A CONTINUOUS MODULATED SINGLE SIDEBAND BANDWIDTH EXTENSION

Frederik Nagel†, Sascha Disch‡, Stephan Wilde†
† Fraunhofer IIS, Erlangen, Germany
‡ Leibniz Universität, Hannover, Germany

ABSTRACT

Bandwidth extension (BWE) is an important parametric technique applied by modern audio coders in order to achieve efficient data rate compression at low bitrates. The perceptual quality of BWE enhanced signals is, however, often hampered by artifacts caused by inharmonicity. We propose hence a bandwidth extension method that avoids inharmonicity, and, at the same time, avoids the costly transmission of additional control parameters for frequency shifts. Harmonicity of the decoded signal is ensured by calculation of the autocorrelation function of the magnitude spectrum. The proposed bandwidth extension method is implemented by single sideband modulation (SSM). Fitting nicely in the general scheme of well established spectral band replication (SBR), the new method has some potential to replace the SBR patching algorithm. Potentially, the SSM can additionally govern the subsequent spectral shaping. Listening test results demonstrate an advantage of the novel scheme compared to SBR as used in high efficiency advanced audio coding (HE-AAC).

Index Terms— Bandwidth extension, audio coding, modulation, spectral autocorrelation

1. INTRODUCTION

Contemporary audio coders solely code the lower spectral part of the signal. Transform coders, such as high efficiency advanced audio coding (HE-AAC) [1], apply a waveform preserving coder; others that mainly rely on a dedicated source model, such as the adaptive multi rate wideband plus (AMR-WB+) [2] coder, employ a prediction model for the lower frequencies. Both have in common that the higher frequencies (HF) are estimated from the lower frequencies (LF) in the decoder driven by additional side information parameters. Subsequently, the spectral envelope of the signal is shaped and tonality is adapted to resemble closely the original. For obtaining the HF, the first uses single sideband modulation (SSM) with fixed modulator, implemented as a copy operation within a quadrature mirror filterbank (QMF) representation, while the latter derives a mirrored upper copy of the lower spectrum by interpolation. [3, 4] give an overview of common bandwidth extension techniques for speech and music.

The algorithms used in HE-AAC and AMR-WB+ do not ensure preservation of the harmonicity of the original signal. This may lead to audible artifacts [5]. Figure 1 shows an exemplary outcome of the patching algorithm of SBR applied to a pitch pipe signal. At the HF-crossover frequency, the inharmonicity problem is clearly visible.

Fig. 1. Amplitude spectra of original (left) and bandwidth extended signal (right) using the SBR of HE-AAC. Audible artifacts occur at the LF cutoff frequency 4 kHz because tonal components are copied into spectral vicinity of each other.

2. HARMONIC PATCHING

2.1. Background

Several solutions have been proposed with the aim to preserve harmonic structure, either based on the estimation of the fundamental frequency \( F_0 \in \mathbb{N} \) (e.g. [6, 7], see also [4]) or by using a mechanism that preserves harmonicity inherently. The latter can be accomplished by applying a non-linear function such as clipping or rectification to the LF, but often at the drawback of introducing unwanted noisy artifacts.

We recently proposed a solution to harmonically extend the bandwidth of primarily audio signals [5], the harmonic bandwidth extension (HBE). HBE is based on phase vocoders to stretch the frequency scale of the band limited LF signal. The method has been shown to be beneficial for many types of music signals, but was still inferior to SBR for the coding of speech [5].

2.2. Continuously modulated bandwidth extension

In this paper, we propose an alternative BWE method that operates in time domain based on single sideband modulation (SSM) to generate the HF information. The modulator is adapted to the signal such that harmonicity is preserved. The adaptation of the modulator does not rely on an accurate estimation of \( F_0 \). One or more fundamental frequencies contained in the signal are considered implicitly by calculating the maximum of the autocorrelation function of the magnitude spectrum. In contrast to BWE methods that rely on pitch tracking as used for instance in [6], the new method is also useful for music or when more then one person talks simultaneously.

The modulation function that controls the SSM of the low band is defined to be frequency modulated (FM) and, additionally, amplitude modulated (AM). Since the method employs modulators...
that are continuously defined across successive temporal blocks, it has been termed continuously modulated bandwidth extension (CM-BWE). Figure 2 provides an overview of the processing steps that are carried out in CM-BWE.

2.3. Theory

Harmonic patching is only necessary, if the signal has a harmonic structure. Therefore, initially the tonality of the signal is calculated as the spectral flatness measure (SFM). No harmonic continuation is desired if the flatness is above a certain threshold. In that case, no frequency modulation of the modulator is necessary and it is kept constant. In the following, we concentrate on harmonic signals.

Let $s(t)$ be a time domain signal, $t \in \mathbb{R}$. It always can be written as sum of sinusoids plus residual $e(t)$ as

$$s(t) = \sum_k \alpha_k \sin(t \cdot \omega_k + \phi_k) + e(t)$$

with frequencies $0 \leq \omega_k \leq 2\pi$, phases $\phi_k < 2\pi$, and amplitudes $\alpha_k \in \mathbb{R}$; $k \in \mathbb{N}$. Every signal can additionally be considered w.r.t. as a superposition of harmonic series of sinusoids, i.e.

$$s(t) = \sum_l \sum_k \alpha_{k,l} \sin(t \cdot \omega_k + \cdot \phi_k) + e(t)$$

with $l \in \mathbb{N}$, $\alpha_{k,l} \in \mathbb{R}$. Let us consider a fixed $l$ and, in the following, focus on the tonal part of the signal that is expressed by sinusoids. One harmonic series consists of a fundamental frequency $F_0$ and its integer multiples $S := l \cdot F_0$. If we shift a part of the spectrum, as done in SBR, a fixed frequency offset $LF_{max}$ is added to the sinusoids, resulting in shifted frequencies $\tilde{S} := l \cdot F_0 + LF_{max}$. This offset is defined to be equal to the crossover frequency between LF and HF part to be generated.

Fine tuning of this offset is needed to recover the original harmonic structure. Therefore, one has to find the minimal $\delta$ that solves the equation

$$i \cdot F_0 = j \cdot F_0 + LF_{max} + \delta; i, j, \delta \in \mathbb{N}, \delta < LF_{max}$$

(3)

It is straightforward that this applies for $\delta = k \cdot LF_0 - LF_{max}$. However, if $F_0$ is unknown, $\delta$ cannot be determined. We therefore propose, instead of estimating $F_0$, to calculate the cross-correlation of the original and the decoded HF spectrum. Since the original HF spectrum is not available at the decoder side, the correlation has to be computed on the LF part by autocorrelation as illustrated in Figure 2b,c. We therefore correlate the HF spectrum with itself but shifted by $LF_{max}/2$.

$$i \cdot F_0 = j \cdot F_0 + \frac{LF_{max} + \delta}{2}$$

(4)

This has the general advantage that, when we observe a superposition of several sinusoids, we are able to identify the $\delta$ with

$$\delta := 2 \cdot \arg \max_{d \in \mathbb{Z}} \frac{\arg \max_{i \in \mathbb{N}} (i \cdot F_0 \leq LF_{max}) - \frac{LF_{max} + \delta}{2}}{2}$$

(5)

In the following, we consider digital signals and write $s[n]$ with $n \in \mathbb{N}$ instead of $s(t)$. We can now solve the equation 4 indirectly without explicit knowledge of $F_0$. The solution for formula 5 is given by the lag of the biased autocorrelation of the magnitude spectrum $S = |F(s)|$

$$R[\nu] := b[\nu] \sum_{\nu_{min} \leq \nu \leq \nu_{max}} S[\omega] \left| S \left[ \omega + \frac{LF_{max}}{2} + \nu \right] \right|$$

with $\nu \in \mathbb{Z}$ at its maximum

$$\delta = \arg \max_{\nu \in \mathbb{Z}} R[\nu]$$

(7)

The bias is chosen by changing the $\beta$s in the weighting function

$$b[\nu] = \frac{1}{|F_{max} - |\nu \nu|}; 1 \geq \beta \in \mathbb{R}$$

(8)

whereas $\beta \approx 0$ leads to a preference for small lags and $\beta \approx 1$ weights all lags equally. One can decide, whether small lags are preferred by using a biased autocorrelation ($\beta = 0$), all lags are handled equally ($\beta = 1$), or anything in between. If there are multiple local maxima of the cross-correlation function, as depicted in Figure 3, the choice of a minimum delta having a negative sign is preferred in order not to produce a gap between LF and generated HF part.

The frequency offsets $\delta$ are used to produce the actual patches as shown in Figure 2d. Subsequently, spectral envelope shaping has to be accomplished. This can be implemented either by shaping the entire patches (Figure 2e), or by scale factors that cover smaller frequency ranges (Figure 2f). The latter, for instance, is utilized by SBR.

Next, the offsets $\delta$ are interpolated, median and lowpass filtered yielding a function $\delta[n]$. Moreover, a hysteresis is introduced, to ensure temporal stability. The final result is a continuously varying modulator to be applied to the entire LF signal. The modulator can be amplitude modulated using frequency dependent weights or a frequency shaping function $\alpha[n]$ in order to additionally restore the spectral shape of the HF signal in an integrated way. By carrying out the resulting SSM, the signal is then shifted in frequency and shaped in its spectral envelope simultaneously. Figure 4a shows a modulator derived by this scheme.

Let the modulator with interpolated shifts $\delta[n]$ be $\mu[n] \in \mathbb{C}$. The analytical signal of the LF part is generated by Hilbert transformation and multiplied with the modulator to produce the bandwidth extension. In order to obtain all patches simultaneously, multiple analytic modulators are added with

$$\mu_k[n] = \alpha_k[n] \cdot \exp \left( \frac{2\pi i}{L} \sum_{l=0}^{n} k \cdot (LF_{max} + \delta[l]) \right)$$

(9)

for patch $k \in \mathbb{N}$ of a signal with sample rate $F_s$. The formula for multiple analytic modulators is accordingly $\mu[n] = \sum_{k=1}^{K} \mu_k[n]$ for $K \in \mathbb{N}$ patches.

Signals may have a DC part or contain LF information of which a transposition in not desired. If the patching should only copy higher frequencies, applying a highpass filter with a lower cutoff frequency $LF_{min}$ is recommended. In formulas 6 and 9, $LF_{max}$ has therefore to be replaced by $LF_{max} - LF_{min}$ with $LF_{min}$. Additionally, the signal has to be highpass filtered prior to examining formula 6.

Lastly, a point-wise multiplication yields the HF time signal $\tilde{s}[n]$ derived from the band limited LF signal $s[n]$ with $K \in \mathbb{N}$ patches:

$$\tilde{s}[n] = \Re(s[n] + i \tilde{\Im}(s[n])) \cdot \mu[n] + e[n]$$

(10)

Lastly, this results in a reconstructed spectrum for each signal, as displayed in Figure 5. In our implementation, however, envelope shaping and noise addition are implemented by usage of the SBR tools. Hence, in equation 9 $\alpha_k \equiv 1$ and in 10, $e \equiv 0$.

The algorithm is carried out on arbitrary temporal positions in the signal in order to correctly align the patch to the LF part. A natural choice would be the core coders block rate. Figure 4 shows
Fig. 2. Steps of bandwidth extension with CM-BWE. (a) Original; (b) Calculating autocorrelation function; (c) Ascertain lag for maximal correlation; (d) Patching by copying with adjusted offset $\delta$; (e) Spectral shaping with one factor per band; (f) Shaping with scale factors with higher resolution.

Fig. 3. Autocorrelation function of the magnitude spectrum of the decoded signal.

the modulator and the resulting signal. Compared to the result of plain SBR, as shown in Figure 1, one can easily see the improved harmonic continuation of the spectrum rather than a mismatch at the LF-cutoff frequency. This lastly results in a reconstructed magnitude spectrum for each signal as displayed in Figure 5.

This algorithm is less complex compared to HBE [5] in which multiple phase vocoders have to be calculated in parallel. The most complex parts of this algorithms are the calculation of the analytical signal and the maxima of the autocorrelation function. Here, one has to find a trade-off between accuracy and complexity.

3. LISTENING TEST

The continuous modulated single sideband bandwidth extension is compared to two alternative BWE, SBR [8] and HBE [5]. For the listening test, we used two different core coders that were incorporated within a unified speech and audio coder (USAC) [9]. USAC uses SBR for speech and HBE for music at 16 kHz mono. We replaced the patching algorithm in USAC for both, speech and music, by CM-BWE. Additionally we applied HBE patching for speech as well. The HE-AAC core coder inside USAC which was used for music content and one mixed item had a bandwidth of $LF_{\text{max}} = 4$ kHz while the prediction coder had a bandwidth of 6.75 kHz. The signals were bandwidth extended by either SBR, HBE, or CM-BWE to 13 kHz for 16 kbit/s; input sample rate was 48 kHz. Only the patching was replaced, both envelope shaping and tonality correction were carried out by SBR.

11 expert listeners participated in the listening test. Ratings were given on a MUSHRA scale [10]. The test consisted of 8 items, 1...
voiced a capella item, 4 music items and 3 items with mixed content (speech + music). The items were presented alongside the original and one lowpass filtered anchor with bandwidth 3.5 kHz. Sound was replayed from a fanless computer equipped with a professional sound card. Stax headphones and amplifier were used.

4. RESULTS

Comparing the absolute ratings as proposed in [10] yielded no significant differences. The difference ratings with its means and confidence intervals were calculated for every item and every listener pairwise between SBR, HBE, and CM-BWE. Figure 6 presents the results of this calculation. Statistical evaluation on the basis of overlapping confidence intervals show that CM-BWE items were rated significantly better ($p < .05$) than SBR for three of eight items. CM-BWE was inferior to HBE for one item. Neither CM-BWE nor HBE are capable of outperforming SBR for speech and a capella voice.

A Bonferroni correction was applied, as multiple testing can easily lead to false positive results. The level of significance was therefore divided by the number of tests $\bar{p} = p/8 = 0.000625$. A paired $t$-test on this basis showed significant improvement for CM-BWE for one case and for HBE for two cases compared to SBR. A non-parametrical Wilcoxon signed-ranks test additionally evidenced that both CM-BWE and HBE overall outperform SBR while HBE is preferred to CM-BWE ($p < .01$).

The current implementation improves SBR by simply replacing the patching algorithm inside SBR. It could not reach the quality of HBE, in particular for one item, but demonstrates the importance of maintaining harmonic structures. An advantage compared to HBE as presented in [5] is a reduced complexity. The CM-BWE can easily substitute patching in SBR, even in existing applications as it can operate without any additional side information.

6. REFERENCES