TOP-DOWN STRATEGIES IN PARAMETER SELECTION OF SINUSOIDAL MODELING OF AUDIO

Toni Hirvonen and Athanasios Mouchtaris

Department of Computer Science, University of Crete, and
Institute of Computer Science
Foundation for Research and Technology - Hellas (FORTH-ICS)
Heraklion, Crete, Greece

ABSTRACT

Sinusoidal modeling of audio requires the model parameters to be selected by analyzing the original signal spectrum. This paper proposes two improvements in sinusoidal selection by considering how psychoacoustic masking curves can be calculated using a top-down strategy in certain situations. First, a non-iterative component selection method to be used in combination with an added residual signal is presented. Tests indicate computational gain and quality increase when the method is used with a noise-synthesized residual. Secondly, the estimation of the masking curve in binaural listening when signals are panned is considered. Tests show that knowledge of the degree of panning is beneficial when heavy panning is applied to simultaneously rendered audio object signals.

Index Terms— audio coding, sinusoidal modeling, psychoacoustic masking

1. INTRODUCTION

Sinusoidal modeling [1] of audio is one of the most popular parametric audio modeling methods, since it has the capacity to represent an audio signal with good quality by only modeling a relatively small number of spectral components. Some types of sounds cannot be accurately represented by the sinusoidal model. For these cases, an additional component is included (residual part), which models the sinusoidal error signal, *i.e.* the difference between the actual signal and its modeled version [2].

In sinusoidal modeling, energetic masking (due to the human auditory system) is utilized to determine the frequencies of the most perceptually important components [3]. This is usually done in an iterative manner; after selecting one component, the residual magnitude spectrum and the masking curve are updated. At each step, the component frequency that minimizes a perceptual distortion measure is selected. The remaining model parameters (sinusoidal amplitudes and phases) are estimated from the original signal after the frequency selection.

State-of-art approaches for sinusoidal selection such as [3] can be thought of as implementing a bottom-up approach, where no information of the signal reconstruction model or playback conditions are exploited. At each step, the method maximizes the energy of the spectrum that is covered by the masker of the new component. No additional criteria, such as naturalness due to the use of a residual model are considered in this process. The purpose of this paper is to refine the sinusoidal parameter selection process to be more fitting to certain applications in a more top-down manner. The term "top-dow" in this paper implies that instead of using the processing tools independently, we introduce a holistic approach of the reproduction process and conditions that can be used to alter the methods in a way which is beneficial for these particular conditions. This paper proposes two contributions regarding the frequency selection in the sinusoidal model: (a) a *non-iterative* process for estimating the perceptually important frequency components, and (b) masking curve estimation when multiple signals are to be panned before reproduction.

The former contribution is useful when a residual signal is used. In this case, the synthesized energy is close to that of the original signal. Consequently, we show that our non-iterative method for sinusoidal component selection offers an improved sound quality compared to current iterative methods, besides the added computational efficiency. The latter contribution indicates how the sinusoidal selection must be implemented in cases when the audio signals are modeled before mixing occurs. The significance of this result relates to the current efforts in the upcoming MPEG Spatial Audio Object Coding (SAOC) standard [4] and to the possibility of applying the sinusoidal model in this context (see for example [5]). In SAOC, the goal is to encode multiple audio signals before they are mixed into a stereo or multichannel reproduction. This offers the advantage of mixing at the decoder, which is expected to enable a variety of interactive audio applications.

2. NON-ITERATIVE COMPONENT SELECTION

This section discusses an improved method for component selection in the sinusoidal model, for the case when using an additive residual signal. Unlike in Section 3, modeling of single-channel audio is considered in this section.

2.1. Psychoacoustic Sinusoidal Matching Pursuit

Current state-of-the-art methods employ perceptual matching pursuit algorithms to determine the sinusoidal parameters of each frame. In [3], an improved frequency masking model was combined with Psychoacoustical Matching Pursuit (PAMP). At each iteration i PAMP minimizes the perceptual distortion measure D

This work has been funded in part by the Marie Curie TOK "ASPIRE" grant, and in part by the PEOPLE-IAPP "AVID-MODE" grant, within the 6th and 7th European Community Framework Programs respectively.

given by

$$D_i = \int_\omega A_i(\omega)|R_i(\omega)|^2, \quad (1)$$

where $R(\omega)$ is the magnitude spectrum of the residual (original signal minus the selected component), and $A(\omega)$ is a frequency weighting function usually set as the reciprocal of the masking curve energy. Thus, updating the residual magnitude spectrum and the masking curve is required at each iteration.

The analysis is performed on windowed overlapping frames, and the synthesis is here done by summing each component to form the frame signal, and overlap-adding (OLA) them.

In sinusoidal modeling, the synthesis often sounds artificial. To remedy this, a noise residual based on the modeling error between original and synthesis can be introduced. It has been shown that good results can be obtained with *e.g.* the Perceptual LPC (PLPC) method [6]. The residual, being random phase, does not synchronize perfectly but creates "comfort noise" that generally increases naturalness.

2.2. Non-Iterative Method for Sinusoidal Component Selection

When using residual techniques, the energy, and thus the energetic masking properties of the final synthesized signal (*i.e.* sum of the sinusoidal part and residual) are close to that of the original signal. In this light, it is reasonable to use the masking curve of the whole original signal instead of updating it at each step. In addition to being faster, it can be hypothesized that this approach produces better selections of the sinusoidal components than the energy maximizing approach that considers only the masking of the previously selected components. Furthermore, using the improved masking model presented in [3] is convenient with this strategy, since it creates the masking curve by integration with critical band filters, and not by tonal/noise analysis. Thus, the model's speed does not depend on the input signal, and the computational cost of the masking calculations can be reduced directly compared to the amount of components.

Simply using PAMP and replacing $A_i(\omega)$ in (1) with $A(\omega)$, the reciprocal of the masker of the whole frame signal, was in preliminary tests found to produce good results than the traditional PAMP method when a PLPC noise residual was added. However, despite not having to update the masking curve, this still leaves the need for updating the residual magnitude spectrum at each iteration step. To remedy this problem, we propose the following method of Non-iterative Peak Picking (NIPP):

Algorithm 1 Non-iterative Peak Picking

1: $S(\omega) \Leftarrow$ original signal magnitude spectrum
2: $M(\omega) \Leftarrow$ masker energy of $S(\omega)$
3: $\omega_m \Leftarrow$ local maxima of $S(\omega)$
4: pick n first frequencies of $sort(|S(\omega_m)|^2/M(\omega_m))$

This algorithm is performed for each signal frame. It is non-iterative in the sense that no updates are required between component selections. Local maxima refers to the frequency bins where the magnitude is greater than the magnitude of adjacent bins. This step is not computationally expensive compared to the masking curve calculation and it is only applied once in each frame. The desired number of components is n.

2.3. Subjective Evaluation

An A/B/Ref listening test was selected to evaluate the new method. Each subject was presented with a reference, non-modeled signal, as well as signals A and B. These were loudness-aligned sinusoid-modeled signals, one made by applying PAMP, and the other by applying the proposed non-iterative method, NIPP. After listening to each triad of signals, each subject was asked to choose whether A or B was more similar to the reference in terms of sound quality.

The test included seven different monaural signals. These were male and female speech, solo trumpet, a Capella singing, and 3 samples of contemporary music. All were sampled at rate 44.1 kHz. In analysis, the frame size was 23 ms with 50% overlap. FFT size was 16384 with both methods from which 25 component frequencies were chosen. Finally, a noise residual signal obtained with the PLPC-method [6] using LPC order 20 was added to all synthesized samples. Each case was repeated, and so the test included a total of 14 double-blind comparisons. All test signals can be found at our website[1].

Although only cases with 25 sinusoids per frame were included in the test, the relative synthesis quality between the two methods seemed roughly similar with component rates 10-40 as well. Using 25 components was decided upon, since with less or more components, the amount of artifacts caused by the residual tends to increase, or the residual effect becomes less noticeable, respectively.

Table 1. Exp. 1: MP vs. NIPP, with added residual.

Sample		Preference for NIPP (%)
1	Female Speech	89
2	Male speech	100
3	Soft rock	89
4	Heavy rock	67
5	Jazz	67
6	Suzanne Vega	78
7	Trumpet	56

Nine volunteers (authors not included) participated in the test. Results are shown in Table 1. Binomial test showed that NIPP was estimated to produce significantly better quality with samples 1, 2, 3, and 6, with samples 4 and 5 being close to significant. Sample 4 was very compressed and spectrally dense and sample 5 was relatively sparse. This probably made the comparisons more difficult. Sample 7 was almost purely tonal and was included to show that the proposed method is not as beneficial with tonal signals.

Subjects reported more natural sound and less artifacts with the new method, which can be attributed to an improved selection of components when the residual is included. Specifically, NIPP was observed to include less higher frequency components than PAMP, leaving the higher frequencies for the noise residual to reproduce. In addition to arguably being more natural, this is also fitting because the hearing system is less accurate at higher frequencies, especially with phase [7]. Furthermore, in the lower and middle frequency regions it is apparently more beneficial to include more sinusoidal components instead of rely on the noise

[1] http://www.ics.forth.gr/~mouchtar/sines/

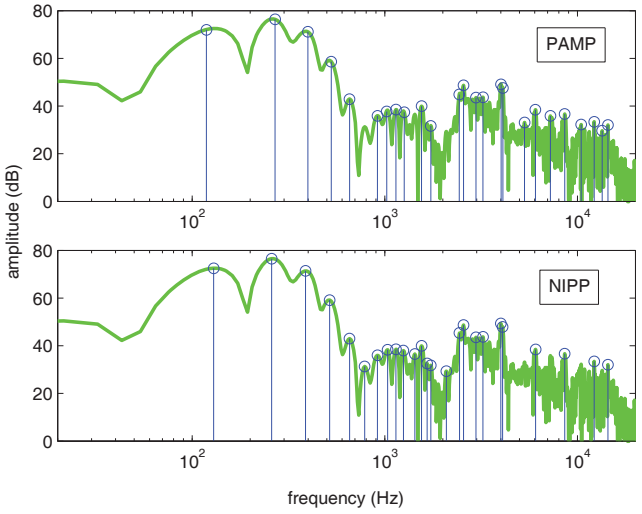

Fig. 1. Comparing selected sinusoidal components by PAMP and NIPP in an example speech frame.

residual, because the phase errors of the noise are more easily detectable and may cause unwanted artifacts.

Fig. 1 illustrates the phenomenon with a typical speech frame; PAMP selects more components at high frequencies, and with NIPP the 0.8-2 kHz region contains more components. The reason for this is that the ratio of some of the high-frequency components to the masking curve of the whole frame, as calculated by NIPP, is not as high as that for the mid-frequency components. With PAMP, the masking curve is only calculated for previously selected components.

3. MASKING CURVES OF PANNED OBJECT SIGNALS

This section considers the need to modify the masking curves in the sinusoidal analysis of two separate auditory object signals, which are subsequently panned and rendered binaurally. Binaural listening in this paper refers to using headphones for reproduction, which makes arbitrary panning possible. It is assumed that the panning factors are known in the modeling process. his implies knowledge of the reproduction conditions. In case of time-varying panning in a real-time application, this knowledge can be accomplished by a feedback process to the encoder.

3.1. Panning in Audio Object Coding

When reproducing largely uncorrelated object signals in a SAOC-type framework, (*e.g.* studio recordings of different instruments), it is possible to manipulate the signals with panning at the receiving end to produce different spatial scenes. For example, given two object audio signals s_1 and s_2, we can create by panning the left and right audio channels (signals c_L and c_R respectively)

$$c_L = s_1 + w_2 s_2 \tag{2}$$

$$c_R = w_1 s_1 + s_2, \tag{3}$$

where it is assumed that initial signals s_1 and s_2 have similar energy level and w_1 and w_2 are attenuation factors (between 0 and 1). In this sense, when reproduced with headphones, signals s_1 and s_2 will arrive at the left and right eardrums of the listener with different levels, and thus with a given (depending on the attenuation factors) spatial image. Although this application considers only the Interaural Level Difference (ILD) and not the Interaural Time Difference (ITD), it is important since in practice most object signals are mixed in this manner.

It can be generalized that typically in SAOC, the masking analysis is done for the sum signal of the objects (*i.e.* the signals are downmixed and compressed with mono or stereo legacy codec) [4]. However, this approach has been applied only in the context of transform coding, while sinusoidal modeling has not, to our knowledge, been considered within the SAOC context. The sinusoidal model has been examined in the context of Spatial Audio Scene Coding (SASC) in [8]. However, in SASC applications the focus is on exploiting channel redundancy in static multichannel audio mixes. Also, [8] did not investigate the binaural synthesis quality or a psychoacoustic-motivated selection of the sinusoid frequencies. Thus, we are interested to test whether using the sum signal for obtaining the masking analysis in sinusoidal modeling is valid for object audio signals. Furthermore, we can consider some theoretical cases when use of the sum signal is not optimal in masking analysis. This is true, for example, when two signals with uncorrelated content are panned to completely opposite headphone channels. Because of minimal inter-channel masking with headphones, the masking should be analyzed separately for each signal. In the following sections we examine whether masking analysis in sinusoidal modeling must be performed based on the sum signal or using some alternative option, when sinusoidal analysis is applied to object signals before panning. Consequently, two different methods are considered for performing the masking analysis when two uncorrelated object signals are panned and reproduced binaurally, and the performance of these methods in relation to the degree of panning is examined.

3.2. Estimation of Mutual Masking Between Panned Signals

We consider two object signals which are analyzed simultaneously but have individual sinusoidal model parameters. As no residual signal is used here, PAMP is utilized for both test methods. The first method examined uses the masker of the sum signal for both objects at each iteration, following the SAOC paradigm. We suggest an alternative method of modifying the masking curves in analysis, according to the knowledge about panning in listening conditions, and compare the two. The proposed modification is to use separate masker curves for both objects. Each masker is the masker of the object signal power summed with an attenuated version of the other object signal masker. This attenuation depends on the amount of panning. When obtaining the masking curve from the sum signal, it can be considered as replacing $A_i(\omega)$ in (1), with

$$A_{1i}(\omega) = A_{2i}(\omega) = 1/(M_{1i}(\omega) + M_{2i}(\omega)). \tag{4}$$

In the second proposed method, $A_i(\omega)$ in (1) is replaced by

$$A_{1i}(\omega) = 1/(M_{1i}(\omega) + w_2^2 M_{2i}(\omega)) \tag{5}$$

$$A_{2i}(\omega) = 1/(M_{2i}(\omega) + w_1^2 M_{1i}(\omega)), \tag{6}$$

where $A_{1i}(\omega)$ and $M_{1i}(\omega)$ are respectively the weighting function and masking energy for the object signal s_1 (and similarly for the

Table 2. Exp. 2: sum (4) vs. masking using panning factors (6).

	Sample	Pan	Preference for use of (6) (%)
1	Female speech and noise	40 dB	78
2	Female and male speech	40 dB	67
3	Classical	40 dB	78
4	Female speech and noise	15 dB	56
5	Female and male speech	15 dB	50
6	Classical	15 dB	72

object signal s_2). Factor w_1 is the attenuation of object signal s_1 in channel c_R caused by panning in (3) (and similarly for w_2). Note that in the extreme case when the two signals are in separate channels, this method reduces to using only the masker of the one object, as is natural. Although the method is applicable to cases where w_1 and w_2 differ, the following analysis, for simplicity, focuses on cases where they are equal. Also, generalization to more than two objects is trivial.

3.3. Subjective Evaluation

An A/B/Ref listening test with nine volunteers was utilized as previously, only now using six different stereophonic test signals. Each stereo test signal was created using two object signals: Female speech plus 2 kHz low-pass noise, female plus male speech, and classical recording with different instruments divided onto two groups. These three cases were mixed with two different panning factors, creating the six stereo signals which were used for the listening test. This arguably small test set was deemed sufficient to illustrate the benefits of the different masking analysis techniques.

The first signal (female speech and noise) represents an artificial extreme case of uncorrelated signals, whereas the other two are more realistic cases of audio object coding. The stereo classical music recording was constructed from individual tracks [9] of the same piece, by panning half of the instruments to left and half to the right. The signal parameters were as in previous test, except no residual and 10 sinusoidal components per object were used. The two objects in all samples were panned to opposite channels either 15 or 40 dB. It is noted that 15 dB is approximately the maximum ILD occurring in free-field listening, whereas 40 dB is closer to total separation of the objects to different channels. Thus there were 6 double-blind comparisons with two repetitions (12 total). As previously, all test signals can be found at our website.

Results are seen in Table 2. According to binomial test, female speech plus noise with 40 dB panning, and the classical sample in both cases produced significant preference towards the proposed method. It can be thus stated that with large interaural level differences are applied to objects, the masker of the sum signal can be sometimes suboptimal. However, with moderate panning, such as 15 dB, the sum masker is a reasonable approximation resulting in mostly similar sound quality as the proposed method. As only two panning values equal to both objects were utilized here, these results can be considered as pilot study. In the future, we plan to make more research to further establish the limitations of using the sum masker.

4. CONCLUSIONS

We have proposed that when information about the listening conditions is available, the psychoacoustic masking analysis in audio coding can be improved in certain situations in a top-down manner. This principle was illustrated in two cases of sinusoidal coding: monaural synthesis with added residual signal, and binaural listening to panned object signals. In the first case, the proposed modifications resulted in notable improvement compared to previous methods. In the second case, we showed that using the sum of the object signals results in a reasonable masking analysis for most practical applications. However, it was found that in case of high amount of panning, the masking analysis will be more accurate if a method which analyzes the signals more independently is used.

5. REFERENCES

[1] R. J. McAulay and T. F. Quatieri, "Speech analysis/synthesis based on a sinusoidal representation," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 34, no. 4, pp. 744–754, 1986.

[2] X. Serra and J. O. Smith, "Spectral modeling sythesis: A sound analysis/synthesis system based on a deterministic plus stochastic decomposition," *Computer Music Journal*, vol. 14(4), pp. 12–24, Winter 1990.

[3] S. van de Par, A. Kohlrausch, R. Heusdens, J. Jensen, and S. H. Jensen, "A perceptual model for sinusoidal audio coding based on spectral integration," *EURASIP J. Appl. Signal Process.*, vol. 2005, no. 1, pp. 1292–1304, 2005.

[4] J. Breebaart et al, "Spatial audio object coding (SAOC) - the upcoming MPEG standard on parametric object based audio coding," *Presented at 124th AES Convention, Amsterdam, The Netherlands*, 2008.

[5] C. Tzagkarakis, A. Mouchtaris, and P. Tsakalides, "A multichannel sinusoidal model applied to spot microphone signals for immersive audio," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 17, no. 8, pp. 1483–1497, Nov. 2009.

[6] R. C. Hendriks, R. Heusdens, and J. Jensen, "Perceptual linear predictive noise modelling for sinusoid-plus-noise audio coding," in *IEEE Int. Conf. on Acoustics, Speech, and Signal Processing, (ICASSP '04)*, 2004.

[7] A. R. Palmer and I.J. Russel, "Phase-locking in the cochlear nerve of the guinea pig and its relation to the receptor potential of the inner hair cells," *Hear. Res.*, vol. 24, no. 1, pp. 1–15, 1986.

[8] M. Goodwin, "Multichannel matching pursuit and applications to spatial audio coding," *Asilomar Conf. on Signals, Systems, and Computers*, 2006.

[9] J. Pätynen, V. Pulkki, and T. Lokki, "Frequency domain acoustic radiance transfer for real-time auralization," *Acta Acustica united with Acustica*, vol. 94, no. 6, November/December 2008.