Reducing the Resolution of the Pitch Lag in the CELT Pitch Prefiltering and Postfiltering Operations

Daniele Giacobello

March 31, 2011

Abstract

This document presents a feasibility study on the reduction in resolution of the pitch lag T_p used in the CELT pitch prefiltering and postfiltering operations. In particular, the purpose of this work is to exploit well-known concepts in *pitch perception* in order to reduce the bits necessary to describe T_p without decreasing the overall quality of audio and speech data coded with CELT. Several experiments have been conducted and a subjective and objective evaluation suggests that the decrease in audio quality is irrelevant even when the reduction in quantization resolution is significant.

1 Introduction

The ability of a listener to distinguish between two different notes on a musical scale follows a logarithmic relation with respect to the fundamental frequency[1]. This logarithmic relation is already exploited in speech coding where different resolutions are used depending on the value of the pitch lag. An example can be seen in the AMR-WB coder where, at a sampling frequency of 16 kHz, the resolution is 1/4 sample in the range $[34, 127]$, 1/2 sample in the range $[128, 159]$, and 1 sample in the range $[160, 231]$. Since the current implementation of the pitch filter in the CELT coder is studied to accommodate several different kind of audio signals with a larger range

[1] In the speech and audio coding community is often useful to assimilate the subjective concept of *pitch* with the objective measure of the *fundamental frequency*. We will follow this convention since, for most practical applications, these two concepts coincides.

Table 1: Pitch lag T_p and pitch frequency $f_p = 1/T_p$ ranges of the five considered octaves.

OCTAVE	T_p	f_p [Hz]
O_1	[32, 63]	[1500, 750[
O_2	[64, 127]	[750, 375[
O_3	[128, 255]	[375, 188[
O_4	[128, 255]	[188, 94[
O_5	[512, 1023]	[94, 47[

for the pitch lag T_p, we can reasonably assume that greater gains in efficiency can be obtained.

This document is organized as follows. In Section 2 we describe our method. In Section 3 we present the experimental analysis for several kinds of audio signals. In Section 4 we conclude our work.

2 Description

The current approach in the CELT coder is to code T_p, where $T_p \in [32, 1023]$ for sampling frequency $f_s = 48$ kHz, uniformly with a resolution of $\Delta T_p = 1$. We propose to reduce the resolution depending on which of the five octave O_i $(i = 1, 2, 3, 4, 5)$ the pitch lag falls (Table 2). We then implement five different quantization schemes. In Table 2 we show, for each method, the resolution used to code the pitch lag depending on the octave.

In our approach, the coarser quantization is implemented directly after the current pitch lag estimator and the gain is then chosen accordingly to the new pitch lag without any algorithmic modification. It should be noted that, since the distance between notes is logarithmic with respect to the fundamental frequency, using a resolution of $\Delta T_p = 1$ when $T_p \in O_1$ should be equivalent to using a resolution of $\Delta T_p = 16$ when $T_p \in O_5$. This is done in in **CTP1**, where 5 bits (32 levels) are always used independently of the octave.

Table 2: Resolution ΔT_p used to code T_p in the different methods implemented depending on the octave.

METHOD	O_1	O_2	O_3	O_4	O_5
ORIG	1	1	1	1	1
CTP1	1	2	4	8	16
CTP2	1	1	2	4	8
CTP3	1	1	1	2	4
CTP4	1	1	1	2	2
CTP5	1	1	1	1	2

3 Experimental Evaluation

We present some of the results of the experimental analysis. The experiments were done on stereo audio signal at $f_s = 48$ kHz. The CELT coder was run with a frame size of $N = 256$ samples (5.3 ms) and with 42 bytes per packet, producing a rate of 63 kbit/s. This experimental setting was chosen as the gain in quality when the prefilter and postfilter are used is higher and it is easier to spot degradations in audio quality. Furthermore, at this rate, the CELT coder was usually not able to compensate for eventual mistakes in the filtering operations.

The experimental evaluation was done with accurate listening comparisons and with objective PEAQ evaluation for which we present the results below. We have divided the experimental results into five sub-groups for clarity: low to mid pitched audio (Table 3), mid to high pitched audio (Table 3), audio without harmonic structure (Table 3), and speech (Table 3 for male, Table 3 for female). A summary of the results with the average between all the analyzed files is given in Table 3. For configuration we show the average PEAQ score, its difference with the original method and the worst-case difference in PEAQ score for the analyzed files. As a general remark, the experimental evaluation has focused especially on low to mid pitched audio (9 files for about 10 minutes of total duration) as it would be the one more affected by the reduction in accuracy (see Table 2).

Table 3: Low to Mid Pitched audio files ($T_p \in [128, 1024]$). Nine files have been considered with a total duration of about 10 minutes. Instruments played include: saxophone, bass, trombone, tuba. Music styles are both solos and ensembles.

METHOD	PEAQ score	Δ	Δ_{wc}
ORIG	-2.7778	-	-
CTP1	-2.8307	-0.0529	-0.3309
CTP2	-2.7888	-0.0110	-0.1219
CTP3	-2.7771	+0.0007	-0.0459
CTP4	-2.7769	+0.0009	-0.0221
CTP5	-2.7772	+0.0006	-0.0007

Table 4: Mid to High Pitched audio files ($T_p \in [32, 127]$). Five files have been considered with a total duration of about 4 minutes. Instruments played include: clarinet, flute, piano. Mostly solo music.

METHOD	PEAQ score	Δ	Δ_{wc}
ORIG	-3.1394	-	-
CTP1	-3.1958	-0.0563	-0.1996
CTP2	-3.1552	-0.0157	-0.0643
CTP3	-3.1430	-0.0035	-0.0279
CTP4	-3.1400	-0.0006	-0.0153
CTP5	-3.1397	-0.0003	-0.0024

Table 5: Audio files without harmonic structure (percussion instruments). Three files have been considered for a total duration of about 2 minutes. Only solo music.

METHOD	PEAQ score	Δ	Δ_{wc}
ORIG	-1.9972	-	-
CTP1	-1.9916	+0.0056	-0.0240
CTP2	-1.9871	+0.0101	-0.0086
CTP3	-2.0164	-0.0192	-0.0399
CTP4	-2.0136	-0.0164	-0.0386
CTP5	-2.0004	-0.0032	-0.0135

Table 6: Male speech (original sampling frequency is $f_s = 48$ kHz) $T_p \in [120, 440]$. Two files have been considered for a total duration of about 2 minute.

METHOD	PEAQ score	Δ	Δ_{wc}
ORIG	-1.8176	-	-
CTP1	-1.8928	-0.0752	-0.0932
CTP2	-1.8197	-0.0022	-0.0032
CTP3	-1.8093	+0.0012	-0.0082
CTP4	-1.8106	+0.0019	-0.0069
CTP5	-1.8176	-0.0001	-0.0001

Table 7: Female speech (original sampling frequency is $f_s = 48$ kHz) $T_p \in [85, 208]$. Three files have been considered for a total duration of about 2 minutes.

METHOD	PEAQ score	Δ	Δ_{wc}
ORIG	-2.6565	-	-
CTP1	-2.7535	-0.0970	-0.1626
CTP2	-2.6928	-0.0363	-0.0494
CTP3	-2.6667	-0.0102	-0.0300
CTP4	-2.6553	+0.0012	-0.0126
CTP5	-2.6475	+0.0090	-0.0029

Table 8: PEAQ score for all the analyzed files.

METHOD	PEAQ score	Δ	Δ_{wc}
ORIG	-2.6674	-	-
CTP1	-2.7176	-0.0502	-0.3309
CTP2	-2.6783	-0.0109	-0.1219
CTP3	-2.6715	-0.0041	-0.0459
CTP4	-2.6692	-0.0018	-0.0386
CTP5	-2.6667	-0.0007	-0.0135

4 Conclusion

This document provides experimental evidence that a coarser quantization of the pitch lag is possible. Several careful listening comparisons have been performed and the degradation in perceptual quality from a subjective point of view is insignificant. The PEAQ tests confirm the results of the listening comparisons showing also very little degradation. Furthermore, in some cases, a coarse quantization actually resulted in improved quality. In several audio samples, it can be seen clearly that the new pitch values help achieving smoother pitch contour in the output, increasing the perceptual quality of the coded signal and compensating for small errors in the pith lag estimation. The quantized values have a lower dynamic range, offering a temporal smoothing of the pitch lag values and, as a consequence, better spectral dynamics in the output.

We would like to recommend the implementation of the quantization scheme **CTP2**, since the difference in audio quality between this method and the other less aggressive quantization methods (**CTP3**, **CTP4**, **CTP5**) is not relevant. **CTP1** might be a bit too aggressive, even if, after carefully listening through the output files, it was impossible to notice relevant distortions in any of them.