Object-Based Analysis/Synthesis Audio Coder for Very Low Bit Rates

Heiko Purnhagen, Bernd Edler, Charalampos Ferekidis

Laboratorium für Informationstechnologie
Uni Hannover, Schneiderberg 32, D-30167 Hannover, Germany
Tel.: +49 511 762 5033, Fax: +49 511 762 5052
E-Mail: `purnhage@tnt.uni-hannover.de`
URL: `http://www.tnt.uni-hannover.de/~purnhage/`

Abstract

An Object-Based Analysis/Synthesis Audio Coder designed for bit rates between 4 and 16 kbit/s is presented. The coding is based on the decomposition of the input signal into audio objects which are described by appropriate source models and represented by model parameters. Object models for sinusoids, harmonic tones, and noise are utilised.

1 Introduction

In the context of evolving multimedia applications new demands for very low bit rate audio coding arise. Besides high coding efficiency required to cope with the limited bandwidth of the internet or mobile transmission channels also new functionalities like flexible access to coded data and manipulation by the recipient are desired. These requirements are addressed by the current ISO/IEC MPEG-4 standardisation activities [1].

Currently there are two well established techniques available for coding of audio signals at bit rates below 16 kbit/s. On the one hand speech coding techniques make use of the specific properties of the speech signal for which they are designed by incorporating models of the signal source – the speech generation in the human vocal tract [2]. On the other hand codecs designed for arbitrary audio signals extensively exploit the properties of signal perception by the human auditory system. Their model assumptions for the signal source usually are restricted to general statistical properties [3].

Investigations in the capabilities of coding arbitrary audio signals based on these two approaches show that there is a gap for music signals with contents of relatively low complexity, like e.g. single instruments. For this type of audio signals both approaches mentioned provide only limited performance [4, 5]. In order to close this gap and to provide higher coding efficiency for such audio signals at bit rates in the range of 4 to 16 kbit/s, a new coding technique was developed. It utilises an object-based approach, assuming that the input signal can be decomposed into several audio objects, like e.g. sinusoids, harmonic tones, or noise. These audio objects are described by appropriate source models and represented by model parameters.

The Object-Based Analysis/Synthesis Audio Coder presented in this paper combines the object-based modelling of the signal source with a model of signal perception. It is an extension of an earlier analysis/synthesis scheme which utilised only one type of audio objects with a sinusoidal source model [6]. Now three different types of audio objects will be used.

Section 2 gives a short overview on the different models for the signal source and for the auditory perception used in known audio coding techniques. After that, an introduction into the basic concepts of the object-based approach and the underlying source models is given in Section 3. Section 4 describes the decomposition of an audio signal into the different objects and the procedures for the estimation of the parameters describing an object. The encoding process for the obtained parameters is described in Section 5 and the decoding and synthesis units are described in Section 6. Experimental results and results of subjective evaluations are presented in Section 7. Conclusions and consideration for further development are given in Section 8.

2 Source and Perception Models in Audio Coding

Data reduction for audio signals is usually based on the reduction of redundancy and irrelevance contained in the input signal. Redundancy reduction can be achieved if an appropriate model of the input signal is available. Based on such a source model an encoder can replace the input signal by a representation which is more efficient in the sense of required transmission capacity than the original PCM representation. On the other hand irrelevance can be reduced if a model of perception of the decoded signal at the receiver is available. Usually the main goal of very low bit rate coding is to achieve the maximum possible subjective sound quality with respect to the original signal. Therefore a model of human auditory perception is required to control the quantisation and coding of the parameters which represent the input signal. General block diagrams of an audio encoder and decoder resulting from these considerations are given in Figure 1.

General purpose low bit rate high quality audio codecs are based on a spectral decomposition by a transform, like for example the MDCT [3], which also can be regarded as a subband filter bank. Since no restriction concerning the input material is desired, no specific source model can be applied for such a codec. However the fact is exploited that most audio signals use to be quasi stationary, i.e. show only slowly changing signal properties. Based on this assumption an analysis consisting of a segmentation into frames and subsequent mapping to the frequency domain results in a signal representation which is more efficient than the original PCM. Since the synthesis block in the decoder consists of an inverse transform, errors introduced by quantisation of a spectral component are limited to the corresponding frequency interval. This enables an efficient implementation of irrelevance reduction by spectrally shaping the quantisation noise by controlling the quantiser step sizes for the individual spectral components. The necessary control information is obtained from a perception model (psychoacoustic model) exploiting frequency dependent masking effects in the human ear. Based on refinements of the source model and the perception model several extensions to transform coding have been introduced.

Transform codecs show very good performance for coding music signals of all complexity levels at bit rates down to 16 kbit/s. Below 16 kbit/s however speech codecs perform clearly better, if the input material is speech, due to their specialised source model.

Many of the recently designed speech codecs are based on CELP (Code Excited Linear Prediction) type speech coding or similar techniques [2]. The fact that the type of input signals for which these codecs are designed is restricted to speech is reflected in a sophisticated, physically motivated source model. Speech generation can be decomposed to an excitation which can be either voiced or unvoiced and to an acoustic resonator which shapes the spectral envelope. This is reflected in the synthesis block of a speech decoder which usually consists of an excitation generator and a LPC synthesis filter. In the corresponding encoder the filter parameters are derived from the spectral envelope of the input signal using LPC techniques, and the optimum excitation is selected from a codebook in an analysis/synthesis loop by minimising the residual error. The codebook often consists of two parts, an adaptive codebook and a stochastic codebook. The adaptive codebook contains previous samples which can be accessed with a time offset in order to model voiced signal parts, while the stochastic codebook mainly has the purpose of modelling unvoiced signal parts. In the loop a perception model can be taken into account by spectrally weighting the error between the synthesised signal and the input signal. However, compared to the sophisticated psychoacoustic models used in transform coding, this weighting filter does not exactly model the masking but is derived from the spectral envelope just in order to let the resulting error spectrum have a similar shape. It is interesting to note that the LPC synthesis filter length usually corresponds quite well to the propagation time of sound in the vocal tract. The described basic CELP codec often is extended in correspondence to refinements of the source model.

CELP type speech codecs are used in different standards and show very good performance for bit rates between 4 and 24 kbit/s as long as the input material is speech. If they are applied to music, it becomes obvious that the specialised source models used in CELP codecs are often not appropriate for general audio signals.

Summarising these considerations, it can be stated that in the design of transform codecs a strong focus is put on the perception model, while in CELP codecs the main focus lies on the advanced and highly specialised source model. Furthermore it should be noted that transform codecs [3] as well as CELP codecs [2] use perception models which describe the audibility of a residual error signal in relation to the original audio signal, where the error signal is defined as the difference between the decoder output and the (delay compensated) encoder input signal. Therefore codecs of both types produce an output signal which gives an approximation of the input signal waveform.

3 Object-Based Audio Coding

In this Section a new approach is presented which combines the use of a sophisticated object-based source model with an appropriately designed perception model in order to achieve a high overall coding efficiency. Due to this new approach, the source and perception models used in Object-Based Audio Coding differ from the models used in the two techniques described in the previous Section.

The basic assumption of Object-Based Audio Coding is that the input signal is composed of different components which can be regarded as audio objects. These objects may have different characteristics. Therefore different types of source models with corresponding model parameters can be appropriate to describe different audio objects. Such a source model for an audio object is denoted as object model here.

Figure 2 shows the block diagram of an object-based audio encoder utilising three different object models:

- harmonic tone
- individual sinusoid
- noise

First the input signal is decomposed into audio objects according to the object models supported. Of course several audio objects might use the same object model. The model parameters for each of the audio objects are estimated and then quantised and coded. To improve signal decomposition and parameter estimation, an iterative analysis-by-synthesis approach is utilised. Since only a limited number of parameters can be encoded in the bitstream at the desired bit rates, the quantisation and coding is controlled by a perception model to ensure that those audio objects most relevant for the sound perception are transmitted.

Figure 3 shows the block diagram of the corresponding object-based audio decoder. At first the object parameters are reconstructed and then the audio objects are synthesised according to their object models. Finally all audio objects are added to compose the output signal of the decoder.

In an earlier implementation of this approach [6] only the object model *individual sinusoid* was supported. This means that the input signal was modelled as a set of individual spectral lines each described by its frequency, amplitude, and phase. Similar to the source models described in the previous Section it was assumed hat the input signal behaves quasi stationary and thus can be processed and represented on a time frame basis.

This earlier approach now has been extended in order to support additional object models. Besides the object model *individual sinusoid* now also the object models *harmonic tone* and *noise* can be utilised to describe components of the input signal. A *harmonic tone* object is characterised by its fundamental frequency and the amplitudes and phases of all harmonic partials. A *noise* object can be described by its power spectral density and therefore is represented by parameters for its amplitude and spectral shape.

In the perception model, which is used to control the quantisation and coding of the parametric representation, there also is a major difference compared to the perception models described in the previous Section. In transform coding as well as in CELP coding the quality criterion is derived from the residual error signal calculated in the time domain. In Object-Based Audio Coding however, the quality criterion is how similar the decoder output signal sounds compared to the original input signal. This more general perception model allows to release the requirement of waveform approximation and thus enables additional irrelevance reduction. E.g. it is often not necessary to transmit the phase of a sinusoidal signal component.

Similar approaches for modelling and coding audio signals are described by McAulay and Quatieri [7], George and Smith [8, 9], and Serra [10]. In [7] and [10] a straight-forward spectral peak detection is used to compute the frequencies of sinusoids while in [8, 9] an analysis-by-synthesis procedure is utilised. Only in [10] a distinct model for noise components is used. But [10] addresses only sound processing and synthesis aspects, not audio coding. However, in none of these approaches a specific perception model is included to control the parameter quantisation and coding.

4 Signal Decomposition and Modelling

The decomposition of the input signal into audio objects and the estimation of object parameters according to appropriate source models for the audio objects constitute the core of the Object-Based Analysis/Synthesis Audio Coder. The signal decomposition and modelling is done on a time frame basis to accommodate the time-variant characteristics of an audio signal. For each time frame a set of model parameters is computed which describe the input signal in this frame. A frame length of typically 32 ms is used and overlapping analysis windows are employed to allow smooth modelling of varying characteristics of the input signal.

Since signal decomposition on one hand and the modelling of the separated audio objects on the other hand are quite interdependent functional blocks, a joint optimisation approach seems to be advantageous. Therefore an iterative analysis-by-synthesis approach is utilised in the object-based coder. The basic idea is to iteratively extract audio objects from the input signal. First an audio object is modelled by choosing an appropriate source model and estimating the model parameters. Then the object is re-synthesised according to the model and the estimated parameters. Finally the re-synthesised object is removed from the input signal before the next object is extracted.

4.1 Sinusoidal source model

The signal decomposition and modelling core of an object-based audio coder only utilising the object model for an *individual sinusoid* with the model parameters frequency, amplitude, and phase is depicted in Figure 4. In this coder core, which can be regarded as an analysis/synthesis loop, the individual sinusoids are extracted from the input signal [6]. A perception model is included in the analysis/synthesis loop to control the order in which the sinusoids are extracted such that those sinusoids most relevant for sound perception are extracted first.

In each iteration of this loop, first a coarse estimate $f_{c,i}$ of the frequency of the most important sinusoid i not yet extracted from the input signal $x(t)$ is computed as described below. Based on $f_{c,i}$ the model parameters frequency f_i, amplitude a_i, and phase φ_i are accurately estimated using a linear regression technique [11]. Finally the sinusoid i is re-synthesised using these parameters and added to the previously synthesised signal $s(t)$. Moreover the parameters f_i and a_i of the extracted sinusoids are utilised in a parametric psychoacoustic model to calculate the masked threshold $M(f)$ evoked by $s(t)$ [12, 13]. By comparing the spectrum of the residual signal $e(t) = x(t) - s(t)$ with the masked threshold $M(f)$, the coarse frequency $f_{c,i+1}$ of the remaining sinusoidal component in $e(t)$ having the highest level above $M(f)$ is found. The number of iterations of the analysis/synthesis loop is determined by the number of model parameters that can be encoded and transmitted in the bitstream at the desired bit rate.

To improve the modelling of transient components in the input signal, the normalised amplitude envelope $a(t)$ of the input signal is computed. This envelope is then approximated by a generic envelope described by the three parameters peak time t_{max}, attack rate r_{atk}, and decay rate r_{dec}. In Figure 5 the envelope of the original signal together with the approximated envelope is shown for one frame of the signal "castanets." To find the envelope parameters a weighted regression technique is used which precludes low-level parts of the original envelope below the RMS level from influencing the estimated envelope parameters [6]. For each audio object it is determined whether applying the approximated amplitude envelope to the synthesised signal improves the modelling or not.

Up to now only the modelling of a single time frame of the input signal was described. But for many audio signals it can be observed that sinusoidal signal components persist for longer periods in time than just a single time frame. To exploit this property for an improved modelling of the input signal, the frequency and amplitude parameters of the sinusoids in adjacent time frames are compared. Thus sinusoids continuing from the previous time frame into the current time frame can be detected. This information about continued sinusoids allows for efficient parameter coding and smooth synthesis in the decoder.

4.2 Extended source model

To improve the coding efficiency of the object-based audio coder, two additional source models for audio objects are introduced. The first new model describes a *harmonic tone* with the parameters fundamental frequency and the amplitudes and phases of all its harmonic partials. This model permits a more efficient representation of a harmonic tone than possible by describing it as a number of *individual sinusoid* objects. The other new model permits to describe noise-like components in the input signal using parameters for the amplitude and the spectral shape of a noise signal.

The signal decomposition and modelling utilising the three object models *harmonic tone*, *individual sinusoid*, and *noise* is done in three steps as depicted in Figure 6.

In the first step, a coarse estimate of the fundamental frequency of the dominant *harmonic tone* in the input signal is computed based on a cepstrum analysis of the input signal. For each harmonic partial of this tone the frequency, amplitude, and phase are accurately estimated using the same technique as for individual sinusoids. Based on these parameters the accurate fundamental frequency of the harmonic tone is calculated. For some musical instruments – e.g. those based on plucked strings, like a harpsichord – it can be observed that the frequencies of higher harmonics are not exact integer multiples of the fundamental frequency. This phenomenon is known as stretching and usually caused by stiff strings [14]. It is taken into account in the *harmonic tone* model by an additional parameter describing the stretching. Since also a harmonic tone can persist longer than just a single frame, the fundamental frequencies in adjacent time frames are compared in order to detect continuing tones and permit smooth synthesis in the decoder.

In the second step the analysis/synthesis loop as described above is used to extract the sinusoidal components of the input signal. Since the input signal was not modified in the first step, all harmonic partials are still present in the signal. For each sinusoid extracted by the analysis/synthesis loop, its frequency is used to decide whether it is a partial of the *harmonic tone* found in the first step or whether it must be regarded as an *individual sinusoid* object. After completion of this second step, all remaining partials of the harmonic tone that were not extracted in the analysis/synthesis loop are also removed from the residual signal $e(t)$.

In the third step it is assumed that after extracting the *harmonic tone* and the *individual sinusoid* objects from the input signal all distinct tonal components have been removed from the input signal. Thus the remaining residual signal $e(t)$ is considered as a noise-like signal. The residual is now modelled by a noise with a spectral shape similar to the magnitude spectrum of the residual signal. To describe the amplitude and spectral shape of the noise, the first few coefficients of a cosine series expansion of this magnitude spectrum are used as parameters. They are calculated using a discrete cosine transform (DCT).

Besides the model parameters found in the signal decomposition and modelling process also parameters describing the relevancy of the different audio objects for the sound perception of the complete audio signal are passed on to the coding and quantisation unit to permit taking them into account in the bit allocation process.

5 Coding

To transmit the model parameters of the audio objects to the decoder adequate quantisation and coding is required. In this process the properties of human auditory perception are taken into account for parameter quantisation as well as for the selection of the parameters and audio objects to be transmitted.

For the object model *individual sinusoid* only the frequency and amplitude of a sinusoid have to be transmitted since the influence of its phase on the sound perception can be neglected. If a sinusoid starts in the current time frame, frequency and amplitude are quantised with a logarithmic law to consider the perception of pitch and loudness. For a sinusoid continued from the previous frame only the frequency and amplitude changes are transmitted since this requires less bits. If necessary also the parameters of the amplitude envelope are quantised and transmitted.

At a bit rate of 6 kbit/s the parameters of about 10 to 20 sinusoids can be transmitted for each 32 ms time frame. The number of transmitted sinusoids varies from frame to frame and depends mainly on the number of sinusoids that can be continued from the previous frame. When selecting the sinusoids that are to be transmitted, the order in which these sinusoids were extracted by the analysis/synthesis loop is evaluated. Thus always those sinusoids are transmitted that are most important for the sound perception.

If the additional object models *harmonic tone* and *noise* are utilised, also the parameters of these objects must be quantised and transmitted.

The fundamental frequency of the *harmonic tone* and the amplitudes of its harmonic partials are quantised with a logarithmic law. Since particularly in case of a low fundamental frequency many (up to 100) partial amplitudes must be coded, starting from the 10th harmonic the partials are grouped and only the average amplitude is transmitted for each group. In this way the decreasing frequency selectivity of the human auditory perception for higher frequencies is exploited.

The parameters describing the spectral shape of the *noise* are first normalised and the corresponding scaling factor required for the reconstruction is quantised with a logarithmic law. Then the normalised parameters are quantised and transmitted together with the scaling factor.

The allocation of the available bit rate to the parameters of the different audio objects *harmonic tone*, *individual sinusoid*, and *noise* is automatically controlled by the relevancy of the different objects for sound perception.

6 Decoding and Synthesis

The decoding of a frame starts with the decoding and dequantisation of the amplitude envelope parameters, if transmitted. The amplitude envelope is then reconstructed based on these parameters.

For each *individual sinusoid* object, its frequency and amplitude parameters are decoded and dequantised according to the coding described in the previous Section. If it is signalled in the bitstream that a sinusoid is to be continued from the previous time frame, the phase is continued and the frequency and amplitude parameters are interpolated linearly from their previous to their current values for the synthesis. A sinusoid beginning in the current frame is smoothly faded in to avoid artifacts and a random start phase is used. In a corresponding way sinusoids that will not be continued are faded out. If for a sinusoid the usage of the amplitude envelope is signalled, the synthesised sinusoid is multiplied with the reconstructed envelope.

For a *harmonic tone* object the fundamental frequency and the amplitudes of the harmonic partials are decoded and dequantised. Then the frequencies of all harmonics are calculated and all harmonic partials are synthesised as individual sinusoids as described above. To obtain smooth synthesis of sinusoids coded as a partial of a harmonic tone in one frame and as individual sinusoid in an adjacent frame, frequency and amplitude differences are evaluated to decide whether a sinusoid should be regarded as being continued between these frames or not.

To synthesise the signal of a *noise* object, first the spectral shape is reconstructed from the decoded and dequantised noise parameters utilising an inverse DCT. Then this magnitude spectrum is converted into a complex spectrum by adding a random phase spectrum. By transforming this spectrum with an inverse DFT the synthesised noise signal is obtained. To obtain smooth transitions between frames, a windowed overlap-add scheme is used for the noise signal.

A very interesting property of the coding scheme presented here arises from the fact that the signal is described in terms of frequency and amplitude parameters. This signal representation permits new functionalities like independent speed and pitch change which can be implemented in the decoder with very little additional complexity. By varying the length of the synthesised time frames, the playback speed of the signal can be modified in the decoder without affecting the pitch of the signal. On the other hand it is possible to multiply all frequency parameters by a given factor prior to synthesis to alter the pitch of the decoded signal without affecting the playback speed.

7 Results

To evaluate the Object-Based Analysis/Synthesis Audio Coder, first the behaviour of the analysis/synthesis loop is considered. In Figure 7 the magnitude spectra of the original signal and the signal synthesised from the parameters of the extracted individual sinusoids is depicted for one frame of the signal "male speech." Moreover the masked threshold calculated from the synthesised signal is shown. It can be seen that in this case 13 sinusoids have been extracted which quite good recreate the relevant spectral components of the original signal.

In Figure 8 the frequency and amplitude parameters of the *individual sinusoid* objects and the *harmonic tone* object are shown for a short excerpt from the signal "female vocal." For longer segments of the signal the fundamental and harmonics of a voiced sound are clearly dominant. Around 2.5 s a strong formant at 2.6 kHz can be observed and at about 2.7 s the fundamental frequency is changing. For short segments around 2.25 s and 3.0 s there is an unvoiced sound present. Here only sinusoids with low amplitudes are synthesised and the signal is dominated by the noise component (whose parameters are not shown in this Figure).

The initial sinusoidal coding scheme [6, 15] was proposed for the MPEG-4 standard currently being developed [1]. Due to its coding efficiency for audio signals at very low bit rates, it was accepted to be part of the MPEG-4 standard. Later the extensions for harmonic tones and noise were proposed to MPEG and evaluated in a so-called core experiment procedure [16]. This included subjective and objective evaluation of the merits or possible drawbacks of each proposal. Results from subjective listening tests showed that significant improvements could be achieved especially for those test signals for which the original codec produced the most annoying artifacts [17].

The results of a listening test comparing the extended codec (A) against the initial sinusoidal codec (B) at 6 kbit/s are shown in Figure 9 and Table 1 (from [17]). In this test 8 mainly non-expert listeners were asked to grade codec A vs. codec B on the ITU-R 7-point comparative scale shown in Table 2. A set of 12 test items was used, consisting of 3 items of each of the signal classes "english/german speech," "single instruments," "simple sound mixtures," and "complex sound mixtures." The test items were presented in randomised order as "Ref,A,B,Ref,A,B" and "Ref,B,A,Ref,B,A", where "Ref" is the original encoder input signal sampled with $f_s = 8$ kHz and "A" and "B" are decoded signals of the respective codecs. As can be seen from the 95% confidence intervals, significant improvements were observed for 6 of the 12 test items.

Because of the gain in coding efficiency observed in the core experiment, the proposed extensions for harmonic tones and noise were also accepted by MPEG and are included in the current Committee Draft [18]. The Object-Based Analysis/Synthesis Audio Coder presented here is a joint contribution to MPEG by the University of Hannover and Deutsche Telekom Berkom. Within MPEG-4 Audio, it is referred to as the "Harmonic and Individual Lines plus Noise" (HILN) parametric coder.

In another core experiment subjective evaluations were carried out for the comparison of a CELP codec, a transform codec, and the initial sinusoidal coding scheme (referred to as a "parametric codec") at a bit rate of 6 kbit/s [19]. The results showed that there was a very strong dependency of the performance of all three coding schemes on the type of audio material coded. As expected the speech codec clearly outperformed the other two schemes if applied to speech signals. For music signals always one of the two other codecs designed for more general audio contents clearly outperformed the CELP codec. But it was interesting to see that there were again significant item dependent differences in the behaviour of these two codecs. For audio material with contents of relatively high complexity the transform codec performed significantly better than the parametric codec. On the other hand for audio signals with contents of relatively low complexity the parametric codec offered significantly better audio quality than the transform codec.

By introducing the extensions described above to the parametric codec the performance disadvantages compared to the transform codec for complex audio signals could be reduced. However, informal subjective evaluation shows that there still is some advantage for the transform codec if the complexity of the input signals is too high for a parametric representation with a relatively small number of objects. Additionally the superposition of the signals of a large number of different instruments in a complex auditory scene may cause problems for the automatic decomposition into audio objects and therefore also for the parameter estimation in the Object-Based Analysis/Synthesis Audio Coder.

8 Conclusions

In this paper an Object-Based Analysis/Synthesis Audio Coder was presented. It combines an object-based source model, where the input signal is modelled as the superposition of several audio objects, with an appropriate perception model, which controls the quantisation and coding of the parameters describing the different audio objects.

This Object-Based Analysis/Synthesis Audio Coder was developed by extending an earlier analysis/synthesis scheme which utilised only a sinusoidal source model [6]. Now, three different types of audio objects are supported, each type with a unique set of model parameters for describing the specific object characteristics. The three object models currently utilised are *harmonic tone*, *individual sinusoid*, and *noise*. A first encoder implementation was presented which uses automatic decomposition of the input signal into audio objects and estimation of corresponding model parameters. Additionally, methods for including a perception model specialised for application in object-based coding were shown.

In comparison to the earlier sinusoidal coding, scheme which already had been accepted to be a part of the MPEG-4 standard currently being developed, the subjective sound quality obtained at a bit rate of 6 kbit/s could be significantly improved. The biggest improvements were obtained for input signals with a high complexity content. Based on these results, this extension also was accepted for MPEG-4 Audio and thus is included in the current Committee Draft [18]. Within MPEG-4 Audio, the Object-Based Analysis/Synthesis Audio Coder presented in this paper is referred to as the "Harmonic and Individual Lines plus Noise" (HILN) parametric coder.

Informal comparisons with other coding techniques operating in the same bit rate range showed that the quality of the different techniques heavily depends on the type of input signal. For a rough classification the coding techniques are categorised to the three types: CELP speech coding, transform coding, and parametric coding. For speech signals obviously the specialised CELP codecs show the best performance. For music signals with relatively high complexity contents transform codecs show clear advantages over the two other techniques, while the parametric approach proves to be most suitable for music signals with relatively low complexity contents.

Besides high coding efficiency, the parametric approach presented here easily provides the additional functionalities of independent pitch and speed change. It is also possible to use this coder as a core in configurations offering bit rate scalability by enhancing the quality using other coding techniques like transform coding. The coder presented here is also used in a dynamically scalable audio coder currently prepared for the internet [20].

Further improvements are expected from refinements of the signal decomposition and parameter estimation and from the extension of the source model towards the support of additional object types [21], e.g. for improved representation of speech by using models applied in parametric speech codecs [22].

The work presented here has been done under contract of and in close cooperation with the Deutsche Telekom Berkom GmbH, Berlin, Germany.

References

[1] ISO/IEC JTC1/SC29/WG11, "Overview of the MPEG-4 Version 1 Standard," ISO/IEC JTC1/SC29/WG11 N2078, Feb. 1998.
`http://www.cselt.it/mpeg/standards/mpeg-4.htm`

[2] P. Noll, "Speech Coding for Communications," *Proc. Eurospeech*, 1993.

[3] K. Brandenburg and M. Bosi, "Overview of MPEG Audio: Current and Future Standards for Low Bit Rate Audio Coding," *J. Audio Eng. Soc.*, Vol. 45, No. 1/2, pp. 4–21, Jan./Feb. 1997.

[4] B. Edler, "Current Status of the MPEG-4 Audio Verification Model Development," *101th AES Convention*, Preprint 4376, Nov. 1996.
`http://www.tnt.uni-hannover.de/project/coding/audio/asac/aes_101.html`

[5] B. Edler, "Very Low Bit Rate Audio Coding Development," *AES 14th International Conference*, June 1997.

[6] B. Edler, H. Purnhagen, and C. Ferekidis, "ASAC - Analysis/Synthesis Audio Codec for Very Low Bit Rates," *100th AES Convention*, Preprint 4179, May 1996.
`http://www.tnt.uni-hannover.de/project/coding/audio/asac/aes_100.html`

[7] R. J. McAulay and T. F. Quatieri, "Speech Analysis/Synthesis Based on a Sinusoidal Representation," *IEEE Trans. Acoust., Speech, Signal Process.*, Vol. 34, No. 4, pp. 744–754, Aug. 1986.

[8] E. B. George and M. J. T. Smith, "Analysis-by-Synthesis/Overlap-Add Sinusoidal Modeling Applied to the Analysis and Synthesis of Musical Tones," *J. Audio Eng. Soc.*, Vol. 40, No. 6, pp. 497–516, June 1992.

[9] E. B. George and M. J. T. Smith, "Speech Analysis/Synthesis and Modification Using an Analysis-by-Synthesis/Overlap-Add Sinusoidal Model," *IEEE Trans. Speech, Audio Process.*, Vol. 5, No. 5, pp. 389–406, Sep. 1997.

[10] X. Serra, "Musical Sound Modeling with Sinusoids plus Noise," *Musical Signal Processing*, C. Roads et al., Editors. Swets & Zeitlinger Publishers, 1997.
`http://www.iua.upf.es/~sms/docs/msm/`

[11] S. Kay, "A Fast and Accurate Single Frequency Estimator," *IEEE Trans. Acoust., Speech, Signal Process.*, Vol. 37, No. 12, pp. 1987–1989, Dec. 1989.

[12] C. Ferekidis, *Kriterien für die Auswahl von Spektralkomponenten in einem Analyse/Synthese-Audiocoder*. Diplomarbeit (diploma thesis), Universität Hannover, Institut für Theoretische Nachrichtentechnik und Informationsverarbeitung, Mar. 1997.

[13] C. Ferekidis, B. Edler, and H. Purnhagen, "Psychoacoustic Model using Object-Based Model Parameters," submitted for *104th AES Convention*, May 1998.

[14] N. H. Fletcher and T. D. Rossing, *The Physics of Musical Instruments*, pp. 61–62. Springer, New York, 1991.

[15] B. Edler, H. Purnhagen, and C. Ferekidis, “Detailed Technical Description of the MPEG-4 Audio Coding Proposal from University of Hannover and Deutsche Telekom AG,” ISO/IEC JTC1/SC29/WG11 MPEG96/0843, Mar. 1996.

[16] H. Purnhagen, B. Edler, and C. Ferekidis, “Proposal of a Core Experiment for extended ‘Harmonic and Individual Lines plus Noise’ Tools for the Parametric Audio Coder Core,” ISO/IEC JTC1/SC29/WG11 MPEG97/2480, July 1997.

[17] H. Purnhagen, B. Edler, “Check Phase Results of Core Experiment on extended ‘Harmonic and Individual Lines plus Noise’,” ISO/IEC JTC1/SC29/WG11 MPEG97/2795, Oct. 1997.

[18] ISO/IEC JTC1/SC29/WG11, “MPEG-4 Audio Committee Draft 14496-3,” ISO/IEC JTC1/SC29/WG11 N1903, Oct. 1997.
`http://www.tnt.uni-hannover.de/project/mpeg/audio/documents/`

[19] T. Moriya et al., “Proposal of the core-experiment of 6 kbit/s (8kHz sample) t/f coder,” ISO/IEC JTC1/SC29/WG11 MPEG97/2327, July 1997.

[20] B. Feiten, R. Schwalbe, and F. Feige, “Dynamically Scalable Audio Internet Transmission,” submitted for *104th AES Convention*, May 1998.

[21] B. Winduratna, “FM Analysis/Synthesis Based Audio Coding,” submitted for *104th AES Convention*, May 1998.

[22] M. Nishiguchi et al., “Technical description of Sony IPC’s proposal for MPEG-4 Audio and Speech Coding,” ISO/IEC JTC1/SC29/WG11 MPEG95/0321, Nov. 1995.

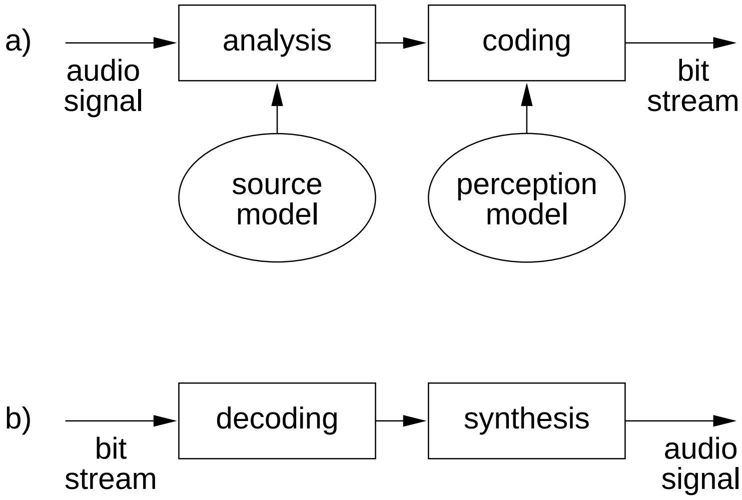

Figure 1: Basic block diagram of an audio encoder (a) and decoder (b).

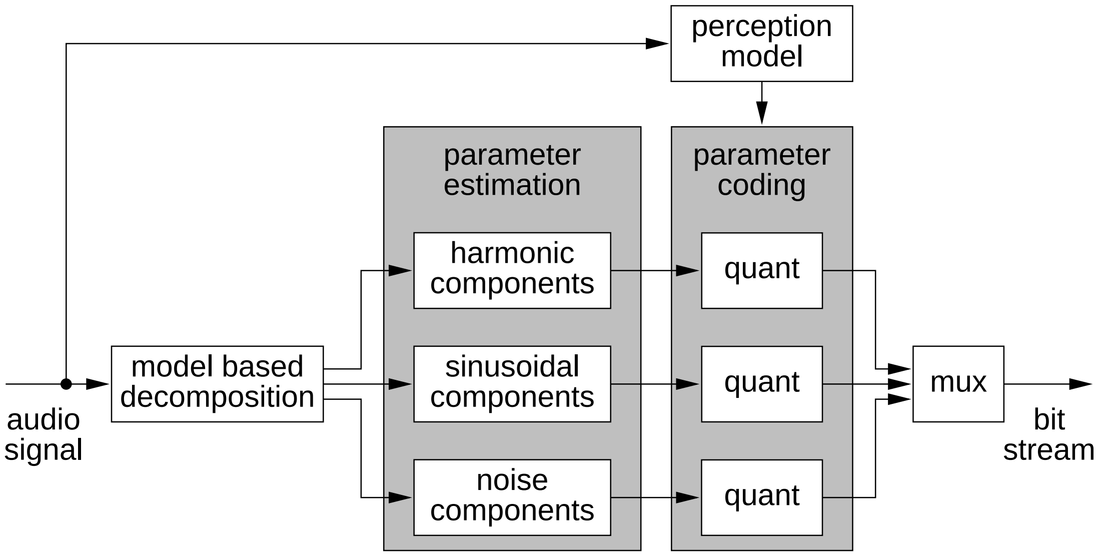

Figure 2: Block diagram of an object-based audio encoder.

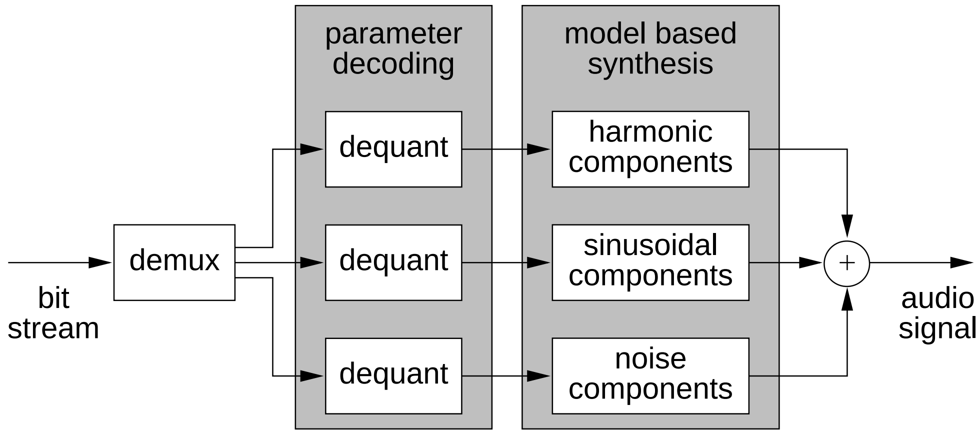

Figure 3: Block diagram of an object-based audio decoder.

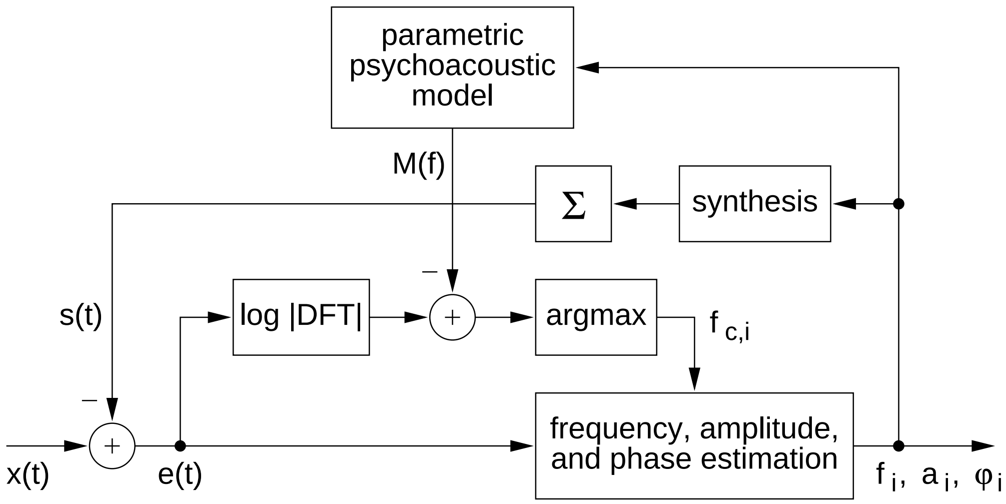

Figure 4: Signal decomposition and parameter estimation based on an analysis/synthesis loop for the sinusoidal source model utilising the object model *individual sinusoid*.

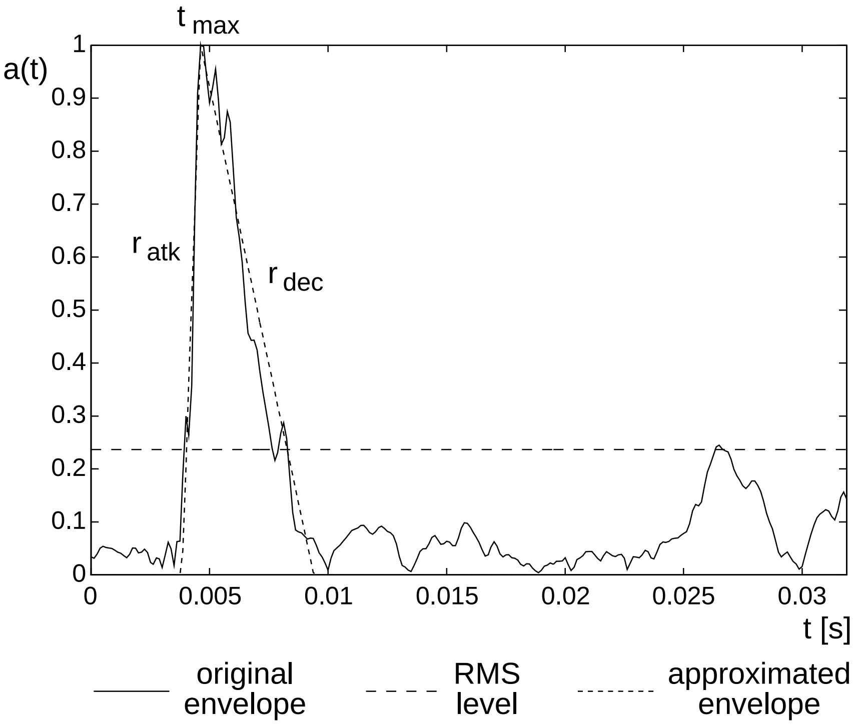

Figure 5: Normalised amplitude envelope $a(t)$ and envelope parameters t_{max}, r_{atk}, and r_{dec} for one frame of signal "castanets."

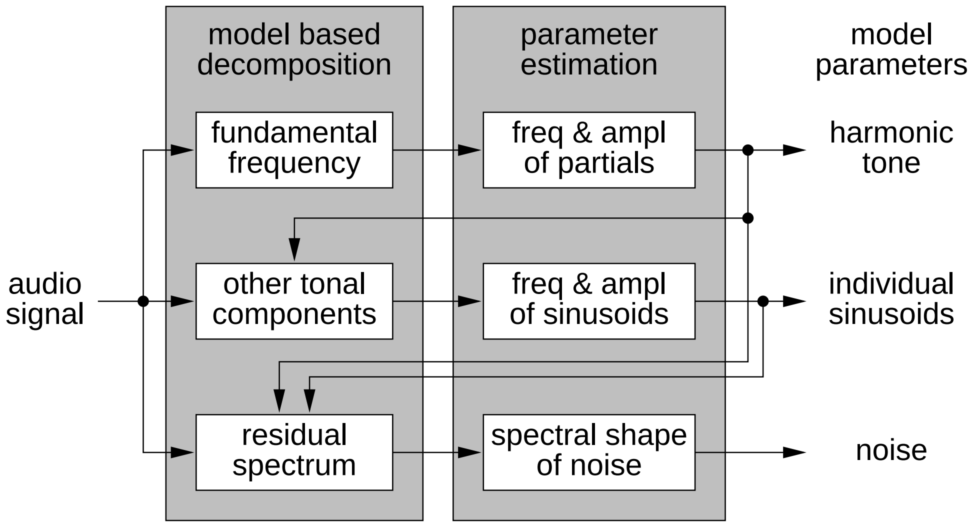

Figure 6: Signal decomposition and parameter estimation for the extended source model utilising the object models *harmonic tone*, *individual sinusoid*, and *noise*.

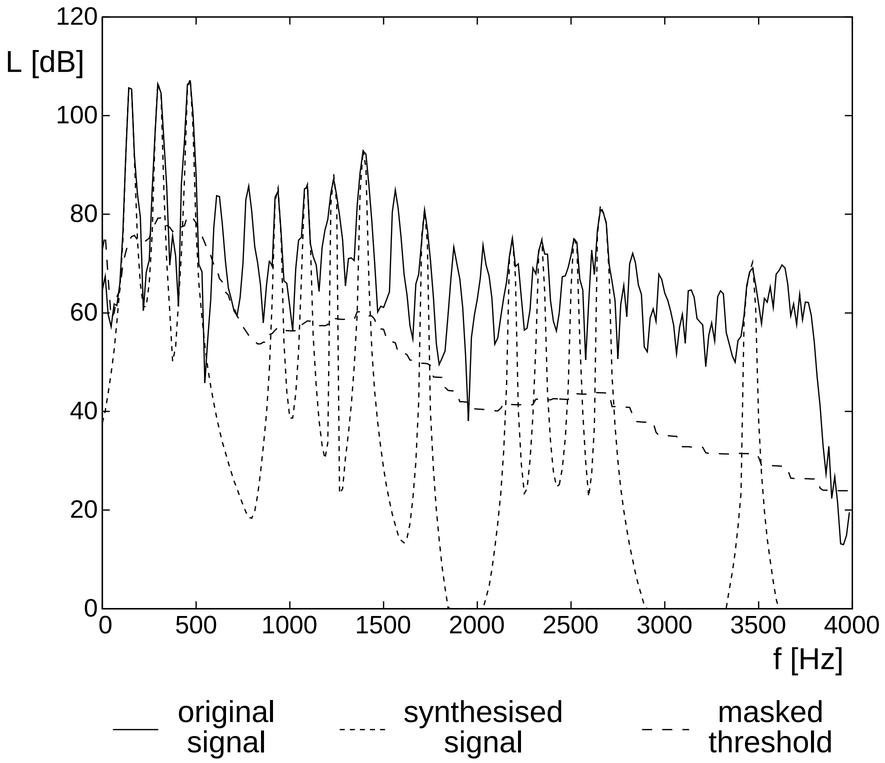

Figure 7: Magnitude spectra of original signal $X(f)$ and synthesised signal $S(f)$ and masked threshold $M(f)$ for synthesised signal using the object model *individual sinusoid* for one frame of "male speech."

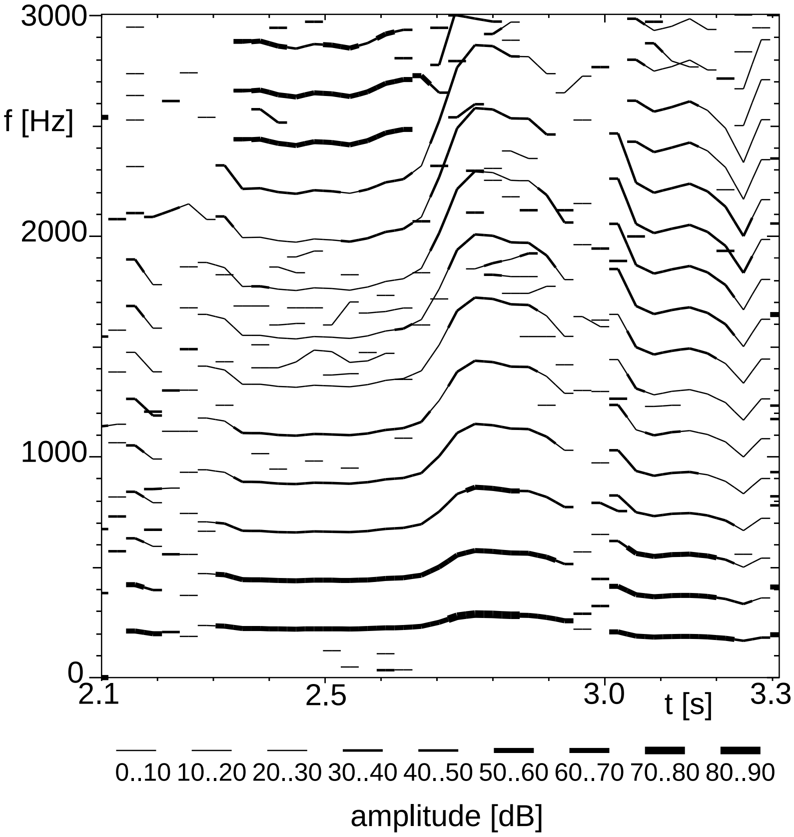

Figure 8: Spectral line pattern visualising the frequency and amplitude parameters of all sinusoidal components modelled by *individual sinusoid* objects and the *harmonic tone* object for a short excerpt from "female vocal."

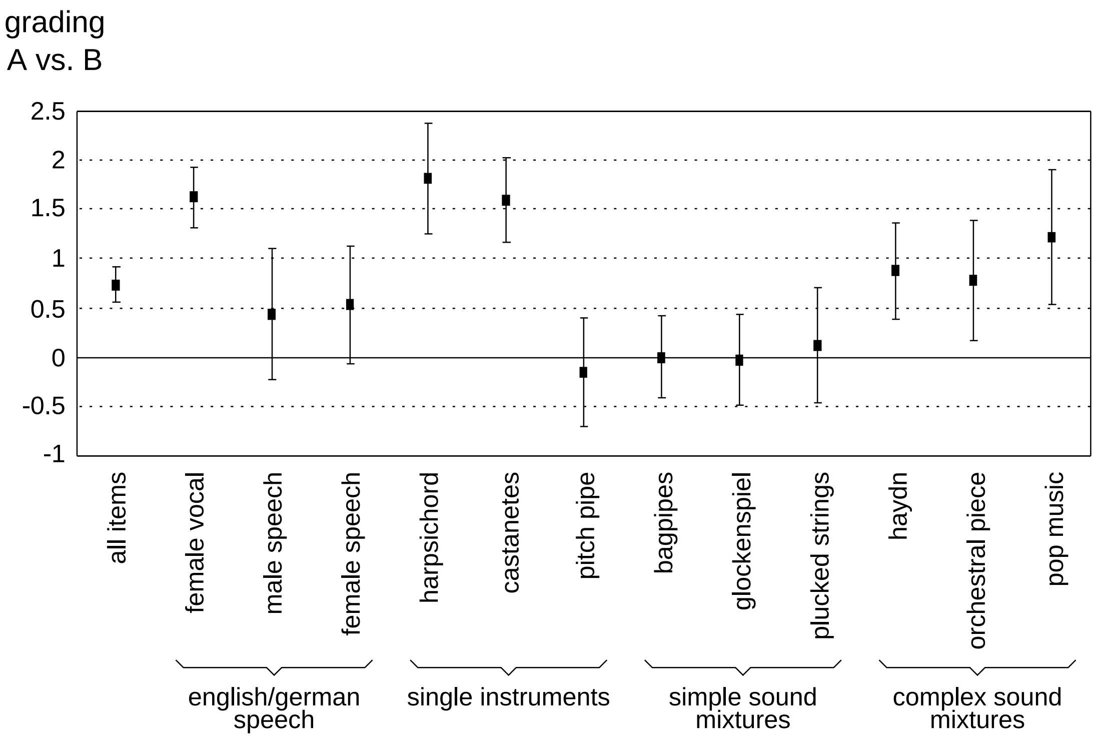

Figure 9: Listening test results for comparison of the extended codec (A) (object models *harmonic tone, individual sinusoid, noise*) vs. the sinusoidal codec (B) (object model *individual sinusoid*) at 6 kbit/s showing average grading and 95% confidence interval based on 8 listeners using the 7-point comparative scale (from [17]).

item	average grading A vs. B	standard deviation	95% confidence interval
all items	0.734375	1.2635716	±0.1787297
female vocal	1.625000	0.6191392	±0.3033722
male speech	0.437500	1.3524669	±0.6626956
female speech	0.531250	1.2311072	±0.6032306
harpsichord	1.812500	1.1383468	±0.5577788
castanets	1.593750	0.8797490	±0.4310685
pitch pipe	-0.156250	1.1212902	±0.5494213
bagpipes	0.000000	0.8563488	±0.4196026
glockenspiel	-0.031250	0.9393038	±0.4602497
plucked strings	0.125000	1.1902381	±0.5832051
haydn	0.875000	1.0082989	±0.4940566
orchestral piece	0.781250	1.2378577	±0.6065382
pop music	1.218750	1.4020075	±0.6869700

Table 1: Listening test results for comparison of the extended codec (A) (object models *harmonic tone, individual sinusoid, noise*) vs. the sinusoidal codec (B) (object model *individual sinusoid*) at 6 kbit/s showing average grading, standard deviation, and 95% confidence interval based on 8 listeners using the 7-point comparative scale (from [17]).

grading	description
3	A much better than B
2	A better than B
1	A slightly better than B
0	A same as B
-1	A slightly worse than B
-2	A worse than B
-3	A much worse than B

Table 2: ITU-R 7-point comparative scale used in the listening test.