A WIDEBAND SPEECH CODER BASED ON HARMONIC CODING AT 16KBS

A. Amodio & G. Feng
Institut de la Communication Parlée, UPRESA 5009
INPG/ENSERG/Université Stendhal
B.P. 25, 38040 GRENOBLE CEDEX 09, FRANCE

ABSTRACT

In this paper, we present an improved harmonic coder for the wideband. In an effort to improve spectrum modelling with harmonic representation, we include in the structure, a post error spectrum modelling that is essentially useful for noisy signals or mixed spectra of high pitch voices. The bit constrained error spectrum modelling procedure coupled with an efficient quantization of the harmonic model parameters allow us to develop a low fixed bit rate coder. The detailed structure of all the analysis, synthesis and quantization steps are outlined for the developed 16kbs coder on the [50-7000 Hz] bandwidth. Informal listening tests show that the quality obtained is comparable to the ITU_T Recommendation G722 at 48 kbs.

1. INTRODUCTION

Wideband speech coding has attracted an increasing amount of interest. The use of wideband [50-7000 Hz], due to the larger bandwidth, improves speech quality such as intelligibility and naturalness and also adds the feeling of the speaker closeness. Several coders using different techniques are being developed on this bandwidth with reduced bit rates: coders with bit rate of 32kbs, 24kbs and 16kbs will soon be normalised in order to replace the G722 ITU-T standards at respectively 64kbs, 56kbs and 48kbs. The harmonic coders and MBE coders have proven to be very efficient coding structures for the telephone bandwidth, providing good quality at low bit rate[1] [2] [4] [7] [9]. However, the quality obtained for both these coders doesn’t seem to be sufficient for a wideband coder which quality constraint is reinforced [3].

We propose a 16kbs wideband speech coder based on a harmonic scheme. To overcome the harmonic coder limitations, especially when modelling female voices complex spectra, we add after the first harmonic modelling a procedure that refines the spectral model by representing the spectral error between the original spectrum and the modelled harmonic one. This spectral error modelling procedure coupled with an efficient Vector Quantization (VQ) of the harmonics magnitudes, allows us to improve quality will keeping a low fixed bit rate.

In section 2, the general structure of the coder is presented, while in section 3 we outline the detailed analysis, synthesis and quantization procedures. In section 4, we present the first evaluation of our 16 kbs coder.

2. GENERAL PRESENTATION OF THE CODER

Schematic diagrams of the coder and the decoder are respectively presented in Fig.1. and Fig.2.

The analysis frame is preliminary classified into Unvoiced one which spectrum only contains “noise like” energy and Mixed one which spectrum can be both harmonic and noise like.

It has been shown [9] that a harmonic model can be appropriate even for noise spectra provided that the fundamental frequency used is less than 100Hz. We use such a harmonic model for the Unvoiced frames: the original speech spectrum $w_S(\omega)$ of the signal $s_w(n)$ is modelled by a harmonic spectrum $w_{Sh}(\omega)$ of $M$ amplitudes $A_m$ centered on the harmonics of $F_0$. As the original noise spectrum has no harmonic structure, $F_0$ actually represents the spectral sampling frequency of $w_S(\omega)$ which is not fixed.

Mixed frame spectrum $w_S(\omega)$, which contains a harmonic structure and/or “noise like” energy is modelled by the sum

$w_S(\omega) = \sum_{m=1}^{M} A_m \sin(2\pi m F_0 t)$

In this model, $A_m$ is the amplitude of the $m$-th harmonic, $\sin(2\pi m F_0 t)$ is the $m$-th harmonic of $F_0$, and $t$ is the time variable.

In the case of the Mixed frames, the spectral error $E_w(\omega)$ is modelled by a harmonic spectrum $w_{Ew}(\omega)$ with amplitudes $A_m$ centered on the harmonics of $F_0$. The detailed structure of all the analysis, synthesis and quantization steps are outlined for the developed 16kbs coder on the [50-7000 Hz] bandwidth.

Informal listening tests show that the quality obtained is comparable to the ITU_T Recommendation G722 at 48 kbs.
\( \hat{S}_n(\omega) \) of a harmonic spectrum \( S_h(\omega) \) and of a noise spectrum \( \hat{E}_n(\omega) \).

\[
\hat{S}_n(\omega) = S_h(\omega) + \hat{E}_n(\omega)
\]

(1)

\( S_h(\omega) \) is modelled by \( M \) amplitudes \( A_m \) and the fundamental frequency \( F_0 \).

\( \hat{E}_n(\omega) \) is built from the error between the original and the harmonic spectra which is defined by

\[
E_n(\omega) = S_n(\omega) - S_h(\omega)
\]

(2)

The bandwidth is divided into sub-bands where \( \hat{E}_n(\omega) \neq 0 \) or \( \hat{E}_n(\omega) = E_n(\omega) \), \( \hat{E}_n(\omega) \) refining the spectral modelling of \( S_n(\omega) \) on sub-bands where the harmonic model is of poor representation; this can occur especially for transition regions and noisy signal of high pitch voices. The procedure of reconstruction of \( \hat{E}_n(\omega) \), will be detailed in section 3.1.2.

3. DETAILED STRUCTURE OF THE CODER

This coder is developed on the [50Hz-7000Hz] bandwidth for speech sampled at 16KHz. The analysis frame is 32 ms long and there is a 50% overlap between two consecutive analysis frames.

3.1. Speech Analysis

3.1.1. Phonetic classification of the frames

We introduce an initial bi-classification of the frames into fully Unvoiced and Mixed frames. The discrimination criteria we use are the signal energy, the zero crossing rate, the SFM (Spectral Flatness Measure), the first autocorrelation coefficient and the spectral error energy between the original spectrum and the synthesized harmonic one. Other classifications have been proposed for the telephone band [3], [4], using different discrimination criteria.

3.1.2. Mixed Frame analysis

Mixed frame spectrum can contain a harmonic structure and a noise structure within the same spectral sub-band, which can be badly modelled if only by a harmonic spectrum or a noisy spectrum. Therefore, after the harmonic spectrum modelling, a noise spectrum \( \hat{E}_n(\omega) \) can be added to \( S_h(\omega) \) to refine the harmonic modelling. A modelling of the error spectrum was mentioned in [1] and [2] in the theoretical structure of the proposed harmonic coders but no application was presented.

The FFT \( S_n(\omega) \) of the signal \( s_n(n) \) is modelled by \( \hat{S}_n(\omega) \) as in equation (1).

- \( S_h(\omega) \) modelling and computing

\( S_h(\omega) \) is a harmonic spectrum represented by \( M \) amplitudes \( A_m \) centered on the harmonics of the fundamental frequency \( F_0 \).

However, to compute the model parameters \( A_m \) and \( F_0 \), a multi-harmonic model described by \( N \) “fundamental frequencies” \( F_{0i} i \in [1, N] \) on \( N \) spectral sub-bands [3] [5], is applied in order to refine the description of the original spectrum and so compute accurate values of the \( A_m \). The procedure used to compute the parameters of the multi-harmonic model is the one described in [6] by Griffin for the MBE coder. Once the amplitudes \( A_m \) and the frequencies \( F_{0i} \) are computed, the average frequency \( F_0 \) is used at the decoder to describe the harmonic model. Only, a harmonic model is used instead of the multi-harmonic one used during the analysis because of the little loss in quality it leads compared with the gain obtained with bit rate reduction [5].

- \( \hat{E}_n(\omega) \) modelling and computing

\( \hat{E}_n(\omega) \) is modelled from \( E_n(\omega) = S_n(\omega) - S_h(\omega) \) and is meant to represent the perceptual or “audible” error between the original and the harmonic spectra and to improve the spectrum modelling in the regions where the harmonic representation failed.

First step procedure: From \( E_n(\omega) \) to \( \hat{E}_n(\omega) \).

Actually, it’s not necessary to reach the mathematical equality between the original spectrum and the synthesized one to reach perceptual equality for these two signals. Due to spectral auditory masking effect, the modelling error \( E_n(\omega) \) can be masked by \( S_h(\omega) \) and become inaudible [11]. The first step of the procedure consists in modelling \( \hat{E}_n(\omega) \) to keep only its “audible” components. The total bandwidth is divided into 14 sub-bands. On each of these sub-bands, the ratio between the energy of the harmonic spectrum and the original one is evaluated and is chosen as the indicator of the good accuracy of the harmonic model representation. The comparison of this ratio to a threshold then discriminates sub-bands into \( N_I \) “audible” ones where \( \hat{E}_n(\omega) \neq 0 \) (\( b_I = 1 \)) and “inaudible” ones where \( \hat{E}_n(\omega) = 0 \) (\( b_I = 0 \)).

On each of the \( N_I \) audible sub-bands, \( \hat{E}_n(\omega) \) is modelled from \( E_n(\omega) \) by 4 equally spaced amplitudes as we assume that \( E_n(\omega) \) only contains noise like energy.

Second step procedure: The bit rate constraint.

Unlike in classical fixed bit rate harmonic or MBE coders, where the same bit rate is used to quantize either a high or a low pitch harmonic spectrum, we assume that the lower the pitch is, the more information the harmonic model contains and so the more the bit rate has to be. Assuming this fact and as we want to develop a fixed low bit rate coder, we can not add more bit rate for the modelling of \( E_n(\omega) \) if the entire bit rate has already been used to quantize a low pitch harmonic model. The second step of our procedure consists in evaluating the bit rate left after the harmonic modelling quantization and to estimate on how many sub-bands \( N_E \) the error can be modelled.

Last step procedure:
The last step consists in comparing the number of sub-bands $N_1$ to be modelled and the number of sub-bands $N_2$ that can be modelled due to bit rate constraint. If $N_2 > N_1$, the supplementary bit rate is reallocated. Else if, the more energetic sub-bands of $\hat{E}_n(\omega)$ are quantized, the others are not transmitted. Indeed, the poor harmonic modelling leading to a high value of $N_1$ often corresponds to high pitch voices that also corresponds to high value of $N_2$ and inversely.

3.1.3. Unvoiced frame analysis

The whole spectrum of unvoiced frame is modelled by a harmonic spectrum of $M$ amplitudes $A_m$ centered on the harmonics of $F_0$. As $F_0$ only represents the fundamental frequency of the spectral sampling it can be fixed a priori as it has been proposed in [10]. We do not proceed in such a way, but compute the number $M$ of amplitudes $A_m$ and their value by applying a procedure used in sinusoidal representation [8] called “peak-picking”. This procedure consists in locating and evaluating all the peaks of energy of the spectrum; the $M$ amplitudes $A_m$ of our harmonic model corresponding in number and values to these peaks. However, we equally spaced the $M$ amplitudes $A_m$ on the bandwidth leading to a harmonic model, that is less accurate but gives a good quality for a gain in bit rate reduction.

3.2. Speech Synthesis

3.2.1. Unvoiced frame synthesis

The unvoiced frame signal is first synthesized in the frequency domain. With the parameters $F_0$ and the $M$ amplitudes $A_m$ a synthesized spectrum is reconstructed at the decoder. To construct the synthesized signal in the time domain, a reverse FFT procedure followed by an overlap add algorithm is applied.

3.2.2. Mixed frame synthesis

The synthesized signal $\hat{w}_m(n)$ for a mixed frame, is obtain by summing a voiced signal $\hat{w}_v(n)$ issued of the modelling of $S_v(\omega)$ with a harmonic model $S_h(\omega)$ and of a noisy signal $\hat{w}_{nv}(n)$ issued of the modelling of $E_n(\omega)$ by $\hat{E}_n(\omega)$.

Voiced signal synthesis

The voiced signal is constructed in the time domain from the spectrum $S_h(\omega)$ described by $F_0$ and the $M$ amplitudes $A_m$. The procedure used is the one proposed for the MBE coder by Griffin [6] [7] that consists in summing sinewaves which frequencies, phases and magnitudes correspond respectively to the harmonics of $F_0$ and to the phases and magnitudes of the amplitudes $A_m$.

Unvoiced signal synthesis

The unvoiced signal is constructed from the spectrum $\hat{E}_n(\omega)$. A reverse FFT followed by an overlap add algorithm is applied to reconstruct the signal in the time domain. This signal is then added to the voiced signal to form the final signal for mixed frame.

3.3. Quantization

The coder developed is an 16kbs coder on the [50Hz-7000Hz] bandwidth. The total global bit rate allocation is given in Table 1.

The parameters to quantize are the fundamental frequency $F_0$ and the complex amplitudes $A_m$ of the harmonic model for Unvoiced frames or Mixed frames, and the spectrum amplitudes vectors of $\hat{E}_n(\omega)$ for Mixed frames.

<table>
<thead>
<tr>
<th>Parameters</th>
<th>Mixed frame</th>
<th>Unvoiced frame</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mixed/Unvoiced</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>$F_0$</td>
<td>9</td>
<td>6</td>
</tr>
<tr>
<td>Harm. Magnitudes.</td>
<td>60 - 120</td>
<td>126 - 205</td>
</tr>
<tr>
<td>Harm. Phases</td>
<td>50 - 126</td>
<td>123 - 44</td>
</tr>
<tr>
<td>$\hat{E}_n(\omega)$</td>
<td>123 - 0</td>
<td>-</td>
</tr>
<tr>
<td>b indicators</td>
<td>13</td>
<td>0</td>
</tr>
<tr>
<td>Total bit rate</td>
<td>256</td>
<td>256</td>
</tr>
</tbody>
</table>

Table 1. Bit allocation for a 16 ms frame for a bit rate of 16kbs.

3.3.1. Fundamental frequency quantization procedure.

We use a uniform scalar quantization procedure to quantize the frequency $F_0$.

For mixed frames, this frequency corresponds to a voice pitch and so has physiological limits. On the contrary, the fundamental frequency for an Unvoiced frame corresponds to the number of peaks in the spectrum in the [50Hz-7000Hz] band, and so has statistical limits.

Therefore, for mixed frame $F_0$ is uniformly scalar quantized with 9 bits for pitch varying from 60Hz to 200Hz and for Unvoiced frames, $F_0$ is uniformly quantized with 6 bits for sampling frequency varying from 60Hz to 100Hz.

3.3.2. Amplitudes $A_m$ quantization procedure

Even if the same harmonic model is applied to noisy spectrum and mixed spectrum, the bit rate allocation procedures differ for unvoiced and mixed frames. Indeed, the properties of the $A_m$ are totally different if they model an unvoiced spectrum or a harmonic spectrum. For example, phase accuracy is more important for Voiced spectrum than for Unvoiced spectrum.

Magnitude quantization procedure : intra-frame predictive VQ.

We propose a very efficient intra-frame predictive partitioned vectoriel quantization procedure for the magnitudes of the $A_m$. 

Magnitude quantization procedure : intra-frame predictive VQ. 

We propose a very efficient intra-frame predictive partitioned vectoriel quantization procedure for the magnitudes of the $A_m$. 

As the number of harmonics can vary from one frame to another, we use partitioned VQ. Actually, the total magnitude vector which dimension can vary from ~30 (high pitch voice) to ~100 (low pitch voice) is decomposed into a variable number of sub-vectors of dimension 3 or 7. Each sub-vector is then VQ quantized.

In order to use the same dictionaries when VQ is applied on a low frequency sub-vector or on a high frequency sub-vector, the logarithmic difference between two consecutive magnitudes within the same frame are VQ quantized instead of the magnitudes.

Bit rate allocation between the different sub-vectors depends on the nature of frame (mixed or unvoiced), on \( F_0 \) and on its frequency position. The maximum bit rate associated a sub-vector is 10 bits allowing a low complexity VQ search procedure.

Phase quantization procedure: uniform scalar quantization.

The phases of the harmonics for Mixed frames are uniformly scalar quantized on 4, 3, 2, 1 bits or predicted, respectively for harmonics of low frequencies to high frequencies. The bit rate allocation depends only on the frequency \( F_0 \).

The phases of the harmonics, for unvoiced frames, are uniformly scalar quantized on 2 bits or have a random value. The bit rate allocation depends on the frequency \( F_0 \) and also on the \( M \) \( |A_n| \); only the phases of the most energetic magnitudes are quantized.

Spectral error quantization procedure

The magnitudes of the 4 amplitudes of \( \hat{E}_w(\omega) \) on a sub-band are vector quantized using 6 bits and each phase is quantized using 1 bit.

4. EVALUATION AND SUMMARY

In this paper, we have presented the development of a 16kbps improved harmonic speech coder for the wideband. This coder was developed using a new approach to model any mixed spectrum. The total bandwidth is at first modelled by a harmonic spectrum and then, the error spectrum, spectral difference between the original spectrum and the harmonic one is added on the sub-bands where the harmonic model is of poor representation. To develop a fixed bit rate coder, we take into account the fact that the accuracy of the harmonic model representation decreases when the fundamental frequency increases (female voices) whereas the necessary bit rate to quantize the harmonic spectrum parameters decreases : free bit rate is then used to refine the harmonic modelling of the original spectrum.

We also propose an efficient partitioned VQ of the harmonic magnitudes that allow us to reach a low bit rate of 16kbps while keeping a low complexity for the quantization procedure.

An informal evaluation of our 16kbps coder has shown to be promising. The tests were based on a set of 15 speech utterances (7 male utterances + 6 female utterances + 2 child utterances) and was an informal MOS-like test. 16 listeners were asked to evaluate the quality (on a MOS scale from 1 to 5) of the IUT-T Recommendation G722 at 48 kbps and of our 16kbps coder.

<table>
<thead>
<tr>
<th>Coder</th>
<th>G722/48kbs</th>
<th>16kbps coder</th>
</tr>
</thead>
<tbody>
<tr>
<td>MOS</td>
<td>3.16</td>
<td>3.0</td>
</tr>
</tbody>
</table>

Table 2. MOS results

The quality obtained with our 16 kbps coder is comparable with the G722 coder at 48kbps. Further work will concern bit rate reduction to 13 kbps since signal degradation is essentially due to the harmonic model (the voice sounds sometimes “metallic”) and not to the quantization step.

5. REFERENCES