HIGH-FREQUENCY TONAL COMPONENTS RESTORATION IN LOW-BITRATE AUDIO CODING USING MULTIPLE SPECTRAL TRANSLATIONS

Imen Samaali[1], *Gaël Mahé*[2], *Monia Turki-Hadj Alouane*[1]

[1]Unité Signaux et Systèmes (U2S), Université Tunis El Manar, ENIT, Tunisia
[2] Laboratory of Informatics Paris Descartes (LIPADE), Université Paris Descartes, France
email: imen.samaali@yahoo.fr, gael.mahe@parisdescartes.fr, m.turki@enit.rnu.tn

ABSTRACT

At reduced bitrates, the audio compression affects high frequency tonal components of signals, which results in a roughness phenomenon. Audio coders are limited in the reconstruction of the high-frequency spectrum mainly because of the potential unpredictability of the structure of the latter, as well as unprecise indicators of tonal to noise ratio. We propose a technique for high-frequency tones restoration, based on the correction of the tonal positions in the decoded signal, using a small set of information transmitted through an auxiliary channel at a very low bit-rate (typically < 2 kbps). The proposed approach is evaluated using objective measures of perceptual roughness. The experimental results with HE-AAC coding at 16 kbps exhibits an efficient preservation of the harmonicity and a significant improvement of the audio quality.

1. INTRODUCTION

In perceptual audio coders, coding at low bit-rates worsens the quality of audio signal. Under 96 kbps for MP3 codec (mono) and 64 kbps for AAC codec (mono), the quantization noise generated by the encoder exceeds the masking threshold and thus generates audible artifacts [1]. To keep a transparent audio quality at reduced bit-rates, several coding schemes have been proposed, including bandwidth extension techniques like Spectral Band Replication (SBR) [1,2]. The latter has been combined with the AAC coder to create MPEG-4 High Efficiency AAC (HE-AAC), also called aacPlus [2].

The SBR technique takes advantage from the high correlation between low and high frequency in audio signals, to reconstruct the high frequency band from the low frequencies. The principle of SBR is to replicate in the high frequencies the fine structure of the low-frequency spectrum and to reshape it thanks to additional parameters transmitted at a low bitrate (1 to 3 kbps), namely the frequency envelope and the tone to noise ratio of the high-frequency band. Hence, only the low frequency band and those parameters need to be coded.

This work is part of the WaRRIS project granted by the French National Research Agency (project n° ANR-06-JCJC-0009) and was supported by the franco tunisian CMCU project n° 08S1414 .

The SBR technique associated with a perceptual audio coder reduces efficiently the number of bits needed to encode the high frequency band, while maintaining a decoded signal perceptually similar to the original one. However, the way of generating the high-frequencies fine structure, consisting in multiple translations of the low-frequency sub-bands, causes major disadvantages.

First, the reconstruction of the high-frequency band does not ensure the preservation of the harmonicity of the original signal. Fig. 3 a-b show an exemplary result of the high-frequency band generated by SBR applied to a trumpet signal. In the high-frequency band of the coded-decoded signal, the inharmonicity problem is noticeable. This may lead to an audible artefact known as roughness [3]. According to [4], a roughness is perceived if the frequency difference between two tones is between 20 and 200 Hz.

Secondly, the reconstruction of the high frequencies is not suitable for non-harmonic tonal signals, for which the SBR generates tonal in high frequencies completely different from the original ones and even new tonal may appear.

In order to enhance the audio quality, some techniques with different complexities were proposed in the literature. The harmonic bandwidth extension (HBE) [5] is based on multiple spectral stretching operations using phase vocoders operating in parallel in order to generate the high-frequency band. The HBE technique has been found to be interesting for reducing roughness. However, the technique has two major drawbacks:

- For harmonic signals with a percussive character like guitar, HBE may induce pre- and post-echoes artifacts [5].
- For strongly harmonic signals like violin, some high-frequency harmonics may not be generated, which results in a non-preservation of the harmonicity.

With the objective of preserving the harmonicity of audio signal, Nagel [6] proposed a second bandwidth extension method called continuously modulated bandwidth extension (CM-BWE) which generates HF information by single sided modulation in time domain. The modulator is adapted to the signal such that the harmonicity is preserved. However, isolated tonal components may appear for non harmonic tonal

978-0-9928626-3-3/15/$31.00 ©2015 IEEE

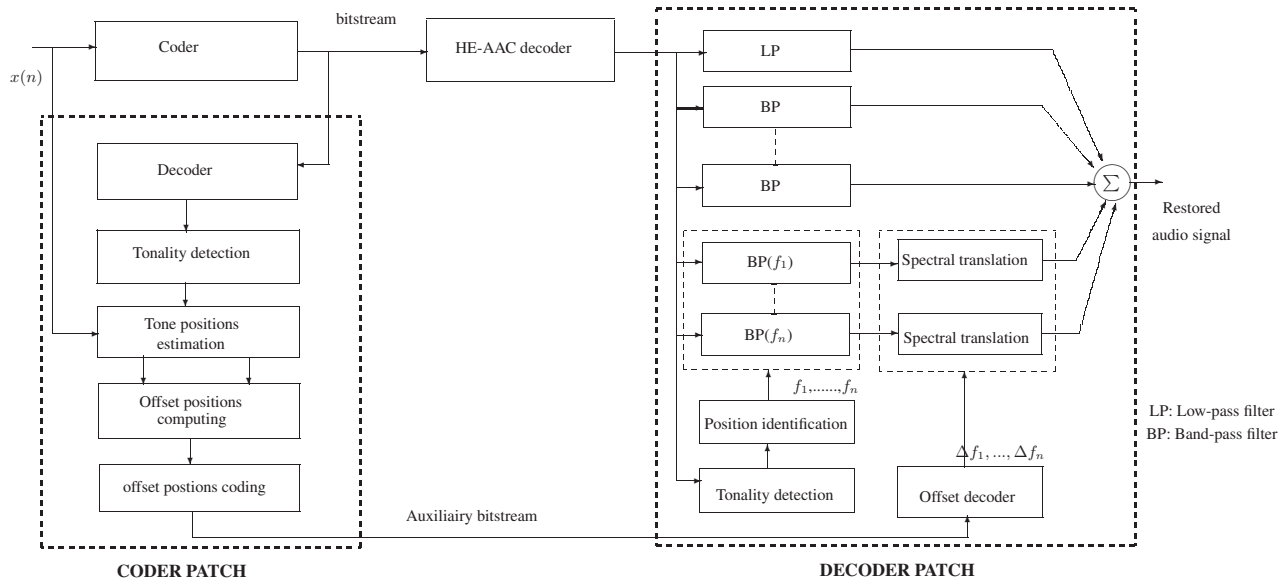


Fig. 1. Block diagram of the proposed approach for tonal frames.

signals like glockenspiel.

In this paper, we propose a novel method aiming at preserving harmonicity and restoring isolated tonal components. The idea is to correct the tonal components positions of the coded-decoded signal, using a small set of parameters transmitted over a very low bit-rate (< 2 kbps) auxiliary channel. Hence, the proposed method is conceived as an external patch that does not change the coder itself. It only needs an auxiliary channel, that can be provided without additional bit-rate by a watermark, given its reduced rate of information.

This paper is organized as follows: in Section 2, we present the new approach dedicated to tonality correction of SBR coded-decoded signals. Section 3 presents a performance evaluation of the proposed algorithm.

2. PROPOSED TECHNIQUE

The proposed restoration approach, depicted in Fig. 1, constitutes a post processing after the decoder. Parameters related to the tonal component positions in the original signal are extracted at the encoder and transmitted to the decoder through an auxiliary channel. In order to minimize the bitrate of this channel, we propose to transmit the frequency offset Δf between each tone position detected on the original signal and its equivalent detected on the encoded-decoded one, instead of transmitting the positions. Therefore, we need to perform "blank" decoding at the encoder.

At the decoder, the tonal positions $f_1 \dots f_n$ synthesized by the SBR decoder are corrected by multiple spectral translations using the respective transmitted and decoded offsets $\Delta f_1, \Delta f_2, ..., \Delta f_n$.

In the following, we will describe more accurately each component of the proposed system, referring to Fig. 1.

2.1. Tonality detection

One way to determine the noise-like or tone-like nature of a signal is to calculate its Spectral Flatness Measure (SFM) [7], defined as the ratio between the geometric mean G_m and the arithmetic mean A_m of the power spectrum $|X(k)|^2$ (where $X(k)$ stands for the Discrete Fourier Transform of the signal):

$$SFM_{dB} = 10 \log_{10} \left(\frac{G_m}{A_m} \right), \quad (1)$$

where:

$G_m = \sqrt[N]{\prod_{k=0}^{N-1} |X(k)|^2}$ and $A_m = \frac{1}{N} \sum_{k=0}^{N-1} |X(k)|^2$.

From the SFM, one can derive the coefficient of tonality:

$$\alpha = min \left(\frac{SFM_{dB}}{SFM_{min}}, 1 \right), \quad (2)$$

where $SFM_{min} = -60$ dB corresponds to pure tones.

The values of the coefficient of tonality α are in the range of [0, 1], where 0 is the value for pure noise and 1 is the value for a pure tone. The coefficient of tonality α is compared to a threshold τ to make a final decision. Based on exhaustive empirical measures, the value of τ is fixed at 0.2. Thus, each frame is considered as tonal if $\alpha > 0.2$, as noise otherwise.

2.2. Tonal component position detection

The method used to detect tonal positions is similar to the one described in the MPEG-1 standard [8]. In the first step, the peaks (local maxima) are identified on the power spectrum $|X(k)|^2$ previously smoothed by a median filter aiming at reducing the number of peaks. A frequency component $X(k)$ is considered as a peak if it is greater than its immediate neighbors $(k \pm 1)$ and if it exceeds by 4 dB its other neighbors

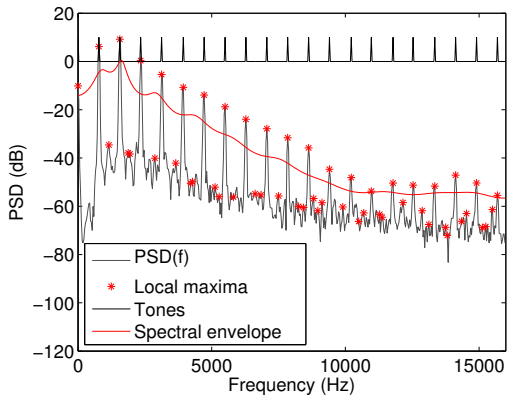
distant of less than a given value Δ_{peak}. These conditions can be expressed as folllows:

$$X(k) > X(k-1) \quad (3)$$
$$X(k) \geq X(k+1) \quad (4)$$
$$X(k) - X(k+j) \geq 4 \text{ dB} \quad \forall\, j \in \{-3,-2,3,2\}, \quad (5)$$

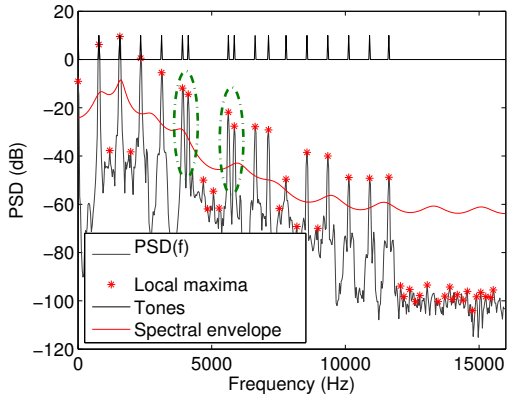
where k represents the discrete frequency.

In the second step, the non-tonal peaks are discarded by thresholding. The considered threshold is an estimate of the spectral envelope by an autoregressive model of order p. The spectral envelope must be smooth enough to provide a general shape of the energy distribution of the signal. For this reason, we chose a prediction order $p = 15$.

To illustrate the effectiveness of the proposed approach, Fig. 2 shows the tone positions detected by our algorithm on a trumpet sequence sampled at 44.1 kHz and its coded/decoded version with HE-AAC at 16 kbps sampled at the 32 kHz. The proposed method reduces significantly the unnecessary local maxima, on both spectra. In addition, close tones due to erroneous SBR (see Fig. 2 b around 4 and 5.6 kHz) are correctly detected.



(a)



(b)

Fig. 2. Tones identification by the proposed algorithm performed: (a) on the original signal, (b) on the coded/decoded signal at 16 kbps.

2.3. Tonal position coding

Once the tone positions are estimated, they must be coded and transmitted through the auxiliary communication channel. For the HE-AAC decoder at 16 kbps for instance, the high-frequency band synthesized by SBR is 4-11.7 kHz, so that coding each tone position accurately on such a range of values would require a high bitrate. To reduce the information rate transmitted to the decoder, we propose to carry out a "blank" decoding process at the encoder and transmit the differences between the original tone positions and those detected on the decoded signal. The offset vector will be noted Δf. The latter is of a variable size, depending on the number of tones in the replicated band.

To determine Δf, each tone of the encoded-decoded SBR band is matched with the nearest tone of the original signal, which must be matched with only one tone of the coded-decoded signal, the closest one. For the decoded tones matching no tone in the original signal, a special value is fixed in Δf (see encoding step), indicating to remove the tone. The tones of the original signal with no equivalent in the decoded signal are not treated.

The components of vector Δf are coded according to the following two steps:

- First, they undergo a uniform scalar quantization on 2^n values (n to be set) in a range $[-f_0, f_0[$ depending on the nature of the signal:
  - for harmonic signals, f_0 is the fundamental frequency;
  - for tonal signals, f_0 is the maximum error on tone positions caused by SBR.
- In a second step, the quantized values of Δf are coded on n bits according to a Gray coding, in order to limit the impact of a bit error in the auxiliary channel (from the perspective of using the watermarking as an auxiliary channel). As the difference in position between harmonics can never reach the value of the fundamental frequency, the code representing $-f_0$ will be used to encode the indication of tonal removing.

Considering the reference note A at 440 Hz and the band 4-11.7 kHz to be corrected, a maximum of 19 tonal positions may be coded per frame of 46 ms. Hence, setting n = 6 in each frame leads to a bitrate of about 2 kbps for the auxiliary channel.

2.4. Spectral translation for tonal components correction

The correction of tonal positions is based on spectral translations according to the offset Δf transmitted and decoded. Using a non-regular filter bank, the decoded signal is divided into sub-bands according to the tonal positions f_i detected in

Audio signal	Original	coded-decoded	Restored
Trumpet	107.49	103.99	108.14
Violin	82.21	49.9	56.74
Pipe	84.96	46.34	83.2
Harmonica	48.56	64.44	57.11
Bagpipe	13.94	21.18	14.34

Table 1. Objective evaluation of the performance of the proposed system.

the decoded signal. We define three types of sub-bands:

- the low-frequency band fully transmitted by the codec;
- tonal high-frequency sub-bands of width of 100 Hz centered around the tone frequencies;
- high-frequency sub-bands of various widths corresponding to the remaining (non-tonal) spectrum.

Only the sub-bands of second type need to be processed.

In each sub-band containing a tonal component, the tone position correction is based on a single sideband modulation (SSB). Let $x_i(t)$ be time domain signal of the sub-band containing f_i. The frequency-translated signal, $y_i(t)$, with tone f_i translated of Δf_i, is given by [9]:

$$y_i(t) = \Re[x_i^a(t)\exp(\mathrm{j}2\pi\Delta f_i t)] \tag{6}$$

where $\Re$ denotes the real part and $x_i^a(t)$ is the analytic signal corresponding to $x_i(t)$, defined by:

$$x_i^a(t) = x_i(t) + \mathrm{j}\mathcal{H}\big(x_i(t)\big)$$

where $\mathcal{H}$ denotes the Hilbert transform.

3. EXPERIMENTAL EVALUATION OF THE PROPOSED APPROACH

The experimental evaluation was performed on five sequences of mono audio signals, from QUASI database [1], that exhibit a remarkable harmonic character. All the considered original signals are sampled at 44.1 kHz and their decoded versions are sampled at 32 kHz . The extension band encoder used is the standard version of HE-AAC encoder (aacPlus). This version offers compression rates ranging from 8 to 160 kbps, with a transparent quality at 24 kbps in mono. The considered rate is 16 kbps. The parameters of the offset vector Δf were coded on 6 bits and transmitted by a low bitrate auxiliary channel (less than 1 kbps), which corresponds to a maximum of 8 tonal positions coded and corrected per frame.

As a primary evaluation of the proposed system, we computed the spectrograms of the signals in three versions: original, coded-decoded and restored (see Fig. 3 and 4). In the coded-decoded trumpet signal (Fig. 3 b), a non harmonic spectrum appears from the sixth tonal around 4 kHz and extends up to 8 kHz. These components come into dissonance and generate a perceptual artefact which can be heard as a buzzing sound. A clear correction of the harmonicity is observed on the restored version of the signal. For the glockenspiel, we note on the decoded signal (Fig. 4 b) isolated synthesized tonals different from the tonal components on the original one (eg tonal framed by the dotted rectangle). Although the analyzed signal is highly non-stationary, a correction of some tonal components is verified in Fig. 4 c.

The audio quality of HE-AAC is evaluated through objective measures provided by PEAQ software (Perceptual Evaluation of Audio Quality) based on the ITU BS.1387 standard. However, the measurements obtained by the free version of PEAQ [2] does not coincide with the measures presented in the literature and confirmed by the listening tests: a transparent quality at 24 kbits/s. Thus, for harmonic signals, the performance of the harmonicity correction were evaluated through the roughness measurement provided by the SRA software [10], which provides an objective evaluation of the perceived impairment due to the loss of harmonicity. For each frame, the measure is based on a list of tonal components with frequency and amplitude (f_i, A_i). For each possible pair of components (*i,j*), the partial roughness is defined as:

$$r_{i,j} = X^{0.1} * 0.5(Y^{3.11}) * Z \tag{7}$$

where

$$\begin{cases} X = A_{\min} * A_{\max} \\ Y = 2A_{\min}/(A_{\min} + A_{\max}) \\ Z = e^{b_1 s(f_{\max} - f_{\min})} - e^{b_2 s(f_{\max} - f_{\min})} \end{cases}$$

where $A_{\min} = \min\{A_i, A_j\}$; $A_{\max} = \max\{A_i, A_j\}$; $f_{\min} = \min\{f_i, f_j\}$; $f_{\max} = \max\{f_i, f_j\}$; $b_1 = 3.5$; $b_2 = 5.75$; $s = 0.24/(s_1 f_{\min} + s_2)$; $s_1 = 0.0207$ and $s_2 = 18.96$.

All partial roughnesses are then summed to provide the total roughness. The roughness is then averaged over frames. Note that this is an intrinsic value that highly depends on the nature of the signal.

We present in Table 1 the roughness values estimated for five strongly harmonic signals and their corrected versions by the proposed system. For each signal, the roughness of the restored version is closer to the original one than that of the non-restored version, particularly for the pipe sequence. Hence, the proposed solution corrects the harmonicity loss also from a perceptual point of view (assuming that the objective measure of roughness is reliable).

4. CONCLUSION

We have proposed a technique of harmonicity correction and tones restoration for bandwidth extension encoders, particularly the HE-AAC encoder. The proposed solution, dedicated

[1] http://www.tsi.telecom-paristech.fr/aao/en/2012/03/12/quasi/

[2] http ://www-mmsp.ece.mcgill.ca/Documents/Software/index.html

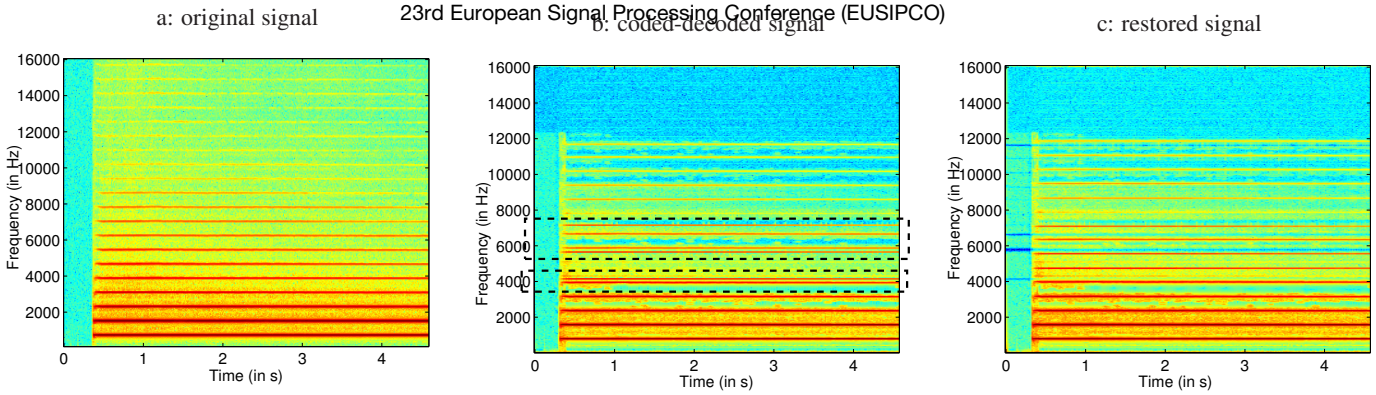


Fig. 3. Illustration of harmonicity correction using the proposed approach for trumpet signal.

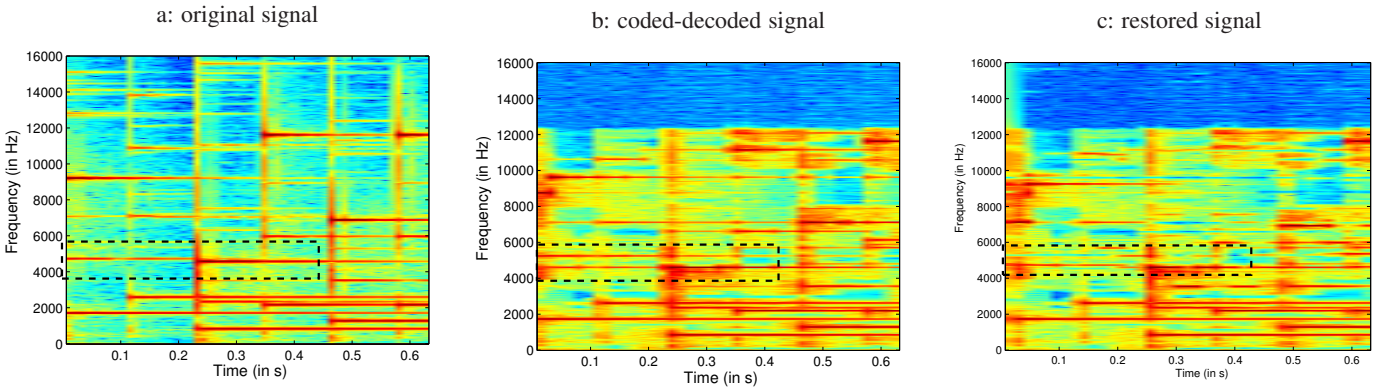


Fig. 4. Illustration of tonality correction using the proposed approach for glockenspiel signal.

to tonal and strongly harmonic audio signals, is based on frequency adjustment of a set of tonal components by multiple spectral translations. These translations are performed in the time domain via single sideband modulations combined with a filterbank, and using a small set of information transmitted through a low bitrate auxiliary channel.

The proposed system was evaluated for mono-instrumental sounds, both by the spectrograms observation and by an objective measurement dedicated to the roughness perception. The spectrograms show a good restoration of the tones positions and, for harmonic signals, the roughness measure indicates a significative quality improvement. Further studies will investigate this method for more complex sounds, particularly multi-pitch multi-instrument sounds.

REFERENCES

[1] K. Kjrling M. Dietz, L. Liljeryd and O. Kunz, "Spectral band replication, a novel approach in audio coding," in *Audio Engineering Society, 112th Convention*, 2002.

[2] ISO (2003), "Bandwidth extension, ISO/IEC 14496-3:2001/amd 1:2003. ISO. retrieved 2009-10-13," .

[3] A. Plomb and W. J. M. Levelt, "Tonal consonnace and citical bandwidth," in *Journal of the Acoustical Society of America*, 1965, pp. 548–560.

[4] V. Helmholtz, "On the sensations of tone," in *Acustica*, 1954, pp. 201–2013.

[5] F. Nagel and S. Disch, "A harmonic bandwidth extension method for audio codecs," in *ICASSP*, Taipei, 2009, pp. 145–148.

[6] F. Nagel, S. Disch, and S. Wilde, "A continuous modulated single sideband bandwidth extension," in *ICASSP*, Dallas, 2010, pp. 357–360.

[7] J. D. Johnston, "Transform coding of audio signals using perceptual noise criteria," in *IEEE Jour. Selected Areas Commun*, 1988, pp. 314–323.

[8] ISO/IEC, "Information technology coding of moving pictures and associated audio for digital storage media at up to about 1,5 mbit/s – part 3: Audio. ISO/IEC 11172-3:1993," in *Joint Technical Committee 1 Subcommittee 29 Working Group 11*, 1993.

[9] Chang Yu-Hsien, "Single sideband modulation assignment 1, digital audio systems, desc9115, semester 1," 2012, Faculty of Architecture, Design and Planning, The University of Sydney.

[10] Pantellis N. Vassilakis, " SRA: a web-based research tool for spectral and roughness analysis of sound signals," in *Proceedings of the 4th Sound and Music Computing (SMC) Conference*, 2007.