INTER-TONE NOISE REDUCTION IN A LOW BIT RATE CELP DECODER

Tommy Vaillancourt[1,2], Milan Jelinek[1,2], Redwan Salami[1], Vladimir Malenovsky[1,2], and Roch Lefebvre[2]

[1]VoiceAge Corporation, Montreal, Qc, Canada
[2]University of Sherbrooke, Qc, Canada

ABSTRACT

In this paper we present a novel technique to enhance music signals encoded using a low bit rate CELP coder. The method is based on reduction of inter-tone quantization noise for decoded music signals without affecting the quality for speech signals. The proposed technique consists of two modules. The first module is used to discriminate between stable tonal sounds and other sounds and the second module is used to reduce the inter-tone quantization noise in the stable tonal segments. The inter-tone noise is reduced by means of spectral subtraction. The proposed method is a part of the newly standardised ITU-T G.718 codec.

***Index Terms*— s**peech coding, CELP, noise reduction, music enhancement

1. INTRODUCTION

In some applications, such as music-on-hold, low bit rate speech codecs are required to operate on music signals. This usually results in poor music quality due to the use of a speech-specific model such as Code-excited Linear Prediction (CELP).

For some music signals, the spectrum exhibits a stable structure with several tones (corresponding to spectral peaks) which are not harmonically related. These signals are difficult to encode with CELP-based codecs which use an all-pole synthesis filter to model the spectral envelope, and a long-term pitch predictor to model the periodicity in the excitation signal. The pitch filter is very efficient for voiced segments of a speech signal where the spectrum exhibits a harmonic structure. However, it fails to properly model tones which are not harmonically related. Further, CELP coders exploit the masking property of the human ear by shaping the quantization noise so that it has more energy in the formant regions where it is masked by the strong signal energy [6]. Unfortunately, the LP-based perceptual weighting, performed in CELP, is not suitable to hide the quantization noise in case of tonal music sequences. Thus, the quantization noise in the low-energy regions of the spectrum (e.g. spectral valleys and inter-tone regions in music signals) becomes audible, especially at low bit rates.

In this paper, we propose an efficient post-processing method based on inter-tone noise reduction which significantly increases the quality of decoded stable tonal music signals. The paper is organized as follow. In Section 2, the issue of inter-tone noise is discussed and the technique of inter-tone noise reduction is described in detail. Section 3 explains how the proposed technique is integrated in the ITU-T G.718 standard. The performance of the proposed approach is evaluated in Section 4 and conclusions are drawn in Section 5.

2. INTER-TONE NOISE REDUCTION

The algorithm operates in the frequency domain as shown in Figure 1 below. Spectral analysis is performed at each 20-ms frame, using a 30-ms transform window with 33% overlap. Inter-tone noise attenuation and gain correction are applied to the spectral parameters. The amount of inter-tone noise reduction is controlled by a signal classifier where more reduction is applied in case of stable music signals. An inverse FFT is used to convert the enhanced signal back to the time domain and overlap-add technique is then used to reconstruct the signal. Prior to the conversion from time to frequency domain, a fixed pre-emphasis with a first order high-pass filter is used to flatten the spectrum. Its inverse, the de-emphasis, is applied at the end, after the IFFT and the overlap-add. These pre- and post-processing operations are not essential, but they are helpful to maintain a reasonable spectral dynamics useful for fixed-point implementation of the algorithm.

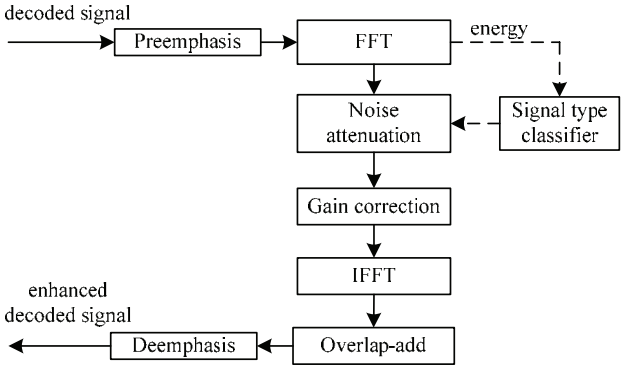

Figure 1: Principle of the inter-tone noise reduction.

2.1. Signal classification

In general, speech signals are well coded with a CELP-based codecs and do not need this type of frequency-based post-processing. Moreover, a serious degradation may appear when applying the inter-tone noise reduction to speech spectrum. A signal type classifier has been designed to maximize the efficiency of the inter-tone noise reduction by identifying the sounds suitable for the algorithm, such as stable music, and rejecting those which are not suitable, such as speech. This signal classifier is based on stability estimation.

One of the most challenging problems in the design of such classifier is its robustness. The classifier must be robust enough to ensure no degradation on clean and noisy speech and yet be able to detect enough stable tonal music items to benefit from the effects

of the inter-tone noise reduction. It was found that the most discriminative feature is related to the variation of the frame energy. The total energy E_{fr}^{t} of the frame t is computed as the logarithm of the sum of the average spectral energies $\bar{E}_{CB}(i)$ in each of the k critical bands [5], i.e.:

$$E_{fr}^{t} = 10\log\left(\sum_{i=0}^{k}\bar{E}_{CB}(i)\right).$$

The total energy E_{fr}^{t} itself cannot sufficiently discriminate stable tonal music frames from other sounds, but its variation in time contains valuable information. The energy difference from frame to frame Δ_{E}^{t} is first calculated, and then averaged $\bar{E}_{diff}$ for the last 40 frames as follows:

$$\bar{E}_{diff} = \frac{\left(\sum_{t=-40}^{t=-1}\Delta_{E}^{t}\right)}{40},\ where\ \Delta_{E}^{t} = E_{fr}^{t} - E_{fr}^{(t-1)}$$

Then, a weighted deviation of the energy variation in the last 15 frames is computed:

$$\sigma_{E} = 0.7745967\cdot\sqrt{\sum_{t=-15}^{t=-1}\frac{\left(\Delta_{E}^{t}-\bar{E}_{diff}\right)^{2}}{15}}.$$

The resulting deviation is then compared to an adaptive threshold to determine whether the frame is suitable for the application of the inter-tone noise reduction. Considering that false detection of stable tonal music on a speech frame could result in serious degradation, the stability decision is soft and split into 5 categories. Each category has its own setup of inter-tone noise reduction tuning. The first category (Category 0) is a non-stable sound where the deviation σ_E is high. At the other end, Category 4 corresponds to a very stable frame where the deviation is very low. Table 1 shows the categories and the associated frequency bands in which the inter-tone noise reduction is applied, along with the maximum allowed level of noise reduction R_{max}. F_s corresponds to the output sampling frequency. Category 0 corresponds to a non-stable sound and no inter-tone noise reduction is performed for this category. Similarly, Category 1 is considered as mostly non-stable and therefore any possible noise reduction is limited only to higher frequencies with a maximum attenuation level of 6dB. Categories 2, 3 and 4 are considered as stable, and the modifications to the spectrum extend down to 400 Hz with a maximum possible attenuation of 12 dB (Category 4).

Category	Enhanced band [Hz]	Max. allowed reduction (R_{max}) [dB]
(0) Non-stable	N/A	0
(1) Mostly non-stable	[2000, $F_s/2$]	6
(2) Quasi stable	[1270, $F_s/2$]	9
(3) Mostly stable	[700, $F_s/2$]	12
(4) Stable	[400, $F_s/2$]	12

Table 1: Frequency band and noise reduction level

The tuning of each category has been performed in such a way that if a frame is miss-classified, it will not be perceptually degraded by the inter-tone noise reduction. The frequency boundaries chosen have been tested against all the speech genres presented in Table 2. The thresholds are further made adaptive to increase the discriminative power of the classifier to split sounds into the different categories mentioned above. Typically, a stable sound like music has a much lower deviation of its energy variation than a non-stable sound like speech and it is unlikely that speech or music content changes from frame to frame. If the number of consecutive frames classified as Category 2, 3 or 4 is greater than 30, the adaptive thresholds are increased to allow for more energy variation. The inverse is also true for frames classified as Category 0 and 1. Thus, the adaptive thresholds favor the current state of the classifier.

2.2. Background noise update

To determine the amount of noise to be removed with the inter-tone noise reduction method, it is necessary to update regularly the background noise energy. This update is performed for each frame using the following formula:

$$N_{CB}^{(t)}(i) = \frac{\left(0.6E_{CB}^{(t)}(i) + 0.2E_{CB}^{(t-1)}(i) + 0.2N_{CB}^{(t-1)}(i)\right)}{16.0},\ i=0,..,16,$$

where $N_{CB}^{(t)}(i)$ represents the estimated background noise energy in a critical band i and $E_{CB}^{(t)}(i)$ represent the current frame energy for the same critical band. Furthermore, $N_{CB}^{(t-1)}(i)$ and $E_{CB}^{(t-1)}(i)$ represent the same quantities for the previous frame.

2.3. Inter-tone noise reduction

The inter-tone noise reduction can start at 400Hz. To reduce any negative impact of the technique, the signal-type classifier module can push the starting frequency up to 2000Hz as is seen in Table 1. Further, the starting frequency can change from frame to frame.

The inter-tone noise reduction is performed in the frequency domain and the enhanced signal is then reconstructed using the overlap-add operation. The noise reduction is achieved by scaling the spectrum in each critical band on which the reduction is allowed. The scaling gain is upper-limited by 1 and lower-limited by a minimum gain g_{min}. The minimum gain is derived from the maximum allowed inter-tone noise reduction R_{max} in dB. That is:

$$g_{\min} = 10^{-R_{\max}/20}.$$

As described above (see Table 1), the signal-type classifier mandates the maximum allowed reduction to be between 6 and 12 dB which results in a minimum gain between 0.25 and 0.5. The scaling gain g_s is derived from the signal-to-noise ratio (SNR) estimated for each frequency bin f. In the computation the background noise energy is updated on a critical band basis, but the scaling is performed on frequency bin basis. In other words, the scaling gain is applied on every frequency bin derived from the SNR for that bin. That is:

$$g_{s}(f) = \sqrt{k_{s}SNR(f) + c_{s}}.$$

The SNR is computed using the current frame bin energy divided by the noise energy of the critical band including that bin. The scaling function $g_s(f)$ is bounded by $g_{min} \le g_s \le 1$. The values

k_s and c_s are determined such as $g_s = g_{min}$ for $SNR = 1$, and $g_s = 1$ for $SNR = 45$. That is, for $SNR = 1$ and lower, the noise reduction is limited to R_{max} dB, and no noise suppression is performed if SNR is 45 or higher.

This feature allows for preserving the energy at frequencies near tones while strongly reducing the noise between the tones. The scaling applied to the spectrum is performed using a smoothed scaling gain updated in every frame as

$$\overline{g}_s(f) = \alpha(f) \bullet \overline{g}_s(f) + (1-\alpha(f)) \bullet g_s(f) ,$$

where the smoothing factor α is inversely related to the gain and is given by

$$\alpha(f) = 1 - g_s(f) .$$

That is, the smoothing is stronger for smaller gains g_s. Temporal smoothing of the gains prevents audible energy oscillations, especially in higher part of the spectrum. Controlling the smoothing using α prevents distortion in high SNR segments preceded by low SNR segments, similarly to what has been implemented in the VMR-WB [4] noise reduction algorithm [1].

Figure 2 shows the effect of the inter-tone noise reduction on a spectrum corresponding to a stable tonal music sequence. It is clearly seen that the noise between the harmonics is reduced while the level of spectral peaks is basically unchanged. Thus, the inter-tone noise reduction method does not affect the tones. However, as will be explained in the next section, while correcting the frame energy, the tone amplitudes are slightly increased to further enhance the signal.

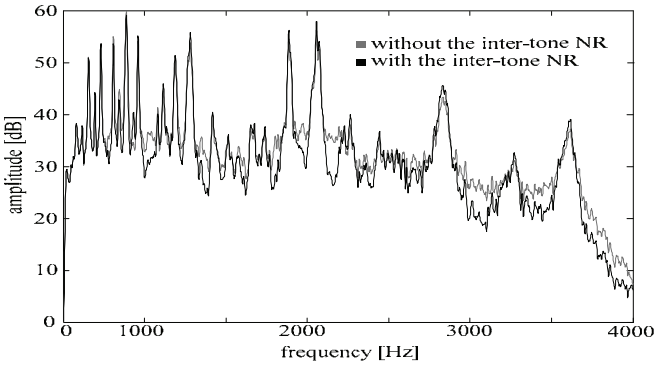

Figure 2: Spectrum modified with the proposed method.

2.4. Gain correction

Performing noise energy reduction between the harmonics may result in an overall drop of the signal energy. To preserve this energy, we further rescale the spectrum in each critical band in a way that the energy in each band at the end of the rescaling is close to the energy before the noise reduction. To achieve this energy compensation, it is not necessary to rescale all the bins but only the bins with highest energy. Thus, the energy removed from the regions between the tones is moved directly to the regions of the tones. Precisely, the energy is moved to the most energetic events of the respective critical band. In this way the enhanced signal will sound clearer, due to the fact that the spectral dynamics is increased.

A first-stage gain g_α per critical band is computed as a ratio of the energy per critical band before and after the inter-tone reduction. That is:

$$g_\alpha(i) = \sqrt{\left(E_{CB}(i) \Big/ E'_{CB}(i) \right)} ,$$

where E_{CB} is the energy per critical band before and E'_{CB} the energy per critical band after the inter-tone noise reduction.

At this stage it has been observed that the CELP model reproduces accurately the energy of the low frequencies but tends to under-estimate the energy of the high frequencies. This under-estimation of the high-frequency content led to a perceived degradation for some musical samples.

To solve this issue a second-stage per-bin rescaling gain is proposed which is based on a ratio of the number of energetic events N_{ev} in a critical band i and the total number of bins N_{bins} in that critical band. An energetic event is defined as a bin to which a scaling gain greater than 0.8 was assigned during the inter-tone noise reduction. This second-stage rescaling gain g_β is calculated as follows:

$$g_\beta(i) = -0.2778 \frac{N_{ev}(i)}{N_{bins}(i)} + 1.2778 ,$$

where the constants were found empirically. This new gain multiplies the scaling gain g_α to obtain the final gain correction g_{corr} per critical band defined as follow:

$$g_{corr}(i) = g_\alpha(i) \bullet g_\beta(i) .$$

The gain correction factor is applied only to the highest bins of a critical band, i.e. the bins where the scaling gain of the inter-tone noise reduction was greater than 0.8.

3. INCLUSION OF THE PROPOSED METHOD IN THE G.718 CODEC

The presented inter-tone noise reduction technique is a part of the new ITU-T G.718 [2][3] speech and audio coding standard. It was introduced to enhance its performance for narrowband music signals. G.718 comprises 5 embedded layers operating in the range from 8 to 32 kbit/s. The lower two layers are based on the CELP technology. The error from layer 2 is further coded by higher layers in a transform domain using the modified discrete cosine transform (MDCT). The codec has been designed to process both wideband (16 kHz) and narrowband input signals (8 kHz).

For narrowband signals, the requirements for the G.718 codec were very high. In particular, the G.718 codec had to perform not worse than G.729E in music [8]. The reference codec G.729E [7] at 11.8 kbps has been designed with a mixed forward/backward LP structure where the backward LP structure was specifically introduced to improve the performance on music signals. The inherent embedded structure of the G.718 codec and its underlining speech model made it difficult to achieve a desired performance for music signals at low bit rates. The inter-tone noise reduction introduced in this paper made it possible to meet these requirements.

4. PERFORMANCE

The performance of the proposed inter-tone noise reduction method has been evaluated with objective and subjective measures. The objective evaluation consisted in the assessment of the signal-type classifier performance and corresponded to a verification of the percentage of its accurate decisions. The rate of

miss-classification has been used to determine the signal-type classifier robustness. As mentioned before, the output from the signal-type classifier of Category 0 and 1 is considered as non-stable, otherwise it is considered as stable. Ideally, the signal-type classifier should classify all speech sequences as non-stable because the speech quality should not be affected with the inter-tone noise reduction method. On the other side, the signal-type classifier should classify most of the tonal music sequences as stable, because these are the items for which the proposed technique brings a significant quality improvement.

The clean speech sequences tested included the following four languages: French, English, Finnish and Chinese. The noise included in the noisy speech database was either office noise or car noise. Finally, the music database included classical, rock, orchestra, pop and vocal genres. In table 2 we can observe that miss-classification in case of speech signal is below 5% and the detection of stable music is higher than 60%. As the classifier is basically a stability detector, it is normal that not all music segments are considered as stable.

Input type category Language type	Non-stable (0 and 1) [%]	Stable (2, 3 and 4) [%]
English	97.84	2.16
French	98.30	1.70
Finnish	98.17	1.83
Chinese	97.18	2.81
Off. noise	95.03	4.97
Car noise	97.41	2.59
Music database 1	15.63	84.37
Music database 2	39.78	60.22

Table 2: Accuracy of the signal classifier

The subjective evaluation of the proposed method consisted in two formal MOS tests comparing performance with and without the usage of the inter-tone noise reduction technique. In case of speech, Figure 3 shows that the technique maintains the same quality. This can be observed by comparing the relative difference of MOS scores between the tested and reference codecs in the upper and the lower part of Figure 3. The reference codecs are the G.729 [9] and the G.729E while the tested codec is the G.718 at 8 and 12 kbps.

In case of music, the subjective performance of the G.718 codec has been significantly improved for both tested bit rates. It can be seen that the G.718 codec with the inter-tone noise reduction technique is performing slightly higher than G.729E.

5. CONCLUSION

In this paper we have shown a new technique to increase the quality music signals encoded by a low bit rate CELP codec. The proposed method is based on inter-tone noise reduction and is applied in the decoder of the CELP codec. The method relies on stability classification of the past decoded signal. To this end, five categories are established, each associated with a different starting frequency and a maximum allowed noise reduction. A highly robust behavior is achieved by applying progressive gain scaling among critical bands based on SNR values calculated for each frequency bin.

The technique is a part of the new G.718 codec, recently approved by the ITU-T. A formal MOS test, similar to the official test conducted by ITU-T, showed that the performance for music signals was improved by approximately 0.3 MOS while the performance for speech signals remained basically unaffected.

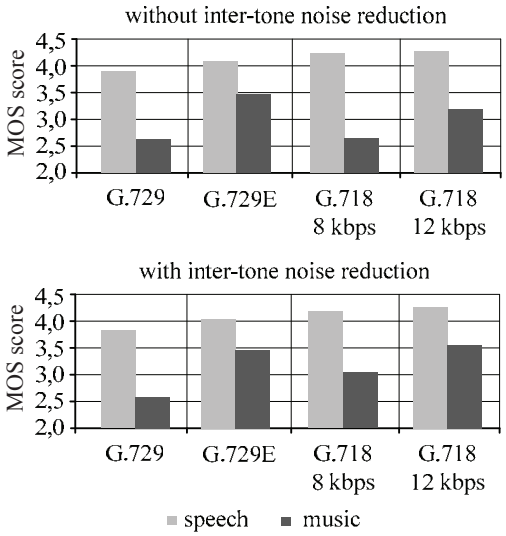

Figure 3: Results of the inter-tone noise reduction method.

6. REFERENCES

[1] M. Jelinek and R. Salami, "Noise Reduction Method for Wideband Speech Coding", *European Signal Processing Conf. (Eusipco), Vienna, Austria, September, 2004*

[2] T. Vaillancourt, and al, "ITU-T G.EV-VBR: a robust 8-32 kbit/s scalable coder for error prone telecommunications channels", *in Proc. European Signal Processing Conf. (Eusipco), Lausanne, Switzerland, Aug. 2008*

[3] M. Jelinek, and al., "ITU-T G.EV-VBR Baseline Codec", *in Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing, Las Vegas, USA, March 2008*

[4] M. Jelinek and R. Salami, "Wideband Speech Coding Advances in VMR-WB standard", *IEEE Transactions on Audio, Speech and Language Processing*, Vol. 15, no. 4, May 2007.

[5] J. D. Johnston, "Transform coding of audio signal using perceptual noise criteria," *IEEE J. Select. Areas Commun*., vol. 6, pp. 314–323, Feb. 1988.

[6] B. Bessette, and al., "Efficient methods for high quality low bit rate wideband speech coding", *IEEE Speech Coding WorkShop, Ibaraki, Japan*, pp. 114-116, October 2002

[7] ITU-T Rec. G.729 Annex E, "Coding of Speech at 8 kbit/s using Conjugate Structure Algebraic Code Excited Linear Prediction (CD-ACELP): 11.8 kbit/s CS-ACELP speech coding algorithm," Sept. 1998

[8] ITU-T Q9/SG16, TD-157, Annex A of Q9/16 Report, "Terms of Reference for the Embedded VBR (EV) Audio Coding Algorithm", Geneva, 6 - 16 April 2006

[9] R. Salami, et al., "Design and description of CS-ACELP: a toll quality 8 kb/s speech coder," *IEEE Trans. on Speech and Audio Processing*, vol.6, no. 2, pp. 116–130, March 1998.