthesized signals (the weights for harmonics are set to 10 and the weights for noise peaks are set to 1). A more rigorous criterion has to be defined and tested to assess this important point, and it will be considered more carefully in our future works. Also, it has been observed that when calculating (15) without the regularization term $\lambda\mathbf{P}$, the inverse matrix frequently happens to be ill conditioned. This results in quite important modelling errors. In [17] the ill conditioning is reported to be due to the non regular structure of the spectrum, with quite large variations in amplitude values (in contrast, in [11], the regularization was not a crucial point because of the regular structure of the fully harmonic spectrum). In the present study, we introduce the noise-like band, which contains erratic values of frequencies and amplitudes, so that the spectrum is more irregular, hence the need for regularization. The value of λ is chosen empirically in this study to ensure sufficiently smooth modelled amplitudes, while not loosing to much modelling accuracy, especially for the first harmonics.

The columns vectors $\mathbf{D}_{\bullet,k}$ of the short term DCM coefficients are concatenated in time, so that we obtain a matrix $\mathbf{D}$ containing all $d_{m,k}$ coefficients of a long-term section:

$$\mathbf{D} = [\mathbf{D}_{\bullet,1} \cdots \mathbf{D}_{\bullet,K}]. \quad (17)$$

Then, during the second step of 2D-modelling, LT modelling can be applied on each of the spectral envelope coefficients $d_{m,k}$,

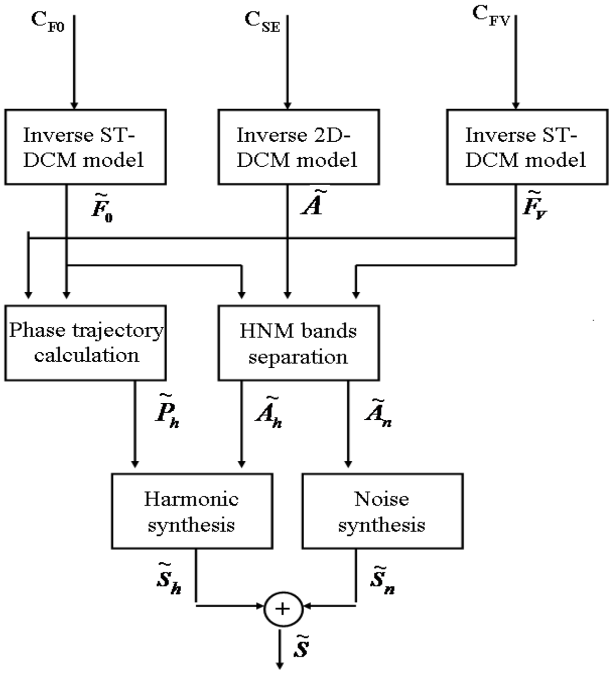


Figure 4: HNM synthesis of speech signal from the DCM coefficients

$m = 0$ to P_1, using a second DCM along the time axis (with the cosine matrix based on the time indexes). More specifically, the time-trajectory DCM model is applied separately on each row vector $\mathbf{D}_{m,\bullet}$ of the matrix $\mathbf{D}$:

$$\mathbf{D}_{m,\bullet} = [d_{m,1} \cdots d_{m,K}]. \quad (18)$$

This provides a second set of coefficient vectors denoted $\mathbf{C_{SE}}$ (SE stands for spectral envelope). Note that, since all vectors $\mathbf{D}_{m,\bullet}$ have the same size, the set of vectors $\mathbf{C_{SE}}$ can be calculated in a matrix form if the order P_2 of the time-dimension DCM is the same for all vectors $\mathbf{D}_{m,\bullet}$ (i.e. in (11) $\mathbf{X}$ can be $\mathbf{D}$ instead of $\mathbf{D}_{m,\bullet}$, and $\mathbf{C}$ would result in the concatenation of row coefficient vectors). In such case, the dimension of the LT coefficients matrix is $(P_2 + 1) \times (P_1 + 1)$.

Note that, as for F_0 and F_V modelling, we do not use weighting and regularizing terms during this modelling of $\mathbf{D}$ in the time dimension. Also, note that P_2 is also chosen empirically in the reported experiments to allow high data compression with acceptable modelling quality.

SYNTHESIS OF THE MODELLED SPEECH SIGNAL

In this section, we describe the synthesis of the modelled speech signal, i.e. generation of speech signal samples from LT modelled HNM parameters. The different steps of this synthesis process are given in Fig.4. The first step consists of obtaining a time trajectory of each parameter of the HNM model (1) from the corresponding long-term model coefficients. The second step is the synthesis of the speech signal with (1) using the modelled parameters trajectories.

F_0 and F_V trajectories from the LT model

The time trajectories of the modelled F_0 and F_V for each sample of the long-term section are simply computed from their respective DCM coefficients using (8), since (8) is actually a synthesis equation.

Harmonic frequencies trajectory

The modelled trajectories of the harmonic frequencies are then obtained by multiplying the modelled F_0 trajectory with the harmonic rank. However, attention is paid to limit the region of interpolation of the harmonic frequencies to the region below the F_V trajectory. As F_V is considered in this study as the upper harmonic frequency, the modelled F_V values at t_k are "rounded" towards the closest multiple of modelled F_0 at t_k, i.e.:

$$\tilde{I}_k = round\left[\frac{\tilde{F}_V(t_k)}{\tilde{F}_0(t_k)}\right], \tag{19}$$

$$\tilde{F}_V(k) \longleftarrow \tilde{I}_k \times \tilde{F}_0(k). \tag{20}$$

Indeed, the beginning and ending locations of the harmonics trajectory depend on the F_V shape. This is handled by including a harmonic "birth" and "death" process similar to the one used in [2] for the sinusoidal partials. The birth and death process is easy to manage since the number of modelled harmonics (or equivalently the modelled F_V value) is perfectly known at each time instant t.

Harmonic phase trajectories

Once the harmonic frequencies are interpolated using the DCM coefficients, the computed instantaneous phase trajectory (i.e. the argument of synthesis cosine functions in (1) at each sample t) of each harmonic is given by cumulative sum of the corresponding modelled frequency:

$$\tilde{\phi}_i(t) = 2\pi i \sum_{n=0}^{t} \tilde{F}_0(n). \tag{21}$$

This enables to directly exploit the interpolative nature of the LT model, and ensures phase continuity. If a harmonic is split into several segments within the same long-term section because of the variations of the F_V trajectory, several birth and death processes can occur, and the phase update is only activated during "living" portions of the harmonic.

Amplitude trajectories extraction from 2D-DCM

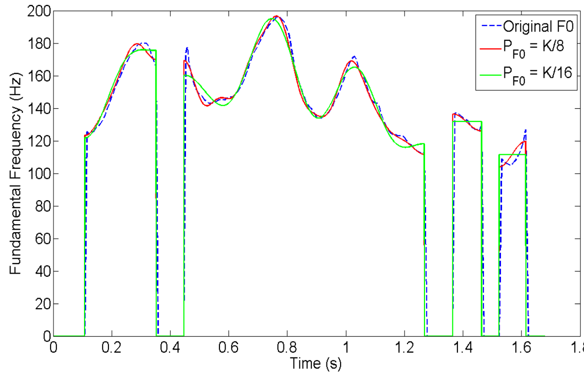
Amplitude (short-term) vectors are "decoded" from the 2D-DCM model by applying equation (10) two times in a cascaded manner, first along the time axis, then along the frequency axis. Applying (10) first using $\mathbf{C_{SE}}$ (and the appropriate cosine matrix) leads to the modelled matrix of spectral coefficients $\tilde{\mathbf{D}}$. Then applying (10) a second time using each transposed column vector $\tilde{\mathbf{D}}^T_{\bullet,k}$ of $\tilde{\mathbf{D}}$ (with the appropriate cosine matrix) leads to modelled amplitude vectors $\tilde{\mathbf{A}}(k)$.

It is important to note that the synthesis time indices can be different from the analysis time indices t_k used in the previous sections. In other words, synthesis can be carried out at arbitrary instants, e.g. using fixed windows/hop size. This enables to avoid the transmission of those indexes between analysis and synthesis for data-rate saving (this also enables interpolation facilities for speech transformations such as time-stretching for example). Nevertheless, the same notation t_k is used for the synthesis instants for simplicity.
During the second step, we resample the spectral amplitude vectors $\tilde{\mathbf{A}}(k)$ from the modelled spectral envelope. In the harmonic band, we resample at $\tilde{F}_0(k)$ and its harmonics. The obtained modelled amplitude vector in frame k is:

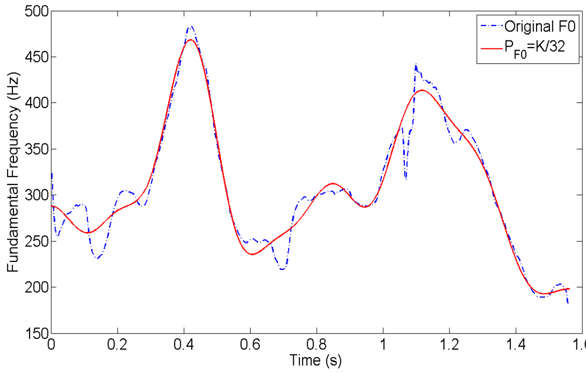
$$\tilde{\mathbf{A}}_\mathbf{h}(k) = [\tilde{A}(1,k),...,\tilde{A}(\tilde{I}_k,k)]^T. \tag{22}$$

In the same synthesis frame, the spectral amplitude in the noise-like band, is regularly sampled from the spectral enveloppe with a fixed frequency step dF to obtain the modelled noise amplitude vector.

$$\tilde{\mathbf{A}}_\mathbf{n}(k) = [\tilde{A}(\tilde{I}_k+1,k)...\tilde{A}(\tilde{I}_k+\tilde{N}_k+1,k)]^T, \tag{23}$$



(a) mixed voiced and unvoiced speech section (signal test: *sig*2).



(b) entirely voiced speech section (speech signal *sig*1).

Figure 5: Time trajectories of original (dashed line) and long-term modelled (solid line) fundamental frequency with different model orders.

where $\tilde{N}_k$ is the number of synthesis noise peaks.

Speech signal synthesis

The modelled harmonic amplitudes $\tilde{\mathbf{A}}_\mathbf{h}(k)$ are linearly interpolated in time from (synthesis) frame to frame in order to have harmonic amplitude trajectories $\tilde{A}_h(i,t)$ defined at each time sample t over the voiced long-term section (i.e. $t = 1$ to N). At the time boundaries of each living harmonic, the amplitudes are extrapolated to zero to manage the "birth" and "death" processes.
Thus, the synthesized harmonic part of the signal is given by:

$$\tilde{s}_h(t) = \sum_{i=1}^{\tilde{I}(t)} \tilde{A}_h(i,t)\cos(\tilde{\phi}_i(t)). \tag{24}$$

Note that the number of harmonics $\tilde{I}(t)$ is variable because of the harmonics birth and death processes.

The modelled amplitudes in the noise-like band are not directly time interpolated. The synthesized noise signal is obtained by an overlap-add technique applied to sinusoids with random phase (uniformly distributed in [0 2π]) as described in [3, 19]. For each frame k, sinusoids of amplitude $\tilde{\mathbf{A}}_\mathbf{n}(k)$, normalized frequencies $\tilde{\mathbf{F}}_\mathbf{n}(k)$ (regularly spaced by dF) and random phase at origin $\mathbf{\Phi_n}(k)$ are synthesized and summed to produce the modelled noise-band signal $\tilde{s}_{n,k}(t)$:

$$\tilde{s}_{n,k}(t) = \sum_{m=1}^{N_k} \tilde{A}(\tilde{I}_k+m,k)\cos(2\pi\tilde{F}_n(m,k)t + \Phi_n(m,k)). \tag{25}$$

Synthesized noise signals for successive frames are then summed using the overlap-add technique [3, 19] (using here a Hanning window), to provide the complete synthesized noise signal $\tilde{s}_n(t)$.

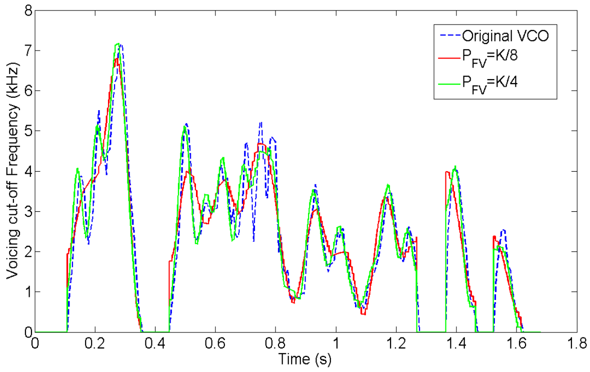


Figure 6: Trajectories of the original (dashed line) and the long-term modelled VCO frequencies of *sig*2 speech signal. P_{F_0} is set to $K/16$.

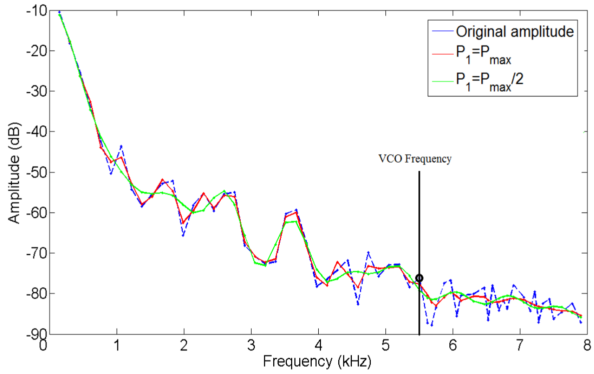


Figure 7: Original (dashed line) and HNM modelled (solid line) spectral envelopes for two model order values, within a voiced/unvoiced frame. $I_k = 36$, $F_0 = 153$Hz. P_{max} is is long-term DCM order given by (16).

The final synthesized signal is the sum of harmonic and noise parts:

$$\tilde{s}(t) = \tilde{s}_h(t) + \tilde{s}_n(t). \tag{26}$$

Note that unvoiced speech sections are processed by simply setting the VCO frequency and harmonic signal to zero, and activating only the noise synthesis (conversely, some voiced sections may be totally harmonic, i.e. the VCO frequency is equal to the Nyquist frequency, and only harmonic analysis/synthesis is performed, but this case is very rare for 16kHz speech signals).

EXPERIMENTATION AND RESULTS

As the proposed long-term HNM is a "compact" version of the short-term HNM, we evaluate in this section its performance in terms of data reduction and listening quality. This is done by comparing the long-term modelled HNM parameters with the short-term ones. We evaluate the listening quality using the objective quality assessment system PESQ[6] and informal subjective listening tests.

For the experimental procedure, we used two sets of 16 kHz speech signals: the first set, denoted *sig*1, is composed of entirely voiced signals produced by a female speaker. The second set, denoted *sig*2, is composed of mixed voiced/unvoiced signals, produced by male and female speakers. Each speech sequence is about 1.5 sec long.

Accuracy of the long-term DCM modelling

In this section, we compare the long-term modelled HNM parameters to those obtained by the short-term HNM analysis stage.

Long-term DCM modelling of F_0

First, we compare the two time-trajectories (original and modelled) of the fundamental frequency. The accuracy of the long-term model depends on the model order P_{F_0} which is here chosen empirically. On Fig. 5(a), P_{F_0} was set to $\lceil K/8 \rceil$ and $\lceil K/16 \rceil$), where $\lceil \cdot \rceil$ denotes the integer part, and we remind that K is the length of the F_0 data vector. We can see that, in both cases, the trajectory of the modelled fundamental frequency $\tilde{F}_0$ fits well the one of the original F_0 in the long sections, e.g. the two first voiced sections of *sig*2 in Fig. 5(a). For short sections, e.g. the two last voiced sections in Fig. 5(a) (where the corresponding data lengths are respectively $K = 13$ and $K = 10$), P_{F_0} is set to 0 when dividing K by 16. Additionnal experiments show that setting $P_{F_0} = 2$ is sufficient to ensure a good fitting of F_0 trajectories in such short sections. A crucial issue lies in the definition of a criterion for an automatic adjustment of the optimal model order depending on the characteristics of the signal section being modelled. This point, that will be considered in our future works, is confirmed by the results plotted on Fig. 5(b): Here the voiced section is very long and a good model fitting is obtained for $P_{F_0} = \lceil K/32 \rceil$.

[6] Perceptual Evaluation of Speech Quality - International Communication Union - Telecommunication Section - Rec. P.862

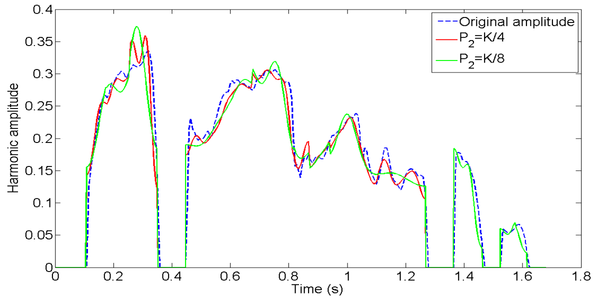
Long-term DCM modelling of F_V

In Fig.6, we can compare the trajectories of original and long-term modelled VCO frequency, F_V and $\tilde{F}_V$, for the same signals as in Fig.5. The choice of the model order P_{F_V} is also set empirically. However, the choice of P_{F_V} seems to be more tricky than for F_0, since the time trajectory of F_V has a more fluctuations than the F_0 trajectory. Thus, the F_V trajectory model requires a higher order value, as shown on Fig.6. Setting $P_{F_V} = \lceil K/4 \rceil$ provides quite faithful modelled trajectories, while $P_{F_V} = \lceil K/8 \rceil$ provides smoothed trajectories. However, the second choice may be better if it can be shown to preserve speech signal quality.

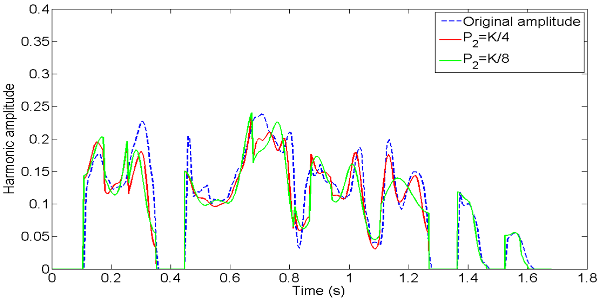
2D long-term DCM modelling of spectral amplitudes A

Fig.7 shows an example of spectral DCM modelling (i.e. first DCM along the frequency axis) for a mixed voiced/unvoiced frame, containing both harmonic and noise-like frequencies. Spectral envelopes modelled with two different model orders are plotted: model order P_1 given by (16), and half of this value. In both cases, the original spectral envelope is well modelled. The second order is used to prove that we can choose a low DCM order -and thus a higher data compression rate- with a good modelling efficiency.

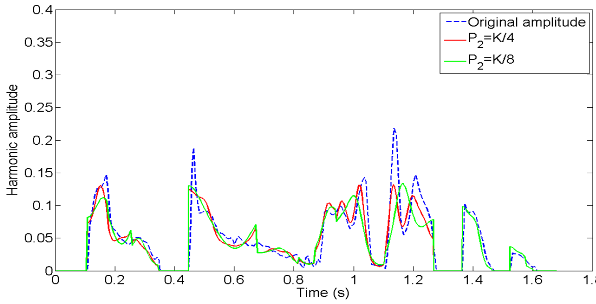
As for the 2D modelling, Fig.8 depicts the amplitude time-trajectories for the four first harmonics of *sig*2 (original trajectories obtained from the short-term analysis are also plotted for comparison). We can see that, along the time axis, a data compression rate of 4 provides a good amplitude modelling efficiency for the four plotted harmonics; In contrast, a data compression rate of 8 seems to significantly alter some amplitudes. Note also that the modelled amplitude trajectories are likely to benefit from a smoothing filtering post-processing, since modelling the trajectories of the spectral enveloppe coefficients using a "smooth" time model do not necessarily guarantee that the resulting amplitude parameters have a smooth time trajectory.



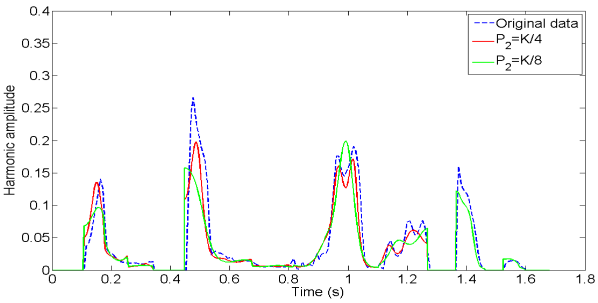
(a) First harmonic



(b) Second harmonic



(b) Third harmonic



(b) Fourth harmonic

Figure 8: Trajectories of original (dashed line) and 2D long-term modelled (solid line) amplitudes of the four first harmonics of *sig*2.

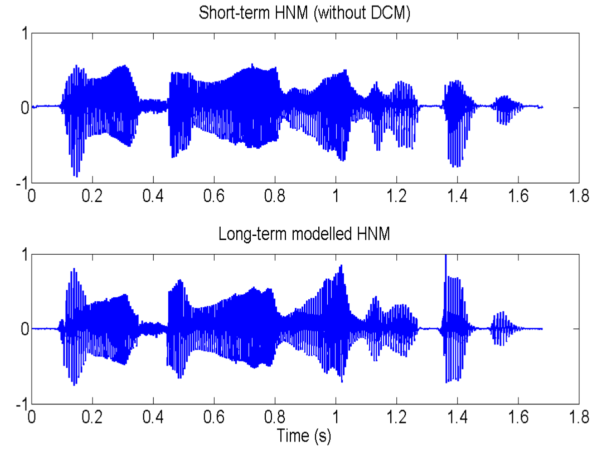


Figure 9: Speech waveforms synthesized from short-term HNM (top figure) and from long-term HNM (bottom).

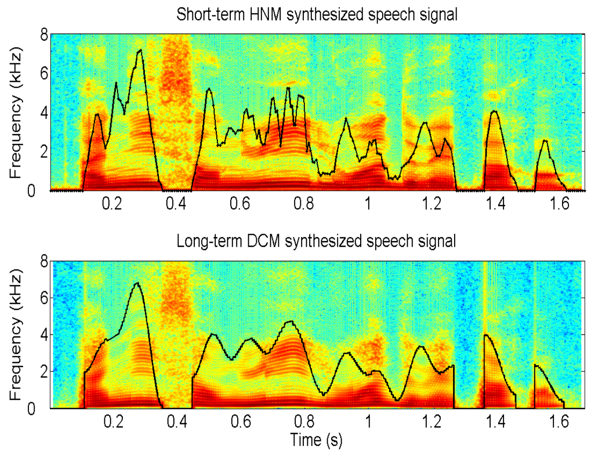


Figure 10: Spectrograms of synthesized speech. Top: short-term HNM. Bottom: long-term 2D HNM.

Synthesized signal

The speech signal synthesized from the long-term modelled HNM parameters is finally compared to the speech signal synthesized directly from the short-term HNM parameters (i.e. without any DCM modelling). Both signals sounds good (this point is further discussed in the following). Fig.9 and Fig.10 show the waveforms and the spectrograms of short-term and long-term DCM synthesized speech (*sig*2). We note the similarity of spectrograms, and also a correct shape for the speech waveform. Differences in the speech waveforms are due to a large extend to the long-term phase modelling (through long-term modelling of frequency trajectories) that do not respect original phase values (i.e. phase values at the origin are not taken into account, and summation of frequencies lead to additional dephasing).

Data compression rate

To evaluate the "compression power" of the proposed long-term HNM, we compare the number of parameters to be transmitted between analysis and synthesis modules, for both short-term HNM and long-term modelled HNM. Those parameters are listed in Tab.1 (a) and (b). For the short-term HNM, and for each frame k, we transmit the analysis frame time index t_k, the fundamental frequency $F_0(k)$, the VCO frequency $F_V(k)$, $I_k + N_k$ spectral amplitudes and N_k frequencies for the noise-like part. For a long-term speech section of K frames, all those parameters are summed. For the long-term HNM, for the same

Parameters	notation	data size
time index	t_k	K
Fundamental frequency	$F_0(k)$	K
VCO frequency	$F_V(k)$	K
Spectral amplitudes	$A(i,k)$	$K \times (I_k + N_k)$
Noise frequencies	$F_n(i,k)$	$K \times N_k$

(a) case of short-term HNM analysis

Parameters	notation	size
section length	N	1
Fundamental frequency	C_{F0}	$P_{F_0}+1$
VCO frequency	C_{FV}	$P_{F_V}+1$
Spectral amplitudes	C_{SE}	$(P_1+1)(P_2+1)$

(b) case of long-term DCM modelling

Table 1: HNM parameters to be transmitted for a given long-term speech signal.

long-term section, we transmit only the length of the section N, $P_{F_0}+1$ DCM parameters for F_0, $P_{F_V}+1$ DCM parameters for F_V, and $(P_1+1)\times(P_2+1)$ parameters for representing the spectral amplitudes. In addition, we do not need to transmit analysis time indices t_k and the N_k noise-like frequencies per frame.

The importance of the obtained data compression depends on the chosen values of the DCM ordres along the frequency and the time axes. In this study, the short-term DCM ordre P_1 is fixed by (16) to the minimum value of I_k+N_k minus 1. Along the time axis, $P_{F_0} = K/16$, $P_{F_V} = K/8$, $P_2 = K/4$. Thus the sort-term analyzed parameters size is $N_1 = K(I_k+2N_k+3)$ but only $N_2 = \frac{7K}{16} + [1+\frac{K}{4}]min_k\{I_k+N_k\}+3$ is transmitted when applying the long-term DCM.

Let us provide an example. The second voiced section of $sig2$ is $821ms$ long, it has $K = 125$ analysis frame, a mean value of $I_k+N_k \simeq 63$ and a minimum value of 42, a mean value of $N_k \simeq 50$. The short-term data size transmitted for this section is $N_1 = 14500$. When applying the long-term DCM with $P1 = 41$, $P_{F_0} = 7$, $P_{F_v} = 15$, $P_2 = 31$, the data size to be transmitted is $N_2 = 1370$. The compression rate is thus larger than 10, while preserving a good signal quality. It is important to note that this compression rate is likely to be significantly increased by applying a (series of) perceptual criterion.

Listening quality evaluation

Informal listening tests were carried out to evaluate the perceived quality of the proposed long-term model, which provides a good quality of the synthesized speech. The objective listening quality tests were carried out with PESQ. The scores are listed in Tab.2.

The long-term DCM is compared to the short-term HNM (without any DCM modelling). Tab.2 shows two PESQ sores: $score1$ evaluates the long-term DCM modelled speech signal compared to the original speech signal and $score2$ evaluates the short-term HNM modelled (without any DCM) speech signal compared to the original speech signal.
We note from Tab.2 that the obtained PESQ scores are nearly 3 which, according to the ACR[7] scale, characterizes a fair quality. The higher scores values are obtained for $seq1$, ($score1 = 2.76$, $score2 = 3.54$), which is an entirely voiced section, with high fundamental frequencies. This can be due to the reduced noise-like part in short-term frames, which can not be perfectly DCM modelled.
In the last column of Tab.2, we calculate a PESQ score difference: ($\Delta_{LT} = score2 - score1$), reflecting the effect of the long-term DCM modelling stage when applied on the short-time HNM parameters. The obtained score difference is around 0.53.

[7] Absolute Category Rating - ITU-T

sequence	score1	score2	Δ_{LT}
seq1	2.76	3.54	0.78
seq2	2.21	2.77	0.56
seq3	2.19	2.94	0.75
seq4	2.69	3.07	0.38
seq5	2.53	2.76	0.23
seq6	2.56	3.12	0.56
mean value	2.50	3.03	0.53

Table 2: PESQ listening quality scores for an entirely voiced ($seq1$) and mixed voices/unvoices speech signals.

CONCLUSION AND DISCUSSION

In this study, we presented a long-term HNM, i.e. a HNM with long-term modelled parameters. The proposed model provides an important data compression in terms of number of parameters to be transmitted from HNM analyzer to HNM synthesizer: Using appropriate (but not optimal) DCM orders, a data compression rate up to about 10 can be obtained compared to the short-term HNM analysis. The proposed long term model provides a good trade-off between data compression and listening quality, as proved by informal listening tests and PESQ scores nearly 3, which indicates an acceptable listening quality according to the ACR scale. However, those performance are obtained on a specific example, and extensive assessment remains to be carried out on a large database. The long-term modelling can be largely improved by adopting perceptual criteria to better estimate the optimal DCM order. Hence, we plan to extend the adaptive approach proposed in previous works for the harmonic model [8, 9, 10, 11] to the HNM framework.

REFERENCES

[1] B.S. Atal, L. Hanauer, "Speech analysis and synthesis by linear prediction of the speech wave", Journal of the Acoustical society of America, vol. 50, pp 637-655, February 1971.

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

[3] E. Bryan George, Mark J. T. Smith, "Speech Analysis synthesis and modification using an analysis by synthesis overlap add sinusoidal model, IEEE Transactions on Acoustics, Speech and Signal Processing", vol. 5, No. 5, September 1997.

[4] D. W. Griffin, J. S. Lim, "Multiband Excitation Vocoder, IEEE Transactions on Acoustics, Speech and Signal Processing", vol. 36, No. 8, August 1988.

[5] Y. Stylianou, "Applying the harmonic plus noise model in concatenative speech synthesis, IEEE Transactions on Acoustics, Speech and Signal Processing", vol. 9, No. 1, January 2001.

[6] N. Farvardin, R. Laroia, "Efficient coding of speech LSP parameters using the discrete cosine transformation", International Conference on Acoustics, Speech and Signal Processing, vol. 1, pp. 168-171, Glasgow, UK, 1989.

[7] S. Dusan, J. Flanagan, A. Karve, M. Balaraman, "Speech compression by polynomial approximation", IEEE Transactions on Audio, Speech and Language Processing 15(2), 387-395, 2007.

[8] L. Girin, M. Firouzmand, S. Marchand, "Long term modeling of phase trajectories within the speech sinusoidal model framework", International Conference on Speech and Language Processing, Jeju, South Korea, 2004.

[9] M. Firouzmand, L. Girin, "Perceptually weighted long term modeling of sinusoidal speech amplitude trajectories", International Conference on Acoustics, Speech and Signal Processing, Philadelohia, USA, 2005.

[10] L. Girin, M. Firouzmand, S. Marchand, "Perceptual long term variable rate sinusoidal modeling of speech", IEEE Transaction on Speech and Audio Processing, 15(3), pp. 851-861, 2007.

[11] M. Firouzmand, L. Girin, "Long-Term flexible 2D cepstral modeling of speech spectral amplitude", International Conference on Acoustics, Speech, and Signal Processing, Las Vegas, Nevada, USA, 2008.

[12] L. Girin, "Long term quantization of speech LSF parameters", International Conference on Acoustics, Speech and Signal Processing, vol. 4, pp. 845-848, Honolulu, Hawaii, USA, 2007.

[13] L. Girin, "Adaptive long term coding of LSF parameters trajectories for large-delay/very-to ultra-low bit rate speech coding", Eurasip Journal of Audio, Speech and Music Processing, volume 2010, Article ID 597036.

[14] P. Boersma, "Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound", Proceedings of the Institute of Phonetic Sciences 17: 97-110. University of Amsterdam, 1993.

[15] K. Hermus, L. Girin, H. Van home, S. Irhimeh, "Estimation of the voicing cutt off fraquency contour of natural speech based on harmonic and aperiodic energies", International Conference on Acoustics, Speech and Signal Processing, Las Vegas, Nevada, USA, 2008.

[16] T. Galas, X. Rodet, "An improved cepstral method for deconvolution of source-filter systems with discrete cepstra: Application to musical sound signals", International Computer Music Conference, pp. 82-84, Glasgow, UK, 1990

[17] O. Cappé, J. Laroche and E. Moulines, "Regularized estimation of cepstrum enveloppe from discrete frequency points", IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, New Paltz, NY, USA, October 1995.

[18] M. Firouzmand, L. Girin, S. Marchand, "Comparing several models for perceptual long-term modeling of amplitude and phase trajectories of sinusoidal speech", European Conference on Speech Communication and Technology, pp. 357-360, Lisboa, Portugal, 2005

[19] M. W. Macon and M. A. Clements, "Sinusoïdal modeling and modification of unvoiced Speech", Transactions on Speech and Audio Processing, Vol. 5, No. 6, pp. 557- 560, 1997.