WIDEBAND SPEECH CODING BASED ON THE MBE STRUCTURE

A. Amodio and G. Feng
Institut de la Communication Parlée, UPRESA 5009
INPG/ENSERG/Université Stendhal
B.P. 25, 38040 GRENOBLE CEDEX 09, FRANCE
Tel. +33 (0)4 76 82 41 20, FAX: +33 (0)4 76 82 43 35, E-mail: amodio@icp.grenet.fr

ABSTRACT

This paper deals with the adaptation to wideband of the MBE coder which was initially developed for the telephone band. As the constraints of quality and bit rate for a wideband and a telephone band coder are different, and as the signal characteristics on these two bands are different too, we must reconsider the coder structure. Several improvements are proposed, some of which were already proposed for the telephone band such as the phonetic classification of the frames or the multi-harmonic modelling of the spectrum. We also propose in order to reach a good quality, especially for high frequency voices, to model and synthesize, as part of the signal, the initial error between the synthetic and original spectra.

1. INTRODUCTION

Currently, wideband speech coding has attracted an increasing amount of interest. The use of wideband (50-7000Hz) allows, due to the larger bandwidth, to improve speech quality such as intelligibility and naturalness and also adds the feeling of the presence of the speaker. Several coders using different techniques have been developed for wideband with bit rates of 32 kbs at a quality equal to the 64 kbs ITU-T (ex CCITT) standard G722 (normalised in 1986 and serving as reference for this band). At the same time, the MBE (Multi-Band Excitation) speech coding structure proposed by Griffin in 1987 [1], has been shown to be efficient for the telephone band, as can be seen from the amount of literature, concerning for example bit rate reduction. In this specific context, we propose the design of a wideband coder based on the MBE structure. Our goal is to design a high quality 16 to 24 kbs coder which incorporates a psycho-acoustic model. This paper presents the first step of our work which consists of the determination of the coder's structure. This requires the analysis and improvements of the different elements constituting the MBE structure and of their suitability for wideband speech. After a brief recap of the MBE coder structure, we present the modifications, choices and improvements we have made as dictated by the analysis of the main problems encountered when adapting this coder to the wideband.

2. THE MBE CODER

The main advantage of the MBE speech coder for the telephone band is to enable a very low bit rate and low complexity to be obtain while providing good quality. This is mainly achieved thanks to a very efficient modelling of the signal spectrum.

2.1. Modelling of the speech spectrum

The Fourier transform $S_w(\omega)$ of the windowed signal $s_w(n)$ is computed and then modelled by the product of a spectral excitation $|E_w(\omega)|$ with a spectral shape $H_w(\omega)$.

$$s_w(n) = w(n)s(n) \quad (1)$$

and

$$\hat{S}_w(\omega) = H_w(\omega)|E_w(\omega)| \quad (2)$$

The bandwidth is divided into N subbands where N can be a fixed number or can vary with the signal. The spectral excitation is modelled either by a periodic spectrum for those subbands considered to be voiced or by a random noise spectrum for subbands found to be unvoiced. The spectral shape is represented by M amplitudes A_m centred on the harmonics of the signal fundamental frequency F_0.

2.2. Analysis and synthesis

Many propositions concerning complexity reduction or bit rate reduction have been made since the first presentation of a MBE coder by Griffin in 1987 [1]. We present here the initial structure and procedures of the MBE coder used on the telephone band [2].

Analysis

The M amplitudes A_m, the fundamental frequency F_0 and the N voiced/unvoiced decisions constitute the coder parameters and are computed by minimising the mean spectral quadratic error ε between the original spectrum and the synthetic spectrum.

$$\varepsilon = \frac{1}{2\pi}\int_{-\pi}^{\pi}\left|S_w(\omega) - \hat{S}_w(\omega)\right|^2 d\omega \quad (3)$$

First, ε is minimised over the frequency considering that all the subbands are voiced. The "all voiced" synthetic spectrum is then evaluated and the voiced/unvoiced decisions are made in the N subbands by comparing to a threshold the error between the original spectrum and this "all voiced" synthetic spectrum. Finally, the M amplitudes A_m are calculated to minimise ε (3).

Synthesis

The synthetic signal is composed by of voiced signal and an unvoiced signal. The voiced signal is a sum of sinewaves, each one corresponding to a spectrum line of the voiced subbands; the amplitudes and phases are determinated by the modules and phases of A_m, and the frequencies of the sinewaves are the harmonics of the fundamental frequency F_0. The spectrum of the unvoiced subbands is modelled by shaping a noise spectrum with the module of A_m, then the unvoiced signal is obtain by computing the inverse Fourier transform of this spectrum. Note that in the unvoiced subbands, only the modules of A_m are used, whereas the complex amplitudes are necessary in voiced subbands.

3. PROBLEMS FOR WIDEBAND SPEECH

The MBE coder has proved its efficiency when used on the telephone band. The modelling of the spectrum allows to obtain a very low bit rate while preserving a good quality. Unlike the CELP coder, the MBE coder applies a simplified model to the spectrum of the signal. Actually, two important hypothesis are made: the first one is that a subband is only voiced or unvoiced and the second one is that when the subband is voiced, the spectrum is only represented by spectrum lines appearing at the harmonics of the fundamental frequency. We thus speculate whether we can obtain good quality for wideband speech with a MBE coder.

3.1. The harmonic spectrum model

With the MBE model we consider that for a voiced subband the spectrum lines are centred on the harmonics of the fundamental frequency. However, the analysis of a whole voiced spectrum and of its corresponding synthetic spectrum shows that even if the spectrum lines fit well for the low frequencies, they can shift for the high frequencies and appear less neatly and so the original and synthetic spectra totally mismatch. This phenomenon was already observed in the telephone band but was judged less important. This mismatch of the lines in the high band can produce two main problems.
First, as the harmonics lines don't match the original spectrum lines, the amplitudes A_m will be mistaken which will add a significant distortion in the high band spectrum.
Secondly, as a subband is declared unvoiced when the error between the original spectrum and the supposed voiced synthetic spectrum is superior to a threshold, a voiced subband in which the lines mismatch can be declared as unvoiced. This error also lead to very annoying artefacts.

3.2. Different numbers of harmonics for different F_0

The spectrum of a windowed signal is modelled by N harmonics, the distance between two harmonics being equal to the signal fundamental frequency; we can view this modelling as a spectrum sampling procedure with a variable sampling frequency. Even with a small number of harmonics, typically when coding a female voice, a good quality can be reached on the telephone band, but can we obtain a good quality for wideband speech?
When coding a male voice, the number of harmonics is, on average, twice the number of harmonics for a female voice although there is no theoretical justification for this. Actually, as the fundamental frequency increases, pertinent information is located elsewhere rather than being centred on the harmonics; however this information is not considered by the MBE coder.

3.3. Quality of the MBE coder for wideband speech

We first adapted the MBE telephone band structure to the wideband without any improvement. The quality we obtained was extremely encouraging when coding male voices but very disappointing when coding female voices. This confirms that the last problem we mentioned is a very critical one.

4 PROPOSED STRUCTURE

The point we will now focus on is the quality, the constraints of the bit rate and complexity are first relaxed even if we aim to obtain a rate of 16 kbs to 24 kbs.

4.1. Phonetic classification of the frames

In the MBE coder, the different subbands of the spectrum are classified as voiced or unvoiced and so the signal is synthesized differently for these two sorts of subbands. When a mistake occurs in the voiced/unvoiced classification very audible artefact can be heard in the synthetic signal that are not acceptable in wideband. Moreover, when coding a whole unvoiced frame, the MBE structure is of no interest. This is why we propose to introduce an initial bi-classification of the frames into fully unvoiced or mixed frames. For the first category of frames, all the band is declared unvoiced and a fixed number of spectral magnitudes is transmitted to the decoder. For the second category of frames, the MBE analysis procedure is applied.
The discrimination criteria that we use are the signal energy, the zero crossing rate and the SFM (Spectral Flatness Measure). The constraints we fixed on these three criteria favours the classification of a frame into a mixed one where all the subbands can still be classified as unvoiced. With this procedure, a bit is transmitted to indicate if a frame is mixed or unvoiced and when an unvoiced frame occurs, all the bits are attributed to quantize the spectrum magnitudes. Other types of classification have been proposed for the telephone band [3], [4], using different discrimination criteria.

4.2. New modelling of the harmonic spectrum

The analysis of the spectrum of a whole voiced frame shows that if the harmonic structure appears clearly for the lower band of the spectrum, the spectrum lines shift in the higher band leading to the problem of false

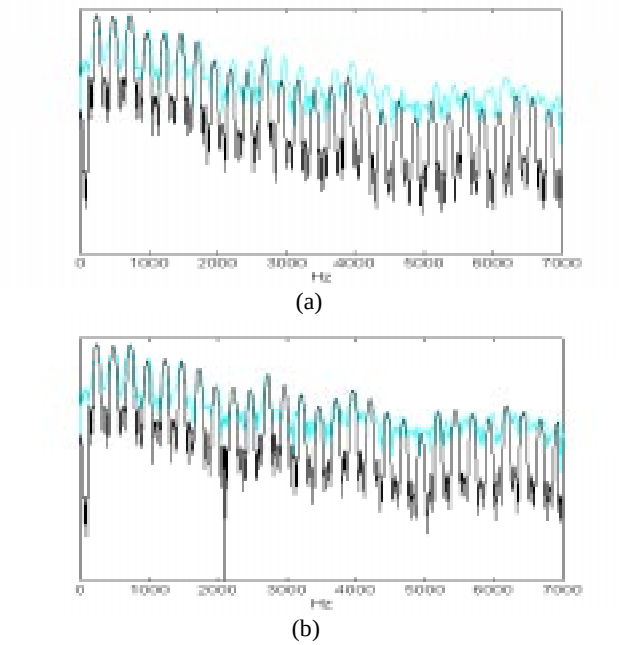
evaluation of the amplitudes A_m and false voiced/unvoiced decisions.As the lines still appear for the high frequencies of the spectrum but seem to be centred on the harmonics of a frequency ($F_0+\Delta f$), we decide to divide the spectrum into L subbands and for each subband l, we find the best frequency ($F_0+\Delta f_l$) whose harmonics match the original spectrum lines. First, the fundamental frequency F_0 is computed and then the l frequency deviations Δf_l.

4.2.1. Determination of F_0

Initially, the fundamental frequency F_0 is computed by minimising the error ε over all the frequencies. As this procedure is computationally expensive, Griffin[2] proposed to first evaluate the fundamental frequency with an autocorrelation method and then improve it by applying a "refinement" procedure for frequencies close to the initial one. We also evaluate the initial frequency F_0 with the classical method of autocorrelation.

4.2.2. Determination of the L ($F_0 + \Delta f_l$) frequencies

On the L subbands, we evaluate the best frequency ($F_0+\Delta f_l$) whose harmonics fit with the original spectrum lines. Actually, we apply L times the "refinement" procedure on the L bands of the spectrum. This new procedure leads to a very noticeable reduction of the spectral distortion especially for high frequencies of the spectrum as can be used in Fig 1.The error between the two spectrum can now be seen as a real voiced/unvoiced indicator whereas before this procedure was applied, a significant error value could be obtained due only to the harmonic spectra mismatch.



(a)

(b)

Fig. 1. (a) Original spectrum (clear line) and modelled synthetic spectrum with a single F_0. (b) Original spectrum and modelled synthetic spectrum with determination of 8 optimal frequencies in 8 subbands.

4.2.3. Evaluation

We divided the spectrum into 8 equal subbands and tested this procedure. The improvement is very audible and eliminates very annoying artefacts in both male and female speech signals. Of course, the complexity and the bit rate increase with this procedure but it has proved to be necessary.

Two important remarks can be made concerning the bit rate. First of all, the L optimal frequencies being correlated, they will not be quantized separately but with a differential procedure so that the rate will not increase a lot.

Secondly, the distortion that produces the worst audible degradation is not a frequency distortion but rather the magnitude distortion of the lines. So, we can only apply this new modelling of the spectrum during the analysis procedure to permit the correct determination of the amplitudes A_m. Then, only the fundamental frequency F_0 is transmitted to the decoder : the positions of the sinewaves are distorted but the magnitudes are correct. With this method, we also obtain a better quality without increasing the bit rate. This observation was already made in [5]: a bi-harmonic modelling of the spectrum was proposed to be used for the telephone band MBE coder.

By considering the rate to obtain we can optimize the number L of bands and decide whether or not we transmit the L optimal frequencies.

4.3. Modelling of the spectral error

The MBE modelling of the spectrum is a very efficient method for the telephone band. However, on the wideband, even with the new procedure of multi-harmonic modelling of the spectrum, it seems that this structure is not sufficient to provide the quality expected.

4.3.1. A theoretical problem

As the fundamental frequency increases, the number of harmonics transmitted to the decoder and so the quantity of information decrease. As the theoretical information has no reason to be lower in a high fundamental frequency voice we think the information must be elsewhere. Actually, if we examine the voiced spectrum of a female voice we see that it presents spectrum lines less sharp than those of a male voice. So, we can hypothesize that for a low frequency voice, the energy is only concentrated on the numerous harmonics whereas, for a high fundamental frequency voice the energy is spread around few harmonics. The initial MBE structure provides an efficient spectrum model for low fundamental frequency voices for which all the information is centred on the harmonics whereas, for high fundamental frequency voices, all the information spread around the harmonics is not taken into account. The observation of the spectral modelling error for male and female voices confirms these hypothesis (Fig. 2.).

4.3.2. Proposition

As, for high fundamental frequency voices, the modelling error between the original spectrum and the MBE modelled one is quite significant, we propose to model this spectral error and transmit it to the decoder. The synthetic signal is now obtained by summing the MBE synthetic signal and the modelling error signal.

After the MBE modelling of the spectrum, we note $E_w(\omega)$ the spectral error between the original signal and the synthesised one.

$$E_w(\omega) = S_w(\omega) - \hat{S}_w(\omega) \quad (4)$$

Since this error presents a quasi harmonic structure (Fig 2.), especially for a high fundamental frequency voices, we model this error by applying the same procedure used to model the harmonic original spectrum $S_w(\omega)$.

This error spectrum $E_w(\omega)$ is modelled by a periodic spectrum $\hat{E}_w(\omega)$ with a fundamental frequency F_1 close to F_0.

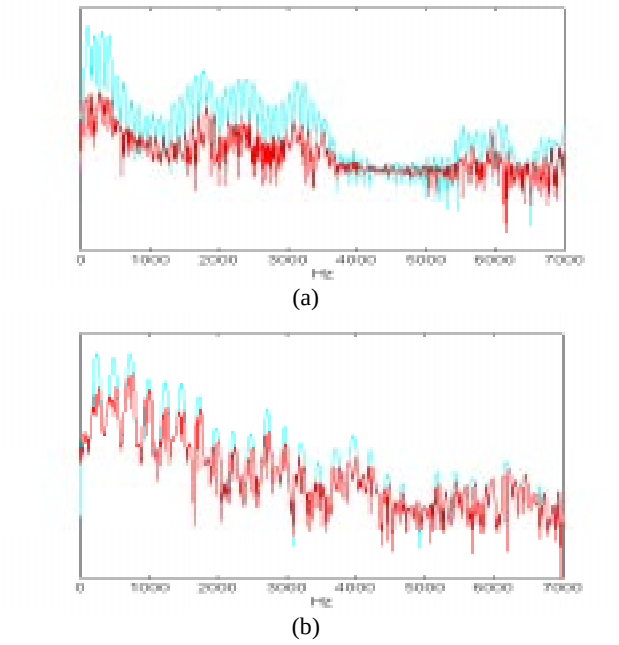
In the time domain, the error signal $\hat{e}_w(n)$ is obtained by summing sinewaves, each one corresponding to a harmonic of F_1, as was obtained the synthetic signal $\hat{s}_w(n)$ with the harmonics of F_0.

The final synthetic signal is now :

$$\hat{\hat{s}}_w(n) = \hat{s}_w(n) + \hat{e}_w(n) \quad (5)$$

and the error spectrum is :

$$F_w(\omega) = S_w(\omega) - \left(\hat{S}_w(\omega) + \hat{E}_w(\omega)\right) \quad (6)$$



(a)

(b)

Fig. 2. (a) Original spectrum (clear line) and error spectrum after the MBE modelling for a low frequency voice. (b) Original spectrum and error spectrum after the MBE modelling for a high frequency voice.

4.3.3.Results

The global evaluation of the energy of the error spectrum $E_w(\omega)$ and of the final spectrum error $F_w(\omega)$ for different speech signals proves that there is an improvement in the modelling of the spectrum. A detailed analysis shows that, as we expected, the gain is much more important for female voices than for male voices. Theoretically, the number of harmonics will be approximately doubled if we decide to transmit all the amplitudes of the spectrum $\hat{E}_w(\omega)$. The main interest of the method we propose is that it is possible to choose whether or not to transmit all the amplitudes. For example, when coding a male voice, if the synthetic spectrum $\hat{S}_w(\omega)$ fits well in a subband with the original spectrum $S_w(\omega)$, that means that the spectrum error is of very small energy, it is not necessary to transmit the amplitudes of $E_w(\omega)$ in that subband. However, when coding a female voice, the spectrum error can be of very high energy and the amplitudes of $E_w(\omega)$ are transmitted. So it is possible to obtain the same number of amplitudes whatever the value of the fundamental frequency is.

When comparing the signals obtained with this new modelling method, as we expected, the gain in quality is much more significant for female voices than for male voices. Listening evaluation have shown that the female voices sound more natural and less metallic.

Improvements can still be obtained when computing the error in the time domain. This signal error is, for the moment, computed as for the synthetic signal $\hat{s}_w(n)$, with interpolation between consecutive frames whereas this signal does not have the properties of the speech.

5 CONCLUSION

We propose a new structure for a wideband speech coder based on the MBE structure that seems to be promising. The constraints of bit rate was not mentioned here and will constitute our further research as the introduction of a psychoacoustic model during the analysis procedure or during the quantization procedure.

6. REFERENCES

[1] D.W. Griffin, "Multiband excitation vocoder", *Ph.D. dissertation*, M.I.T. Cambridge, MA, 1987.

[2] D.W. Griffin and J.S. Lim, "Multiband Excitation vocoder", *IEEE Trans. ASSP-36*, no. 8, pp 1223-1235, A., 1997.

[3] C.Garcia-Mateo, F.J. Casajus-Quiros, and L.A. Hernandez-Gomez, "Multi-band excitation coding of speech at 4.8 kbps", *Pro. ICASSP*, paper S1.4, 1990.

[4] A. Das and A. Gersho, "Enhanced Multiband Excitation Coding of Speech at 2.4 kb/s with Phonetic Classification and variable Dimension VQ", *Signal Processing VI*, pp 943-946, 1994.

[5] C.Garcia-Mateo, J.L. Alba-Castro, and E. R-Banga, "Speech Coding using Bi-harmonic spectral modeling", *Signal Processing VII*, pp 391-394, 1994.