Joint Speech and Audio Coding Combining Sinusoidal Modeling and Wavelet Packets

Márk Fék†‡, Annamária R. Várkonyi-Kóczy†‡ and Jean-Marc Boucher†‡

†Budapest University of Technology and Economics, Dpt. of Measurement and Information Systems
Budapest, Műegyetem rkp. 9. H-1521, Hungary
fek@mit.bme.hu, koczy@mit.bme.hu
‡ENST de Bretagne, Dpt. Signal et Communications
Technopôle de Brest Iroise, BP 832, 29285 Brest CEDEX, France
JM.Boucher@enst-bretagne.fr

1. Introduction

Some applications require the encoding of both speech and generic audio (music) inputs. One example is Internet radio broadcasts, where the successions of commentator speech and music-recordings is transmitted. Another application is the digital archiving of already existing mixed speech and audio recordings, such as musical tales for children.

Speech and audio coding algorithms rely on two principles: – the use of a source model to eliminate redundancy in the sampled waveform representation; – the use of a perception model to eliminate perceptually irrelevant parts, which cannot be heard. State-of-the-art speech coding algorithms (CELP, parametric coders) use a speech specific source model, which does not apply for generic audio signals. Thus, these algorithms fail to encode music with good quality. State-of-the-art audio compression algorithms (MPEG family) use uniform transform coding and psychoacoustic masking models to eliminate perceptual redundancies. Long blocks are used for encoding stationary parts while short blocks are used for encoding transient parts of the signal. Applying long blocks to encode the rapidly varying speech signal leads to artifacts known as pre-echos. The oversegmentation of short blocks increases considerably the required bit rate.

One solution is to use separate speech and audio coders for the different types of input, which is feasible only if we have the two source types separately. Speech/music discrimination is used in [1] to select the specific encoding for a given input segment. Erroneous decisions, however, lead to coding artifacts on misclassified segments. The other solution, which we pursue in this paper, is to develop a joint algorithm that provides good quality for both speech and generic audio signals. The latter approach can be advantageous, even if we have a perfect speech/music discriminator. Namely, in the case of segments containing overlapping speech and music, a joint algorithm can provide better quality than applying a speech or music specific method to encode the mixed segment.

Our method is motivated by recently developed speech and audio coding algorithms. In [2], the authors propose a perceptually adapted Wavelet Packet Transform (WPT) to encode high quality speech signals. The specificity of the method is that it does not rely on a speech production model. Hence, it is suitable to encode generic audio signals. Music inputs often contain slowly changing sinusoids, for which the WPT does not provide a compact representation [3]. As a consequence, the bit rate required to encode such signals is rather high. To overcome this problem, [3] and [4] propose a combined Sinusoidal-WPT (S+WPT) model. The sinusoidal model can be considered as a source model for certain audio signals. The combined model extracts the sinusoidal components of the signal, then encodes the residual using a WPT.

In this paper, we present and evaluate a joint speech and audio coding algorithm combining sinusoidal modeling and WPT based coding. Our method differs from previous S+WPT methods in several respects. We use the WPT suggested in [2], as it is well suited to speech signals. Furthermore, we use the classical version of sinusoidal analysis, originally introduced in [5], reinforced by a Sinusoidal Similarity Measure (SSM), described in [6]. The SSM based sinusoidal analysis requires less computation than the methods used in [3] and [4]. Moreover, by adjusting the SSM threshold, we can control the extent of sinusoidal extraction. Hence, we can optimize the final bit rate by controlling how much of the signal is encoded as sinusoids, and how much is encoded by the WPT. To reduce the bit rate further, we use a novel pitch harmonics based encoding for sinusoidal frequency lines. Furthermore, we apply Perceptual Noise Substitution (PNS) [7] to encode noisy WPT sub-bands.

The paper is organized as follows: Section 2 presents the overall codec architecture. Section 3 describes the quantization and coding of the sinusoidal parameters, introducing the novel pitch harmonics based frequency encoding. Section 5 summarizes the encoding
of the residual by WPT. Section 6 describes the quantization and coding of WPT parameters, along with the PNS applied to noisy sub-bands. Section 7 presents the bit rate and quality results for various music and speech inputs.

2. CODEC architecture

![Codec Diagram]

Figure 1 shows the overall structure of the algorithm. The input is band-limited to 50-7000 Hz and sampled at 16 kHz using 16 bits per sample. The encoding and decoding works on a frame-by-frame basis. The encoder carries out a sinusoidal analysis to identify the stable sinusoidal components of the input. The frame size of the sinusoidal analysis is 512 samples with an overlap of 256 samples between two consecutive frames.

A masking model is also calculated, based on the MPEG1 psychoacoustic model 2 implementation [8]. The sinusoidal components below the masking threshold are not extracted. The parameters of the sinusoidal components are transmitted to the decoder after quantization and coding. The encoder re-synthesizes the sinusoidal components and subtracts them from the original signal to form the residual.

The residual is processed by the WPT. To eliminate the perceptual redundancy, the masking thresholds are used again to quantize the WPT coefficients. The frame size of the WPT is 256 samples. There is no overlap between consecutive frames, but the effective filter lengths stretch beyond the block size. For both the sinusoidal and the WPT analyses, the same psychoacoustic model is used with a frame size of 512 samples and overlap of 256 samples providing a masking threshold for every 256 samples. The residual consist of transients, noise, and artifacts of the sinusoidal extraction. Noisy sub-bands are identified and encoded applying PNS. The quantized masking levels along with the quantized WPT coefficients are encoded and sent to the decoder. The decoder decodes the sinusoidal and residual bit-streams, then it re-synthesizes the two signal components. The re-synthesized sinusoidal and residual components are added together to form the decoded signal.

3. Sinusoidal modeling

The sinusoidal model works as follows. First, a Kaiser windowed DFT (zero padded to 1024 points) of the current frame is calculated. Next, the SSM [6] is computed as the correlation between a spectral pattern corresponding to the main lobe of the Kaiser window, and the magnitude spectra. For further details on the SSM see the Appendix. A peak in the SSM is considered to represent a valid sinusoid, only if it exceeds a certain threshold. The amplitude, frequency, and phase parameters corresponding to valid peaks are extracted from the magnitude spectra.

The sinusoids are re-synthesized using the parameters extracted from two consecutive analysis frames. The re-synthesis follows the trajectory matching and synthesis procedures described in [5]. Sinusoids not associated with other sinusoids found in the preceding or following frame, are considered to originate from noise components, hence they are eliminated. Peaks having an amplitude below the masking threshold are also eliminated. Then, the re-synthesized sinusoids are extracted from the original signal.

4. Sinusoidal parameter quantization and coding

The sinusoidal amplitudes are quantized on a logarithmic-scale using 6 bits, the frequencies on a Bark-scale using 10 bits, and the phases on a linear-scale using 5 bits. Sinusoidal components exist for several frames with a limited change in their amplitudes and frequencies. Hence, there is a considerable redundancy between parameters extracted from consecutive frames. As suggested in [9], we use differential Huffman coding for the amplitude and frequency parameters for components continued from previous frames. An additional bit per component indicates whether it is continued from the previous frame.

The new amplitudes are differentially Huffman encoded. There is no easily exploitable redundancy between the phase parameters, thus we do not use any special encoding.

Frequencies within a frame are usually harmonically related to a pitch frequency. To exploit this redundancy, we encode new (not continued) frequencies using a novel pitch-based approach. Figure 2 illustrates the method. For every frame, we determine a pitch frequency using the new frequencies found in that frame. Each new frequency is considered to be in the proximity of a pitch harmonic. To encode the new frequencies, it is sufficient to encode the pitch frequency, the number of closest pitch harmonic and the frequency difference to it. To encode the numbers of the closest pitch harmonics, we use differential Huffman coding, whereas simple Huffman coding is used to encode the frequency differences from pitch harmonics. The pitch is calculated by minimizing the total number of bits required to encode the new frequencies. If the total number of bits exceeds the number of bits required to encode the new frequencies individually, the individual encoding is used. An additional bit indicates whether pitch-based or individual encoding is used.
The method reduces the bit rate required to encode new frequencies by 20 – 30%. It can be applied for all (continued and new) frequencies, if we do not want to use differential encoding between frames because of contingent channel errors.

5. Residual modeling

The residual is formed by subtracting the re-synthesized sinusoids from the original signal. We apply a WPT to process the residual. The WPT realizes a non-uniform filter bank simulating the critical (Bark-) band model of the Human Auditory System. Only the first 21 Bark-bands lying in the input frequency range are considered. We use the decomposition described in [2], as it was designed specifically for speech inputs. The WPT is implemented by cascading Quadrature Mirror Filters in a tree structure. The Daubechies filter of length 10 is used as the prototype filter.

We calculate the masking thresholds using the MPEG1 psychoacoustic model 2 [8]. The masking thresholds are calculated with a resolution of 0.12 frequency points, as required by the sinusoidal analysis. They must therefore be converted to give values for each critical band. The masking in a critical band is determined as the minimum masking threshold value in that critical band. We do not used temporal masking, but it can be introduced later to reduce the bit rate further.

6. Residual parameter quantization and coding

The WPT quantization follows the method described in [2]. For each sub-band, a quantization step is derived from the masking threshold. The 21 quantization steps are quantized on a logarithmic scale using 3 bits. The number of levels required constructs the quantization steps are differentially encoded. The first value is encoded on 3 bits. The symbols (i.e. the differences) are encoded using prefix codes. We use shorter codewords if there are fewer than three symbols. Two bits are used to indicate the number of symbols (1, 2, 3, or more).

The quantization levels are encoded using prefix codes. We use shorter codewords if there are less than 8 symbols. The number of symbols (1, 2, . . . , 7, or more) are encoded on 3 bits. Occasionally, large parts of the signal are below the masking threshold, hence the quantization levels contain long runs of zero symbols. We apply run-length coding to encode the zero symbols, if it requires fewer bits than the separate encoding. An additional bit indicates whether run-length coding was applied in the given frame.

The residual consist of transients, noise, and artifacts of the sinusoidal extraction. Sub-bands corresponding to noisy segments are perceptually redundant, because the ear is not sensitive to the actual coefficients, but to the noise energy only. We apply PNS in noisy sub-bands, and transmit the sub-band energy only. For sub-bands wherein PNS was performed, the decoder generates a random sequence of coefficients having the same energy as the original coefficients.

We use two criteria to determine the noisy nature of a sub-band. First, we measure the sinusoidality of the sub-band based on the SSM values in the given sub-band. The SSM values in the sub-band must be lower than a threshold to assure that no sinusoids are getting replaced by noise. This step is necessary because the sinusoidal analysis does not extract all of the sinusoidal components. Second, we verify whether there is a great change in the sub-band energy from the last frame. Such changes indicate the presence of transients, which should not be encoded as noise. Only the last four sub-bands (containing half of the coefficients) are examined for PNS. Four additional flag bits indicate whether PNS was applied in these sub-bands. The sub-band noise energies – similarly to the coefficients – are transmitted as a quantization step and a quantization level.

7. Results

The quantization steps are differentially encoded. The first value is encoded on 3 bits. The symbols (i.e. the differences) are encoded using prefix codes. We use shorter codewords if there are fewer than three symbols. Two bits are used to indicate the number of symbols (1, 2, 3, or more).

The quantization levels are encoded using prefix codes. We use shorter codewords if there are less than 8 symbols. The number of symbols (1, 2, . . . , 7, or more) are encoded on 3 bits. Occasionally, large parts of the signal are below the masking threshold, hence the quantization levels contain long runs of zero symbols. We apply run-length coding to encode the zero symbols, if it requires fewer bits than the separate encoding. An additional bit indicates whether run-length coding was applied in the given frame.

The residual consist of transients, noise, and artifacts of the sinusoidal extraction. Sub-bands corresponding to noisy segments are perceptually redundant, because the ear is not sensitive to the actual coefficients, but to the noise energy only. We apply PNS in noisy sub-bands, and transmit the sub-band energy only. For sub-bands wherein PNS was performed, the decoder generates a random sequence of coefficients having the same energy as the original coefficients.

We use two criteria to determine the noisy nature of a sub-band. First, we measure the sinusoidality of the sub-band based on the SSM values in the given sub-band. The SSM values in the sub-band must be lower than a threshold to assure that no sinusoids are getting replaced by noise. This step is necessary because the sinusoidal analysis does not extract all of the sinusoidal components. Second, we verify whether there is a great change in the sub-band energy from the last frame. Such changes indicate the presence of transients, which should not be encoded as noise. Only the last four sub-bands (containing half of the coefficients) are examined for PNS. Four additional flag bits indicate whether PNS was applied in these sub-bands. The sub-band noise energies – similarly to the coefficients – are transmitted as a quantization step and a quantization level.
We have evaluated the performance of our method using different speech and music samples. Sound files demonstrating the compressed audio quality are available at www-sci.ens-bretagne.fr/~jeke/eurospeech01.

Our first test aimed to determine the optimum SSM threshold that minimizes the overall bit rate. The upper part of Figure 3 shows the results. The first three columns show the results using different SSM thresholds \( \mu = 0.85, 0.90, 0.95 \). The fourth column corresponds to WPT encoding without sinusoidal modeling. The quality of the compressed signal is perceptually the same for the five different settings. By extracting more sinusoids (using lower thresholds), the bit rate required by the WPT encoding is lower for all signals. However, for the castanets and English male signals, the usage of the sinusoidal model leads to a higher bit rate than the WPT encoding alone. These signals cannot be modeled well by slowly changing sinusoids. The biggest reduction is achieved for the bagpipe signal which can be well modeled by slowly changing sinusoids.

Our second test aimed to evaluate the performance of the PNS. The lower part of Figure 3 shows the results. A bit rate reduction is achieved for all signals. However, the compressed audio quality is degraded for the English male signal, where clear signal parts became noisy. The slowly changing sinusoidal audio quality is degraded for the English male signal, where reduction is achieved for all signals. However, the compressed PNS. The lower part of Figure 3 shows the results. A bit rate - no PNS

Our second test aimed to evaluate the performance of the PNS. The lower part of Figure 3 shows the results. A bit rate - with PNS

9. Appendix: Sinusoidal Similarity Measure

The SSM is calculated as follows. First, the windowed DFT spectrum \( S[\omega] \) of the input frame \( s[n] \) is calculated:

\[
S[\omega] = \sum_{n=-N}^{N} s[n]e^{-j2\pi \omega n},
\]

where \( h[n] \) denotes the normalized window function. The spectrum is then normalized by the local spectral energy:

\[
\tilde{S}[\omega] = \frac{S[\omega]}{\sqrt{\sum_{\omega_m} S^*[\omega] S[\omega]}}
\]

where the main lobe of the window’s spectrum is situated between \(-\omega_{\text{null}}\) and \(\omega_{\text{null}}\). The sinusoidality measure \( \mu[\omega] \) is calculated by convolving the normalized spectrum \( \tilde{S}[\omega] \) with the main-lobe of the conjugated normalized window’s spectrum \( H^*[\omega] \):

\[
\mu[\omega] = \sum_{\omega_m} |\tilde{S}[\omega]H^*[\omega - \Omega]|.
\]

where \( H[\omega] \) is bounded by \(-\omega_{\text{null}}\) and \(\omega_{\text{null}}\).

10. References


