A NEW POST-FILTERING FOR ARTIFICIALLY REPLICATED HIGH-BAND IN SPEECH CODERS

Guillaume Fuchs and Roch Lefebvre

University of Sherbrooke, Dept. of Electrical Eng., Sherbrooke,Québec, J1K 2R1 Canada
{guillaume.fuchs,roch.lefebvre}@usherbrooke.ca

ABSTRACT

Pure estimation in bandwidth extension of bandlimited speech does not provide transparent quality in the extended band. Combining coding and estimation of the missing band can achieve higher quality by transmitting additional information. However, substantial artifacts can remain in the extended band. In this paper, we introduce a new postfiltering designed to reduce these artifacts. Based on a source-filter model, the adopted bandwidth extension transmits spectral envelope parameters, while estimates the excitation at the receiver. The perceptual quality is then enhanced by processing the estimated excitation by a short-term and a long-term postfilter. Objective and subjective assessments show substantial gains when using the new postfiltering. It confirms that low bit-rate wideband coding based on bandwidth extension can get closer to the performances of the single band wideband coders.

1. INTRODUCTION

Telephonic networks are on the verge of switching from narrowband (0-4000Hz) to wideband voice communications (0-8000Hz). It is well known that the extended bandwidth enhances the speech quality especially for unvoiced phonemes [1]. Both speech intelligibility and naturalness are increased by adding bandwidth. Mainly two methods exist to transmit wideband speech : single band wideband coders and narrowband coders associated with artificial bandwidth extension. The last solution has the great advantage to be easy to implement and backward compatible to narrowband speech coders already deployed. Ideally, the receiver recovers the missing high-band without needing any additional information from the transmitter.

One the other hand, purely artificial bandwidth extensions are insufficient to obtain a high speech quality. These methods are based on the assumption that the transmitted low-band and the missing high-band share mutual informations. Indeed, the missing band is regenerated from informations extracted from the transmitted narrowband speech. But it can be shown that for important components like spectral envelope, the mutual information is limited [2]. Then, the use of explicit coding for such components along with estimation can highly improve the reconstructed high-band quality [3].

Our bandwidth extension is based on an auto-regressive (AR) speech production model to extract two components, the spectral envelope and the spectrally-flat excitation signal. The spectral envelope and energy are explicitly coded because of the lack of mutual informations shared with the low-band. Moreover, the spectral envelope is an important perceptual component compared to the excitation. In contrast, the bit-rate saving comes from the estimation of the high-band excitation generated by spectral folding at the decoder.

Spectral folding by oversampling aliasing is an efficient method for excitation estimation. While the technique is presumably popular for its simplicity, it shows surprisingly efficiency compared to other more sophisticated solutions. Indeed, it was selected as a part of a very recent audio codec standard, the AMR-WB+ [4]. The method outperformed in several cases the other challenging solutions especially for speech coding at low and medium rates. This can be explained by its noteworthy proprieties. First of all, spectral folding by oversampling aliasing is a totally temporal processing. Thus, it makes the replicated excitation totally phase coherent and makes the high-band inherit the temporal envelope from the low-band. Secondly, low and high-band excitations share common spectral characteristics exploited by folding. The tonality is about the same near the 2500Hz-5500Hz frequency range and the harmonic distance is the same over all the spectrum for speech. Naturally, it exhibits some differences and folding introduces some distortions. However, they are attenuated by the fact that human ear is less sensitive to the high frequencies and at the same time to the fine structure of the spectrum. Furthermore, small magnitudes in high frequencies for voiced segments attenuate the perceptual artifacts. However, the perceptual quality is spoiled by two main drawbacks. First, the folding introduces over-estimated periodic components in the highest frequencies. Secondly, the replicated excitation must be truly flat to not introduce high-band spectral distortion before envelope shaping.

Based on those observations, we introduce a new postfiltering for enhancing the replicated excitation when using spectral folding. As in the postfiletring for low bit-rate speech coders [5], a short-term postfilter is combined with a long-term postfilter. The short-term postfilter whitens the quantized excitation coming from the narrowband coding. The long-term postfilter takes on the high frequency over-periodic components.

In the next section, we will describe in more details the wideband coder using a combined estimation and coding scheme for the missing band. Section 3 outlines the artifacts occurring when using such a bandwidth extension and then introduces the new postfiltering to overcome them. Objective measurements and subjective evaluations pointing out the effectiveness of postfiltering are presented in section 4. Finally, our conclusions are given in section 5.

2. WIDEBAND CODING USING BANDWIDTH EXTENSION

2.1. Wideband Coding Scheme

As illustrated in Fig. 1, a band split coding scheme is used in our wideband coder. The input signal sampled at 16kHz is decomposed in two 8kHz downsampled bands using a quadrature mirror filter (QMF) banks. The low-pass filtered band is encoded by a CELP narrowband coder like G.729 [6]. At the receiver, the information

from the CELP encoder can be decoded independently to generate a narrowband synthesis. On the contrary, the higher band is encoded by a highly parametric coding scheme and can only be recovered by using in addition of the transmitted parameters, the information from the narrowband synthesis.

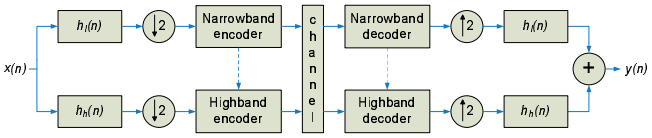


Fig. 1. Wideband coding by splitting the input into two equal bands.

2.2. High-Frequency Band Coding

We use an AR source model to generate the high-band signal. The AR filter representing the shape of the vocal tract is perceptually very important. As a result, the coefficients of the filter are explicitly transmitted. On the other hand, the excitation is simply estimated at the receiver.

The AR model is estimated by a linear predictive (LP) analysis of order 10. The generated coefficients are then quantized in the LSFs domain. Because they have different statistics from the lower band, we designed codebooks using a training sequence on the high-band parameters. The LSFs are transmitted at every 20ms frame, and are linearly interpolated every 5ms. The rate from LSFs is of 1 kbps.

The excitation is estimated by a spectral folding of the low-band synthesized excitation. It is done in the temporal domain by using oversampling aliasing propriety. With the QMF structure, it consists simply to copy at the receiver the low-band synthesized excitation in the high-band. The energy ratio between the low-band and the high-band is vector quantized at every 20ms frame and interpolated each 5ms. This results in an additional rate of 0.4 kbps.

In the end, we obtain a side information for the high-band coding of only 1.4 kbps while doubling the bandwidth. The saving allows to obtain a low bit-rate wideband coder or could be devoted for improving the low-band description.

3. POSTFILTERING OF THE EXTENDED BAND

Postfiltering was first designed for speech coders with low bit-rates [5]. For these coders, noise spectral shaping is not sufficient to make noise inaudible over the entire spectrum. Therefore, valley regions are sacrificed to preserve perceptually important formants and harmonic peaks. In order to minimize the Mean Square Error (MSE), the postfiltering tries to follow the optimal Wiener filter $H(\omega) = S(\omega)/[S(\omega) + N(\omega)]$. By this way the postfiltering has a gain near unity in the frequency range where the SNR is large, and low where the SNR is small. In the replicated band, such a postfiltering cannot be directly applied because the signal is only a parametric estimation not correlated to the original high-band.

3.1. Short-term Postfilter

At low bit-rates, it is well-known that speech coders like CELP favor the formant regions of the spectrum. Therefore, the spectrum of the synthesis signal has more energy in the formant regions even in the excitation signal. Furthermore, the LP synthesis filter will not in general produce a truly flat spectrum. Thus, it is necessary to flatten the replicated excitation coming from a low bit-rate coding.

The short-term postfilter is a whitening filter. This is done by performing an LP analysis on the replicated excitation to obtain an all-zero filter $A_{exc}(z/\gamma_1)$. To add more flexibility to the short-term postfiltering, we add an all-pole portion $1/A_{exc}(z/\gamma_2)$. By this way we can control the generated spectral tilt and the formant expansion.

$$H_{st}(z) = \frac{A_{exc}(z/\gamma_1)}{A_{exc}(z/\gamma_2)}, \quad 0 \leq \gamma_2 \leq \gamma_1 \leq 1 \qquad (1)$$

We use an order of 10, like the narrowband LP analysis. Experimentally we chose parameters γ_1 and γ_2 to be 1 and 0 respectively when using the short-term postfilter alone, and 1 and 0.3 with it is associated with the long-term postfilter.

The Fig. 2 shows the frequency responses of the short-term narrowband synthesis filter $1/A(z)$ and the short-term postfilter generated from the replicated excitation of the G.729. It can be seen that the short-term postfilter formants correspond to the valleys of the synthesis filter. In Fig. 3 the magnitude spectrum of a voiced segment is shown and compared with its syntheses before and after posfiltering. We notice that the short-term postfiltering makes the synthesis magnitude spectrum closer to the original.

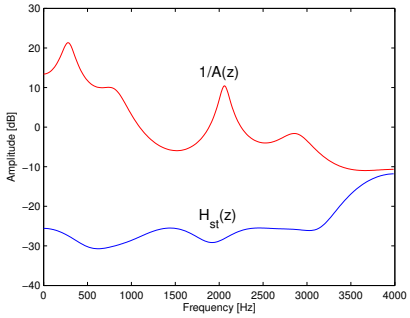


Fig. 2. The short-term postfilter frequency response compared to the narrowband synthesis filter.

As an alternative, for complexity reasons we can use the $A(z)$ instead of $A_{exc}(z)$ in $H_{st}(z)$. This is less effective and adaptive, however it can produce similar effects if handled with care.

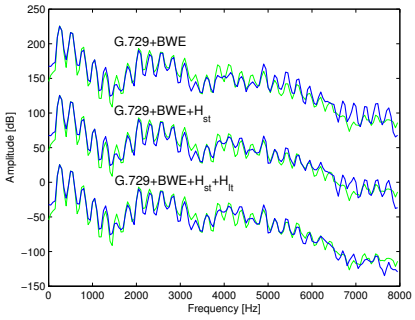


Fig. 3. Magnitude spectrums of the G.729 combined with bandwidth extension, without postfiltering, with short-term postfiltering, and with short and long-term postfiltering (the magnitude spectrums are superposed with the original and shifted to enhance visibility).

3.2. Long-term Postfilter

Spectral folding introduces an over-estimated voicing strength in the high frequencies of the missing band. Indeed, the low frequencies of the low-band excitation are replicated in the high frequencies of the missing high-band, and vice versa. Thus, the degree of voicing in the low frequencies of the high-band is considered well estimated while in the high frequencies over-estimated.

Consequently, we design for the replicated excitation from the low-band a filter combining an anti-harmonic response $H_a(z)$ for the low frequencies and an all-pass response $H_b(z)$ for the high frequencies. The anti-harmonic filter given in (2) is the inverse of the comb filter introduced in [5] which models the pitch-harmonic structure of the excitation.

$$H_a(z) = \frac{1-\beta_1 z^{-P}}{1+\beta_2 z^{-P}} \quad (2)$$

where $0 < \beta_1, \beta_2 < 1$ are function of the voicing strength in the present speech frame. The voicing strength can be estimated from the weight b of the one-tap pitch predictor. The pitch delay P and the voicing parameter b are used directly from the transmitted parameters of the narrowband coder.

$$\beta_1 = C_{\beta_1} f(b), \quad \beta_2 = C_{\beta_2} f(b), \quad 0 < C_{\beta_1}, C_{\beta_2} < 1 \quad (3)$$

with

$$f(b) = \begin{cases} b & \text{if} \quad 0 \le b \le 1 \\ 1 & \text{if} \quad b > 1 \end{cases} \quad (4)$$

Since the transition must be smooth between the two parts of the anti-harmonic and all-pass responses, we use simple first-order low-pass and high-pass filters, respectively $H_l(z)$ and $H_h(z)$, as the subband decomposition. Furthermore, to simplify the filter, and to confine the memory propagation, we only use the all-zero portion of $H_a(z)$. Consequently, the anti-harmonic structure model is less accurate but easier to tune, and then often more effective. Moreover, because we want to attenuate almost all the periodic components of the replicated excitation in low frequencies, we chose to fix C_{β_1} closed to 1. Finally, we obtain:

$$H_{lt}(z) = H_l(z)H_a(z) + H_h(z)H_b(z) \quad (5)$$

$$= \frac{1}{2}(1+z^{-1})(1-\beta_1 z^{-P}) + \frac{1}{2}(1-z^{-1}) \quad (6)$$

$$= 1 - \frac{\beta_1}{2} z^{-P} - \frac{\beta_1}{2} z^{-(P+1)} \quad (7)$$

The spectral characteristics of the long-term postfilter is illustrated in Fig. 4. Its effect on the pitch-harmonic structure $1/H_a(z)$ is shown too. The Fig. 3 gives an example of the postfiltering effects on a voiced segment.

It is worth notifying that gain adjustment after postfiltering is not a need because the postfiltered excitation is spectrally shaped and energy normalized afterward.

4. PERFORMANCE EVALUATION

4.1. Objective Measurements

We use the Spectral Distortion (SD) and the Spectral Flatness Measure (SFM) on the excitation for the objective assessment. It is known that they are both perceptually important criteria. The SD is made by comparing the spectral envelope of the estimated excitation $\tilde{A}_k$ with the original A_k. Accurate spectral envelopes are calculated

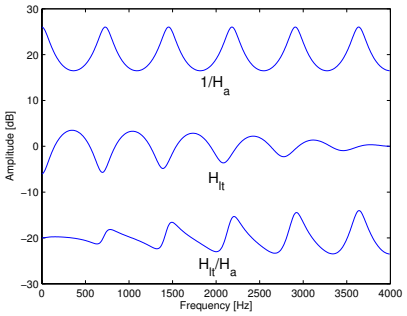


Fig. 4. The amplitude frequency responses of an harmonic excitation, the associated long-term postfilter and the filtered harmonic excitation.

from cepstral transformation. Finally, the SD is calculated over N frames as follows:

$$SD = \sqrt{\frac{1}{N}\sum_{n=1}^{N} \frac{1}{k_2-k_1+1} \sum_{k_1}^{k_2} \left[20\log_{10}\left(G_H \frac{A_k(\omega)}{\tilde{A}_k(\omega)} \right) \right]^2} \quad (8)$$

with

$$G_H = \frac{1}{k_2-k_1+1} \sum_{k=k_1}^{k_2} \frac{\tilde{A}_k}{A_k} \quad (9)$$

where k_1 and k_2 are the lower and higher indexes for the frequency range evaluated.

The SFM [7] is an important feature for psychoacoustic models that can be used as a tonality coefficient.

$$SFM = -\frac{10\log_{10}(\gamma)}{15\text{dB}} \quad (10)$$

with

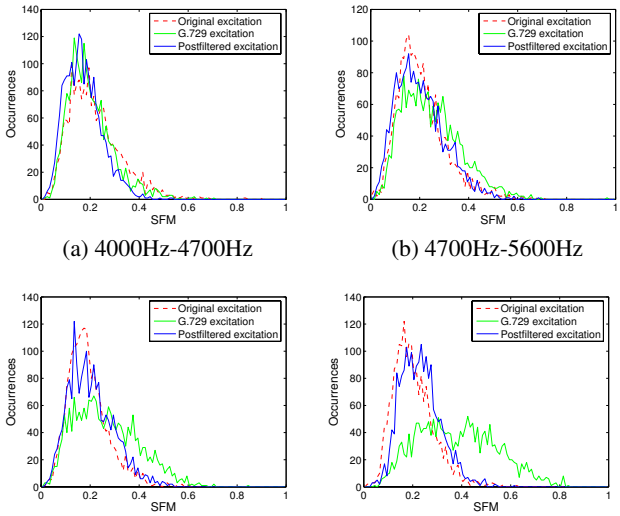
$$\gamma = \frac{\left[\prod_{k=k_1}^{k_2} S_k\right]^{(k_2-k_1+1)}}{\frac{1}{k_2-k_1+1}\sum_{k_1}^{k_2} S_k} \quad (11)$$

where S_k is the sampled Power Spectral Density at the k frequency index, and 15dB is a constant experimentally fixed to constraint SFM between 0 and 1. If $SFM = 1$ the signal is totally tonal, otherwise if $SFM = 0$ it is noiselike. SFM is calculated using coefficients of a Discrete Fourier Transform.

The objective assessment is processed over a 2 minutes speech sequence composed of mixed english and french sentences uttered by female and male voices. The measurements are calculated every 20ms frame. The SD is computed over the frequency range 4000Hz-8000Hz. SFM is calculated for the same band but for each critical bands. To corroborate the choice of this two criteria, we apply the Wideband PESQ [8] objective evaluation. As shows in Table 1, the short-term postfilter improves the SD. On the other hand, the combination with the long-term postfilter increases a little bit the distortion. However, the Wideband PESQ MOS grade is higher when using both filterings. This can be explained by the other objective criterion. Indeed, Fig. 5 shows the important improvement of the excitation tonality characteristic after postfiltering. The tonality is well estimated for the first two critical bands. Conversely, spectral folding fails to model properly the tonality in the higher frequencies and long-term postfilter overcomes a great part of this problem.

G729+BWE	SD [dB]	Wideband PESQ [MOS]
without postfiltering	6.05	2.691
short-term postfiltering	5.54	2.720
combined postfiltering	5.83	2.728

Table 1. SD and Wideband PESQ for different bandwidth extensions



(a) 4000Hz-4700Hz (b) 4700Hz-5600Hz

(c) 5600Hz-6700Hz (d) 6700Hz-8000Hz

Fig. 5. Histograms of the SFM in the critical bands of the high-band.

4.2. Subjective Evaluation

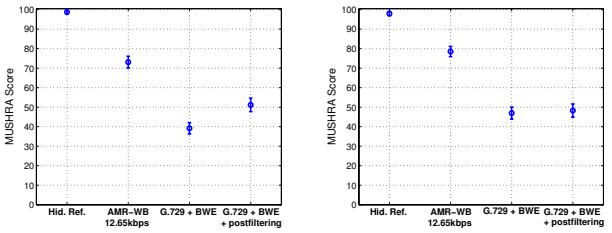
We conducted a listening test for subjective assessment using the MUSHRA methodology. The test was designed in order to grade the degree of artifacts in the bandwitdh extended signal. As a reference, the AMR-WB at 12.65 kbps was included in the test. 14 trained listeners evaluated 12 sentences uttered by men and women in english and in french. All the signals were high-pass filtered with a cutoff frequency of 200Hz. This emphasizes the high-band quality evaluation. The purpose was not to evaluate the improvement over the narrowband coder. This improvement was already stated in many works. Moreover, we did not want to make the bandwidth difference between syntheses the foremost importance.

The results of the test shown in Fig. 6, reveal that the postfiltering is more efficient for female voices. In this case the improvement is as high as 12 Mushra points. For male voices, the improvement is not so important, but in no case the postfiltering degrades the synthesis. Finally, we notice that using G.729 as narrowband coder, our wideband coder produces good speech quality. However, estimation artifacts even after postfiltering, handicaps the subjective quality in comparison with the AMR-WB at 12.65 kpbs. It reveals the difficulty to obtain a good high-band estimation from a low bit-rate description of the low-band.

5. DISCUSSION

We presented a new method for enhancing the quality of bandwidth extension based on spectral folding. The method consists in a short-term and a long-term postfilter which improve spectral characteristics of the replicated excitation. We applied our enhancing method to a bandwidth extension transmitting as low as 1.4 kbps of side-information used to code the high-band spectral envelope. Using ITU-T recommandation G.729 as the narrowband core, the total bit-rate is thus 9.4 kbps to encode a wideband signal.

The new postfiltering removes part of the high-band artifacts and brings the synthesis signal closer to the original as demonstrated by objective and subjective measures at no additional bit-rate. As expected, at 9.4 kbps the subjective quality remains below the performance of the AMR-WB at 12.65 kbps. With the new postfiltering method, it would be interesting to investigate what bit-rate should be allocated to the narrowband coding to achieve a quality equivalent to a single band wideband coder.



(a) Female voices (b) Male voices

Fig. 6. Mushra scores for female and male speech.

6. REFERENCES

[1] S. Voran, "Listener rating of speech passbands," *IEEE Workshop on Speech Coding*, pp. 81–82, 1997.

[2] M. Nilsson, S.V. Andersen, and W.B. Kleijn, "Gaussian mixture model based on mutual information estimation between frequency bands in speech," *IEEE Int. Conf. on Acoustics Speech and Signal Processing*, 2002.

[3] Yannis Agiomyrgiannakis and Yannis Stylianou, "Combined estimation/coding of highband spectral envelopes for speech spectrum expansion," *IEEE Int. Conf. on Acoustics Speech and Signal Processing*, 2004.

[4] J. Makinen and al., "AMR-WB+: a new audio coding standard for 3rd generation mobile audio services," *IEEE Int. Conf. on Acoustics Speech and Signal Processing*, pp. 1109–1112, Mar. 2005.

[5] Juin-Hwey Chen and Allen Gersho, "Adaptive postfiltering for quality enhancement of coded speech," *IEEE Trans. on Speech and Audio Processing*, Jan. 1995.

[6] R. Salami and al., "Design and description of CS-ACELP: a toll quality 8 kb/s speech coder," *IEEE Trans. on Speech and Audio Processing*, vol. 6, pp. 116–130, Mar. 1998.

[7] N. S. Jayant and P. Noll, *Digital coding of waveforms*, Prentice-Hall, 1984.

[8] "Proposed modification to draft P.862 to allow PESQ to be used for quality assessment of wideband speech (BT, UK, and KPN, the Netherlands," *ITU-T SG12 Delayed Contribution COM-D007-E*, Feb. 2001.