AN ADAPTIVE TECHNIQUE FOR MODELING AUDIO SIGNALS

Ismo Kauppinen and Kari Roth

University of Turku
Department of Applied Physics
FIN-20014 Turku, Finland
iska@utu.fi, kari.roth@utu.fi

ABSTRACT

In many applications of audio signal processing modeling of the signal is required. The most commonly used approach for audio signal modeling is to assume the audio signal as an (autoregressive) AR-process where the audio signal is locally stationary over a relatively short time interval. In this case the audio signal can be modeled with an all-pole IIR (infinite impulse response) filter, which leads to LPC (linear predictive coding) where the current input sample is predicted by a linear combination of past samples of the input signal.

However, in practice the relatively short time interval (*i.e.* a frame) where the signal is stationary will vary significantly in the audio signal data stream. Also the information content of the frames will show considerable variation. For a proper modeling of an audio signal it is essential that a suitable frame size and appropriate number of model parameters is used instead of a constant frame size and model order.

In this paper we present an adaptive frame-by-frame technique for modeling audio signals, which automatically adjusts the optimal modeling frame size and the optimal number of model parameters for each frame.

1. INTRODUCTION

Linear prediction has been a very popular technique for modeling audio signals for the purpose of *e.g.* effects, speech compression, spectral modeling, and signal reconstruction. The signal is assumed to be an AR-process which is an approximation in the case of real audio signals. In LPC it is necessary to assume that the audio signal is stationary, which is not exactly true for real audio signals. This problem has been overcome by assuming that the audio signals are locally stationary over a short period of time. Music and speech signals contain fast transients and voiced sounds that remain stationary over a relatively long time interval. The signals can be divided into locally stationary sections *i.e.* frames: in the case of a transient or a stop consonant, the length of the stationary frame is very short but for a voiced sound the length can be several times longer. If the signal is modeled frame-by-frame by using a constant predetermined frame length, it is highly possible that the frames will contain several different locally stationary sections which cannot be properly modeled by using the same model coefficients.

The optimal number of the model coefficients is achieved when they contain all the information that the mathematical model is able to receive from the given signal frame.

The remaining text is organized as follows. In Section 2, the mathematical model of the signal is introduced and some of its constraints are pointed out. In Section 3, the optimal frame length decision is presented. In Section 4, the optimization of the model order for a given frame is presented. Conclusions are drawn in Section 5.

2. A MODEL FOR THE SIGNAL

2.1. The mathematical model

A mathematical model for each locally stationary frame of the input signal $x(n)$ is given by

$$x(n) = \sum_i A_i(n\Delta t)\cos(2\pi f_i n\Delta t + \phi_i) + \varepsilon(n\Delta t), \qquad f_i \geq 0, \tag{1}$$

where A_i is the amplitude envelope, ϕ_i is the phase of each frequency f_i, Δt is the sampling interval, and ε is noise. The linear prediction model for an AR-process is given by [1]

$$x(n) = -\sum_{k=1}^{p} a_k x(n-k) + e(n) = \hat{x}(n) + e(n), \tag{2}$$

where a_k are the prediction error coefficients, p is the model order, $\hat{x}$ is the estimated sample and $e(n)$ is a noise-like signal which ideally is uncorrelated and statistically independent of $x(n)$. If the prediction coefficients are known, we can predict the current sample from p previous samples and the forward prediction error is given by

$$e(n) = x(n) - \hat{x}(n) = x(n) + \sum_{k=1}^{p} a_k x(n-k), \tag{3}$$

which we call the residual. For a given signal $x(n)$, where $n = 1, 2, ..., j$ Eq. 3 gives the residual $e(n)$ where $n = p+1, p+2, ..., j$. The first p values of the residual can be approximated by using the backward prediction error given by

$$e(n) = x(n) + \sum_{k=1}^{p} b_k x(n+k) \tag{4}$$

and choosing the backward prediction error coefficients $b_k = a_k$. This can be done when the prediction error coefficients are calculated by using Burg algorithm [2]. There are several possibilities to obtain the prediction error coefficient a_k from the signal. We shall use the Burg algorithm.

2.2. The modeling constraints

It is known that a single noiseless cosine wave with constant amplitude can be perfectly modeled by using two model parameters. The number of the model parameters for a signal consisting of a sum of cosine waves with constant amplitudes is twice the number of the waves. If the amplitude envelope of a cosine wave is not constant in time (which often is the case for music signals), the number of model parameters is higher. For example, a noiseless signal consisting of a single cosine wave with a quadratic amplitude envelope can be perfectly modeled by using five model parameters. If a signal consisting of several cosine waves with constant amplitudes is modeled by using fewer model parameters than is required to perfectly model the entire signal, the cosine waves with the strongest amplitudes will be modeled. [3]

3. ADAPTIVE OPTIMIZATION OF THE FRAME LENGTH

Our method for detecting the optimal frame length is based on forming a long term residual and comparing it to short term residuals obtained from the signal samples given by a sliding window. These detection residuals are achieved by forming low order $(q \sim 50)$ prediction error filters.

The long term residual is formed by using prediction error coefficients $a^{(lt)}$ calculated from the first W samples of the frame whose length is to be optimized. W is also the minimum frame length $(W > q)$. The long term residual is given by

$$e_{lt}(n) = x(n) + \sum_{k=1}^{q} a_k^{(lt)} x(n-k), \qquad n = W+q+1, ..., M. \tag{5}$$

This equation is used to calculate the long term residual beyond the minimum frame length up to the maximum frame length M.

After the first W samples, a sliding window of length N is used $(q < N < M - W)$. In each position of the window new prediction error coefficients $a^{(st)}$ are calculated from the N samples inside the window and they are used to calculate the short term residual within the window. The short term residual is given by

$$e_{st}(n) = x(n) + \sum_{k=1}^{q} a_k^{(st)} x(n-k), \qquad n = m+q, ..., m+N-1, \tag{6}$$

where m is the sample number of the first sample in the current position of the sliding window. A detection value is obtained by comparing the energy of the long term and short term residuals in the position of the sliding window

$$\eta = \frac{\sum_{i=m+q}^{m+N-1} [e_{lt}(i)]^2}{\sum_{i=m+q}^{m+N-1} [e_{st}(i)]^2}, \tag{7}$$

The q term in the summation index is due to the fact that a forward prediction error is used and the first q samples in the window are needed in order to compute the first sample of the short term residual.

If the value of η exceeds unity the modelling parameters of the short term residual will give a better model of the signal than the parameters of the long term residual. If the value of η exceeds a given threshold value λ, the frame end is set to the sample before the current position of the sliding window. Otherwise the window is shifted forward by a step size μ. A new prediction error filter and a short term residual are computed at the new position of the window and η is updated. This procedure is repeated until η exceeds the threshold value λ or the maximum frame size M is reached. When the length of the current frame is found, the whole procedure will start again from the first sample beyond the current frame.

The maximum frame length sets the minimum of the latency in real time applications. The method can be further optimized by using a growing window for the calculation of the long term residual.

In Figs. 1 and 2, the same signal from a guitar is modeled in two different ways. In Fig. 1 the frame length is optimized, resulting in 26% smaller total energy in the residual. In both figures the signal is modelled by using the same amount of frames (21) and in both cases also the same model order $(p = 500)$ for each frame is used. The length of the frames and the model order are printed inside the frames in the signal graph and the residual graph respectively in both figures. The parameters for the frame length decision are presented in table 1.

Table 1: The parameters for the optimal frame length decision.

	Parameter	value
max. frame length	M	10000
sliding window size	N	200
min. frame length	W	1000
step size	μ	50
threshold	λ	2.5
filter order	q	50

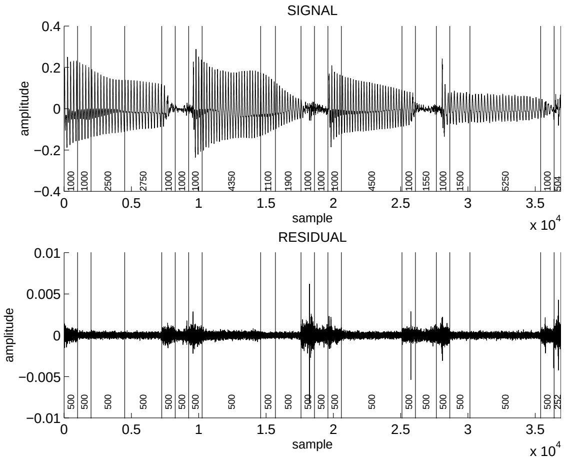

Figure 1: *A guitar signal modeled by using optimized frame lengths and constant model order ($p = 500$). The total energy of the residual is 9.78 in arbitrary units.*

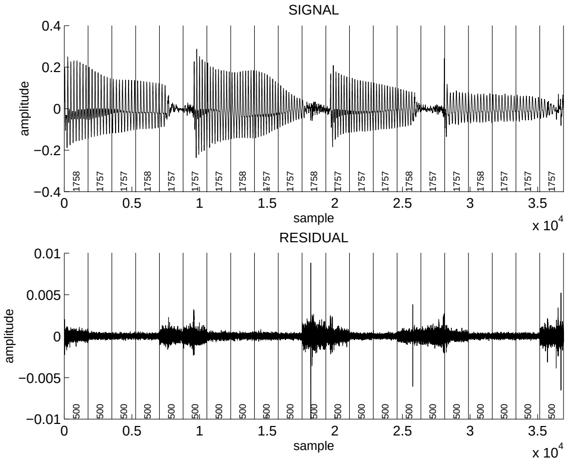

Figure 2: *A guitar signal modeled by using constant frame length and constant model order ($p = 500$). The total energy of the residual is 13.2 in arbitrary units.*

4. CHOOSING THE OPTIMAL MODEL ORDER FOR A GIVEN SIGNAL FRAME

Our method for obtaining the best model order (*i.e.* the number of the modeling parameters) for a given signal frame is based on observation of the power spectrum $E(f)$ of the residual $e(n)$. When the modeled signal is a random autoregressive process of order p, the residual $e(n)$ is white noise [1].

When calculating the model coefficients by using the Burg algorithm, the whole residual vector can be obtained within the process by combining the forward and the backward prediction error. If we model an AR-process of order p by using a smaller number of model parameters $l < p$, then according to the modeling constraints all the frequencies will not be modeled, and therefore the frequencies that could not be modeled will be present in the residual. They will be shown as peaks in its power spectrum. When increasing the model order of the same signal frame and observing the power spectrum of the residual, the optimal model order is achieved as the peaks vanish.

4.1. Detection of the peaks in the power spectrum

A novel peak-detection method is introduced here to enable accurate detection of the peaks in the residual power spectrum $E(f)$. The peak detection is based on observing the absolute value of the derivative of the residual power spectrum given by

$$D(f) = \frac{|E_{f+1} - E_f|}{\Delta f}. \tag{8}$$

Eq. (8) gives the derivative of the power spectrum in a middle point between the two frequencies. To make the peaks more distinct from the rest of the residual power spectrum, the differentiation can be applied several times. In practice, the fourth order derivative has been proven to be very good. By applying Eq. (8) four times in succession and by compensating the shift resulting from the differentiation we obtain

$$D^{(4)}(f) = \frac{|E_{f-2} - 4E_{f-1} + 6E_f - 4E_{f+1} + E_{f+2}|}{(\Delta f)^4}. \tag{9}$$

The optimal model order has been reached when all the peaks fall below a given threshold value. However, a constant threshold level is not very good, since there might be a background in the residual power spectrum.

An adaptive threshold curve can be formed by applying a median filter to the absolute value of the fourth derivative of the residual power spectrum. The median filter is a nonlinear signal enhancement technique for smoothing of signals. The median of a set is defined as the middlemost value of an ordered table of the set values. The median filter has been used for impulsive and random noise suppression of image data [4]. The adaptive threshold curve formed by using the median filter is given by

$$T(f) = a + k(\text{median}(Y_f)) \tag{10}$$

where k is a threshold scaling factor, a is the threshold offset, and $Y(f)$ is a subset of the absolute value of the fourth derivative of the residual power spectrum given by

$$Y(f) = \{D^{(4)}_{f-i}, ..., D^{(4)}_{f-1}, D^{(4)}_{f}, D^{(4)}_{f+1}, ..., D^{(4)}_{f+i}\}, \tag{11}$$

and $2i + 1$ is the length of the median filter.

In the upmost graph in Fig. 3 $D^{(4)}(f)$ is plotted with the model order $p = 0$, *i.e.* the residual is the signal itself. In the middle graph, where $p = 50$, most of the strongest frequencies have vanished from the residual. In the lowest graph all the peaks in $D^{(4)}(f)$ are decreased below the adaptive threshold curve $T(f)$ and the optimal model order is reached. The length of the median filter is 51.

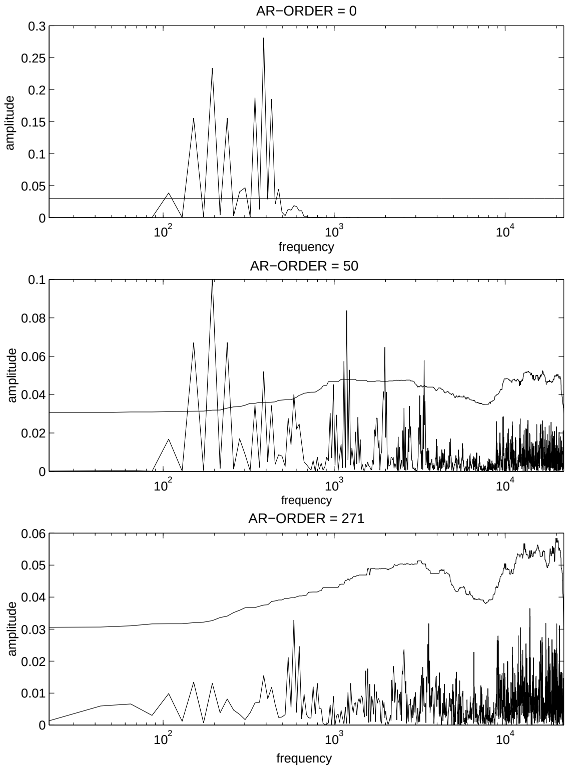

Figure 3: *The optimal model order has been achieved when the frequencies are decreased below the adaptive threshold curve in the fourth derivative of the residual power spectrum.*

5. EXPERIMENTS

In Figs. 4, 5, and 6 the adaptive modeling technique, presented in this paper, is applied to different types of source material. The signals are modeled by optimizing the length of the processing frames. The length of the frames is printed inside the frames in the signal graphs in each figure. Within each frame the minimum number of modeling parameters, which will give optimal results, is searched. Further increasing the number of model coefficients would not result in significant improvement of the model. It would, however, increase the computational complexity proportional to the square of the number of the model coefficients $O(p^2)$. The number of the model parameters for each frame is printed in the residual graph in each of the figures.

In Fig. 4 the automatic decision of the frame size divided the speech signal in frames that contain different sounds in spoken words. For example the third frame contains the letter ”s” from the word ”sound”. This section of the signal is not very close to an ideal AR-process and the optimal model order for this frame is as low as 70. In Figs. 5 and 6 the the frames in guitar and music signals are basically divided from places where the note of the signal changes or a transient, such as drum hit, occur.

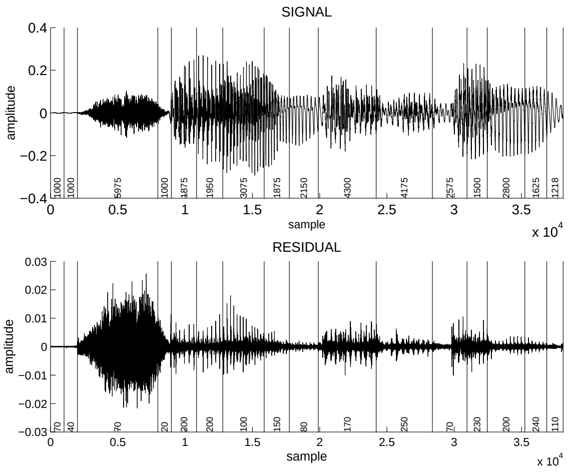

Figure 4: *Speech signal modeled by optimized frame lengths and optimal number of model coefficients for each frame.*

6. CONCLUSIONS

In this paper, we presented an adaptive frame-by-frame modeling technique for audio signals and applied it to different types of source material. The input signal is divided into variable length frames to obtain better starting point for the mathematical model. The signal frames are modeled with LPC prediction error filters. The number of the modeling parameters for a given frame is increased until all the significant information is obtained from the signal.

This modeling technique is suitable for a frame-by-frame real time application, where modeling of the signal is needed *e.g.* audio signal coding and compression, effects, spectral estimation, and noise reduction.

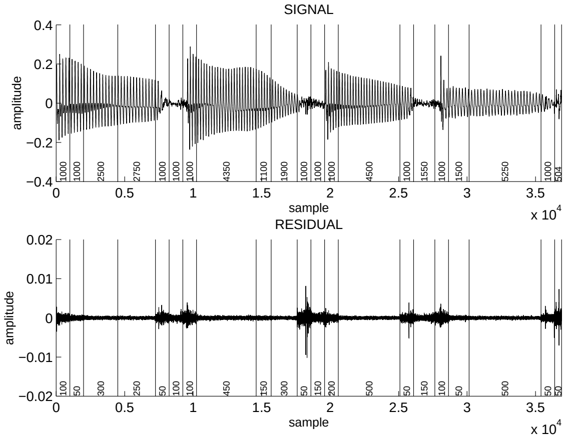

Figure 5: *Guitar signal modeled by optimized frame lengths and optimal number of model coefficients for each frame.*

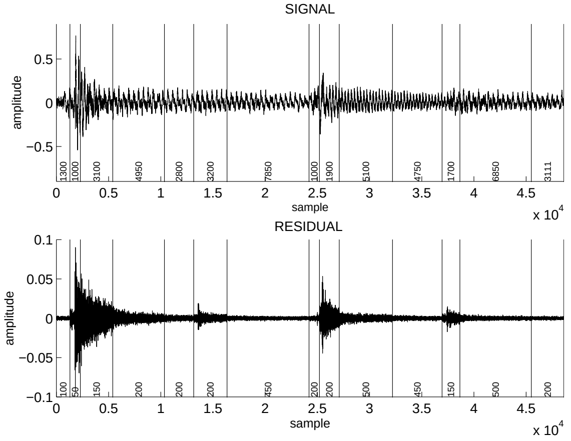

Figure 6: *Music signal modeled by optimized frame lengths and optimal number of model coefficients for each frame.*

7. REFERENCES

[1] Proakis, J. G. and Manolakis, D. G., Digital Signal Processing, Third Edition, Prentice Hall, New Jersey, 1996.

[2] Haykin, S., Nonlinear Methods of Spectral Analysis, Springer-Verlag, Berlin, 1983.

[3] Kauppinen, I., Kauppinen, J. and Saarinen, P., ”A Method for Long Extrapolation of Audio Signals”, accepted to J. Audio Eng. Soc., 2001.

[4] Sebastiani, G. and Stramaglia, S., ”A Bayesian approach for the median filter in image processing”, Signal processing 62, 1997, pp. 303-309.