Adaptive Prediction Order Scheme for AMR-WB+

Fan Zhang*, Takehiro Moriya†, Yutaka Kamamoto†,
Stanisław A. Raczyński*, Noboru Harada†, Nobutaka Ono *, and Shigeki Sagayama *
* The University of Tokyo, Japan
E-mail: {zhang, raczynski, onono, sagayama}@hil.t.u-tokyo.ac.jp Tel: +81-3-5841-6902
† NTT Communication Science Labs, Nippon Telegraph and Telephone Corp., Japan
E-mail: {moriya.takehiro, kamamoto.yutaka, harada.noboru}@lab.ntt.co.jp Tel: +81-46-240-3141

Abstract—**In this paper, we present an adaptive linear prediction order scheme – a simple and effective method to facilitate variable rate coding. We have applied it to the Extended Adaptive Multi-Rate Wideband (AMR-WB+), a state-of-the-art generic speech and audio coder with fixed-length coding designed for wireless transmission. By introducing this scheme into the codec, bit rate can be reduced with a negligible degradation of the quality. The order of linear prediction is adapted based on the audio content, and can be determined by a closed-loop or an open-loop optimal order selection module. The resulted audio quality is examined by average segmental SNR and Perceptual Evaluation of Audio Quality (PEAQ) measure. For this evaluation, speech, music and mixed content audio signals are used. The result proves the effectiveness of this method.**

I. Introduction

Recently, the popularity of multimedia sharing service is growing exponentially, especially in Asian countries. New websites emerge quickly, attracting tens of millions of visitors every day[1]. In these services, bandwidth available for audio content is limited for various technical reasons, so low bit rate codecs that maintain fair audio quality are highly desired. This poses a significant challenge to the audio codec designers: they need to behave well for a vast variety of audio content (speech, music, mixed signals).

Adaptive Multi-Rate (AMR) family codecs – AMR [1], its wideband version AMR-WB [2], and the extended version AMR-WB+ [3] – are one of the state-of-the-art codecs. AMR is designed for GSM and UMTS transmissions of audio signals, therefor are under strict restriction of fixed bit rate, due to high risk of transmission errors in the wireless transmission channel. So for this historical reason, AMR family codecs are all optimized only for fixed-rate coding. On the other hand, we are witnessing a shift from traditional audio transmission systems (such as cellural telephony) to systems designed over the Internet Protocol (IP), which started with the advent of Voice over IP (VoIP) technologies. Internet transmissions allow for variable bit rates and it seems logical that the next extension of AMR family of codecs should include a variable bit rate one.

The focus in this study is the possibility of the quality enhancement of the coders for internet protocol transmission by introducing variable rate and variable length coding into the AMR-WB+. In this paper the state-of-the-art coder AMR-WB+ is used to demonstrate the effectiveness of our adaptive prediction order scheme. AMR-WB+ audio codec is developed by 3rd Generation Partnership Project (3GPP) for GSM and Third Generation (3G) cellular systems. As the name implies, AMR-WB+ is extended from AMR-WB (codified as G.722.2) which is based on AMR. AMR-WB applies Algebraic Code Excited Linear Prediction (ACELP). It is optimized for speech coding. The main improvement by AMR-WB+ is the support for stereo signals, and a transform coding mode (transform coded excitation, TCX), which is included for generic audio coding. This greatly improves music and mixed audio coding performance. Both ACELP and TCX are based on liner predictive coding, and the prediction order is fixed to $r = 16$. So the variable prediction order scheme is beneficial to both of the coder modes. In this paper without going into too much detail about AMR-WB+, the discussion is limited to modules of the codec relevant only to the linear prediction analysis.

II. Background

R. Salami et al. presents in [3] an overall introduction of AMR-WB+. A detailed technical report of AMR-WB is given in [2], which is the same as in AMR-WB+ for the linear prediction analysis modules, and almost the same for the ACELP module. [4] is the standard documentation for AMR-WB+.

A. Linear Prediction in Audio Coding

Human cochlea is believed to act as a short-time spectrum analyzer and most audio coding methods try to immitate its behavior by modeling the audio signal both in frequency- and time-domain. These two aspects – the spectral structure and its temporal envelope – must be properly represented. For a speech signal, spectral fine structure carries information about the pitch and formants, and the envelope is considered to contain most of the linguistic information. Linear Prediction (LP) is one of the most common ways.

LP parameter is attractive in speech and audio coding, because it models the spectral peaks very well, which is an important feature in human auditory perception [5]. In speech particularly, LP coefficient is significant not only to the quality but also the intelligibility of the reconstructed speech.

[1] E.g. youku.com: 140 million daily page views, established 2004; tudou.com: 94 million daily page views, established 2003; 56.com: 28 million daily page views, established 1998; etc. Data obtained from wolframalpha.com

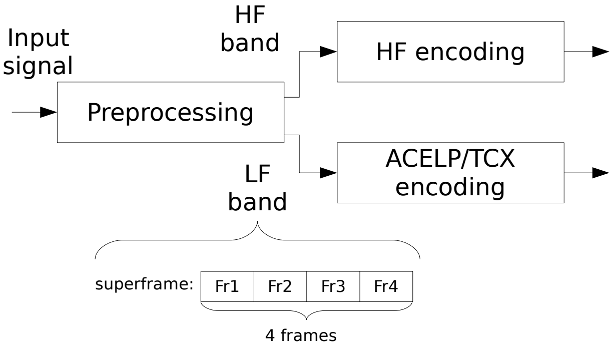

Fig. 1. The high-level structure of the AMR-WB+ encoder, monaural operation modules.

One way to interpret LP analysis is that it removes the redundancy from the input audio by taking out the information from the spectral envelope (which in effect is the linguistic information in case of speech input), and ideally leaves only excitation information for further processing.

To remove the spectral envelope information properly, LP analysis of sufficient order is performed. By selecting this order as low as possible, a fair amount of bits can be saved later in the LP coefficient quantization.

Fig. 1 depicts the high level structure of the AMR-WB+ encoder. The input sequence is first resampled to the internal sampling frequency F_s and then segmented into blocks containing $2N$ samples. Each block is then separated into two complementary bands: the low-frequency (LF) signal and the high-frequency (HF) signal band, both critically sampled at $F_s/2$ and containing N samples. These bands are called superframes. The LF band is treated using the core coder based on an ACELP/TCX switching structure, where the ACELP mode is the same as AMR-WB, and the HF band is encoded with only few bits by means of the bandwidth extension (BWE) method.

B. AMR-WB+ LP Analysis Procedure

The low frequency (LF) band superframe is divided into four frames and every frame is further divided into four subframes. For every superframe, AMR-WB+ core coder chooses between several coding modes: ACELP (one frame), short TCX (one frame), medium TCX (first two or last two frames), and long TCX (all four frames, i.e. the whole superframe). Table I lists all 26 possible mode combinations. Each cell in Table I represents one superframe; the four numbers (m_0, m_1, m_2, m_3) represent the mode combination options for each frame, where:

- $m_k = 0$ means the mode for frame k is ACELP
- $m_k = 1$ means the mode for frame k is short TCX
- $m_k = 2$ means the mode for frame k is medium TCX (Two consecutive frames)
- $m_k = 3$ means the mode for frame k is long TCX (Four consecutive frames)

TABLE I
POSSIBLE MODE COMBINATIONS IN A SUPERFRAME

(0, 0, 0, 0)	(0, 0, 0, 1)	(2, 2, 0, 0)	
(1, 0, 0, 0)	(1, 0, 0, 1)	(2, 2, 1, 0)	
(0, 1, 0, 0)	(0, 1, 0, 1)	(2, 2, 0, 1)	
(1, 1, 0, 0)	(1, 1, 0, 1)	(2, 2, 1, 1)	
(0, 0, 1, 0)	(0, 0, 1, 1)	(0, 0, 2, 2)	
(1, 0, 1, 0)	(1, 0, 1, 1)	(1, 0, 2, 2)	
(0, 1, 1, 0)	(0, 1, 1, 1)	(0, 1, 2, 2)	(2, 2, 2, 2)
(1, 1, 1, 0)	(1, 1, 1, 1)	(1, 1, 2, 2)	(3, 3, 3, 3)

By default the mode combination is decided in a closed-loop fashion, i.e. all mode combinations are attempted and the one with the best Segmental SNR (SSNR) is chosen.

The LP analysis module is common for both ACELP and TCX modes. A good reference about LP analysis can be found in [6], [7]. For any of the four modes, LP parameter is encoded and transmitted once. That is, for a superframe, a maximum of four LPs and a minimum of one LP is actually encoded. So, when calculating the average bit reduction, the coder mode information also needs to be taken into account.

C. LP Coefficient Quantization

The quantization of LP coefficient is one of the central issues in the adaptive prediction order scheme design. It is crucially important that the LP coefficient quantization process is accurate. [8] includes a comprehensive overview and comparison of various LP parameter representations as well as quantization schemes. In AMR-WB+, LP parameter is transformed to Immittance Spectrum Frequency (ISF) [9] for quantization and interpolation purposes.

III. Adaptive Prediction Order Scheme

The basic idea of adaptive prediction order is that a fixed prediction order ($r = 16$ for the case of AMR-RW+) is not optimal for a considerable portion of the target signal for almost any popular multimedia audio broadcast content. For those frames (proved to contain silent, noise or music), the use of a lower order in the linear prediction analysis is sufficient to model the power spectral envelope and therefore the use of a lower prediction order will not affect the perceived audio quality. As a consequence, bit rate reduction can be achieved, because the quantizer can use fewer bits to store for lower order compared to the original split-multistage vector quantizer.

A. Optimal Order Selection Methods

Two different criteria are tested in order to decide if lower order LP is acceptable: SSNR-based order selection method and a method based on log-spectral distortion measure (LSDM).

In the SSNR-based selection, the encoding-decoding process of every optional LP order is performed for every frame, and the LP order that yields a better SSNR is chosen. I.e. if the SSNR from the lower prediction order exceeds a threshold of the SSNR produced by the original sixteenth order LP

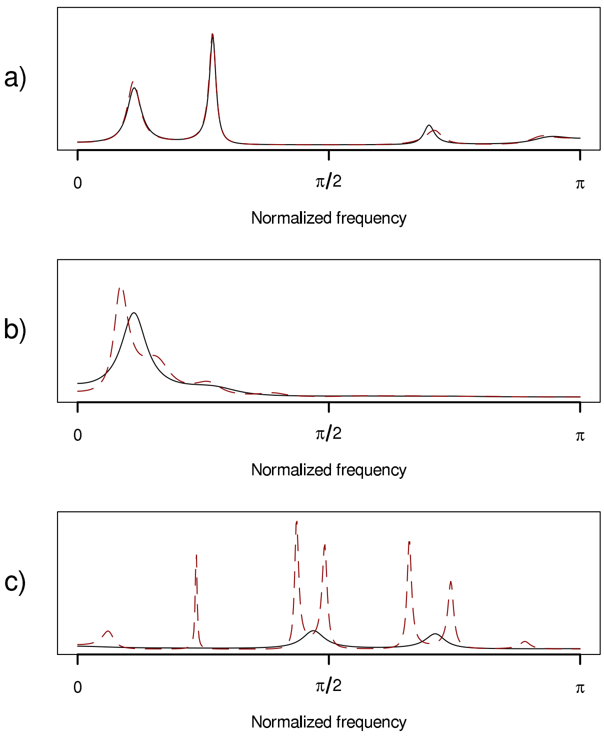

Fig. 2. Comparison of power spectra obtained with LP of order 16 and 8. Dashed line represents the results of the higher prediction order, and solid line the lower order. Three plots show three cases when lower order is acceptable (a), the threshold case, i.e. when switching between orders occur (b), and the case when higher order is chosen (c).

analysis, then the lower prediction order is considered acceptable. The reason that SSNR-based criterion is used instead of perceptually motivated measure, is because of computability considerations. In the original coder, SSNR is used for coder mode (ACELP/TCX) selection. The same routine can be easily reused for the optimal order selection with little increase in computational cost.

In the SSNR based selection method, the encoding-decoding process is performed for every possible LP order for each frame and the LP order that yields a sufficient SSNR is chosen: if the SSNR from the lower prediction order exceeds a threshold set relative to the SSNR produced with the sixteenth order LP, then the lower prediction order is deemed acceptable. Even though a more sophisticated method, like e.g. a perceptually motivated evaluation measure, would be more precise in assesing the performance of different LP orders, we have chosen SSNR for its computational simplicity.

The LSDM-based order selection method compares the spectral difference between the sixteenth and a lower order LP coefficients. Such mean square log-spectral distortion measure is very commonly used to evaluate LP coefficients quantization performance [8], [10]. If the two spectra are similar enough, then the lower prediction order is selected. Plots in Fig. 2 give an intuitive illustration of this criterion.

The SSNR-based approaches bears an intrinsic computability issue. Obviously, in order to compare the SSNR, a single audio segment have to be encoded and decoded several times (for each possible option). On the other hand, LSDM-based method is an open-loop selection, which means that the optimal order is selected before the encoding process, and encoding is performed only once using the optimal order is chosen for each frame. In this way, the computational cost is reduced greatly when compared to SSNR-based approach, which makes this method much more practical for real-world applications.

The order selection is performed after mode selection step for both SSNR and LSDM-based methods. So even though experimental results suggest that for some frames it can happen that the eighth order TCX is better than the sixteenth order ACELP, this is not considered in the mode selection process. The main reason is that if we considered the effect of prediction order selection when selecting the encoder mode, the computational cost would increase explosively with the number of available LP orders.

IV. Practical Issues

A. LP Parameter Quantizers

AMR-WB+ uses split-multistage vector quantization with 46 bits for ISF quantization. It is the same as the quantizer used in AMR-WB, and the detailed description is presented in [1]. The eighth order ISF quantization uses a 16-bit split vector quantizer.

Fig. 3 shows the LSDM distribution for both the original AMR-WB+ sixteenth order LP parameter quantizer (in solid line), and the lower order LP parameter quantizer used in the experiment (in dash-dotted line). The average spectral distortion of the split-multistage vector quantizer in AMR-WB (which is the same as in AMR-WB+, but only for speech signal) is estimated to be 0.894 dB [8]. As expected it is a little worse for generic audio contents. Compare to the highly optimized quantizer used in AMR-WB+, the performance of the lower order quantizer is not as good, but still is adequate to show the advantage of adaptive prediction order scheme in the experiment.

B. Order Selection Based on SSNR

SSNR-based order selection chooses the lower order if the SSNR exceeds a threshold of the higher order result. Informal listening tests show that a threshold of 0.95 renders audio qualities that are almost the same as the original coder. (Which means that the lower prediction order is selected, if the SSNR is larger than 95% of the SSNR resulted from the original sixteenth prediction order mode.) An obvious drawback of this approach is that the threshold is not directly related to the bit reduction result. By lowering the SSNR threshold, more percentage of frames will be decided as suited for lower prediction order, and bit rate will be reduced consequently. Table II shows the percentage (of frames) that's selected as lower order, based on the 95% threshold and the bit rate used. This table is the most important result from the SSNR-based method experiment, which is the average lower order usage

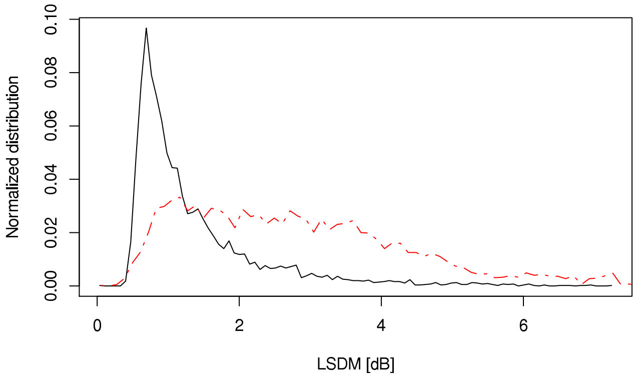

Fig. 3. The distribution (normalized histogram) of log-spectral distortion measure values for the lower order LP coefficient quantizer (the dash-dotted line) and for the original LP coefficient quantizer (the solid line).

TABLE II
AVERAGE PERCENTAGE OF FRAMES SELECTED AS LOWER PREDICTION ORDER

bit rate (kbps)	6	9	12	16
Lower order (%)	25.12	22.51	22.22	23.96
bit rate (kbps)	19	24	26	32
Lower order (%)	23.37	21.77	22.80	22.45

depending on the bit rates. This result is used as a reference parameter in the LSDM-based implementation.

Fig. 4 is an example of the SSNR-based order selection. This plot is calculated from a part of a German speech test sequence output.

Two implementations are experimented for the SSNR-based order selection scheme. First is closed-loop re-encoded order selection. In this implementation after order selection the audio is encoded yet again using the optimal order, in order to solve most of the order-switching-caused problems such as improper memory updates. The encoding process is noticeably slower than AMR-WB+, which is expected since the actual encoding-decoding process is performed several times in the order selection. The second SSNR-based order selection, in comparison, does not re-encode the input signal, but simply copies parts of resulted encoded parameters from sixteenth or lower order mode according to the optimal order selected. This implementation is faster than the first one, but bears the problem that the encoder and decoder do not have the same sets of parameters when LP order switches.

The LP parameter interpolation in particular, is a problem for both methods. In the decoder when the prediction order switches, the new order LP parameter of the previous frame is unknown. This is solved by estimation or approximation. And if the prediction order switches very frequently, a significant drop in SSNR and audible switching artifacts is resulted.

C. Order Selection Based on LSDM

LSDM-based order selection chooses the lower order if the lower order LP spectrum approximates higher order spectrum close enough. The threshold is calculated based on a wide range of input signals, speech (of four languages, with and without environmental noise, male and/or female, etc), music (pop, classic, vocal, clean, noisy, etc), and mixed content (speech and music, speech over music, under different conditions). First the different orders of LP parameters are calculated, and then the LSDM values are collected, based on the histogram of the LSDM values, a threshold of the LSDM is determined according to TABLE II. That is, for each bit rate, a LSDM threshold can be selected as the threshold, (to control the percentage of frames that is going to be selected as lower order.) In a first experiment, the LSDM threshold is set to a closest larger value of the reference percentage from Table II. Table III shows the selected threshold percentage, the LSDM threshold (an internal value similar to dB but not), and the corresponding bit reduction results.

TABLE III
THE PERCENTAGE OF LOWER ORDER FRAMES, AND RESULTED AVERAGE BIT REDUCTION OF LSDM-BASED EXPERIMENTS.

bit rate (kbps)	6	9	12	16
Lower order (%)	25.67	22.69	22.69	24.53
Reduction (%)	3.22	2.79	2.42	2.09
bit rate (kbps)	19	24	26	32
Lower order (%)	23.82	22.07	23.27	22.69
Reduction (%)	1.83	1.46	1.44	1.21

The LSDM threshold resulted from Table III for various bit rates are very close, and from the experiment it seems to be safe to rise the threshold a little. In an other experiment, the threshold is chosen to be the smallest upper bound of all the thresholds, and the average lower order adoption rate is 26.67% for all bit rates.

D. Switching Effect

As is the case for any switching scheme based coder, the problem of switching artifact must be considered. In this study switching artifact results from two causes, one is the parameter mismatch for different orders, the other is because of some of the procedures in AMR-WB+ codec (e.g. several empirical coefficients) are optimized specifically for the sixteenth order LP.

The first cause of switching artifacts, parameter mismatch, refers to the ISF parameter interpolation between current and previous frames (for each subframe). Obviously when prediction order switches, the current order LP coefficient is not available to decoder, estimation or approximation has to be used. Several approaches are tested.

The simplest one is, if the LP order is switched, then no interpolation is performed, instead the current parameter is used repeatedly for the first four subframes. Other types of solutions try to estimate the parameters of the current order from the previous frame parameters (of a different order). This simple solution gives better performance for the SSNR-based approach. It appears that it is difficult to estimate the current order parameters from the previous ones with a satisfactory accuracy. No one method is better than the others all the time, unpredictable anomalies occur from time to time, causing

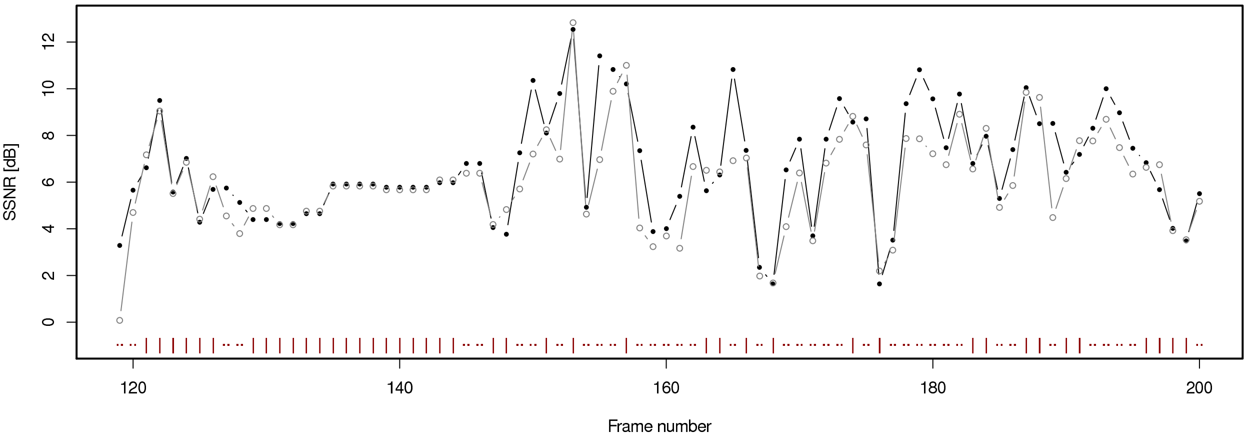

Fig. 4. An visualization of SSNR-based order selection for an example data. Solid line shows the SSNR obtained for sixteenth order prediction, while dash-dotted line for the eighth order. The vertical bars at the bottom indicate frames where lower prediction order is selected, i.e. eighth order SSNR is larger than 95% of the sixteenth order SSNR.

audible artifacts in the reconstructed audio. Although the order estimation is not giving the best result, it is obviously a meaningful task for the adaptive LP order scheme. To improve the accuracy of lower order LP parameter estimation from a higher order version is one of the most interesting tasks for the adaptive LP scheme.

For LSDM-based approach, restriction on the choice of prediction order for long TCX frames performs well. By setting the prediction order consistent to previous frame, the parameter mismatch is avoided and the resulted SNR, Perceptual Evaluation of Audio Quality (PEAQ) [11] results are fairly good.

The second cause of switching artifacts, is the sixteenth order specific empirical parameters. The only option is the ad-hoc style solutions, which mean to observe the parameters resulted from different LP orders, compare and analyze for the correct output, then modify the empirical parameter to perform different process, or use different empirical parameters according to the current LP order. This process takes a long time and is difficult to verify.

V. Experimental results

A. Quality Assessment with Average SNR

Fig. 5 illustrates the bit reduction for both the SSNR and LSDM-based approaches. The first 8 bars in each group show the SSNR-based bit reduction, corresponding to 8 different sampling frequencies. For lower sampling rates, like 8 kHz or 11.025 kHz, adaptive order scheme actually increases the bit-rate for higher bit rate encoder modes. We believe that the reason sampling frequency is relevant to bit-reduction is because for lower sampling rates there is less correlation in consecutive samples, so it is more difficult to predict a sample value using a small number of previous samples (i.e. using a lower prediction order), and so there is a dramatic decrease in the number of frames encoded with lower order, but 1 bit is still needed to signal the order selection, these two factors together cause the bit rate increase. The order selection indication bit could be reduced to 1 bit for each encoder mode (1 bit/frame for ACELP and short TCX, 1 bit for two frames for medium TCX, and 1 bit/superframe for long TCX).

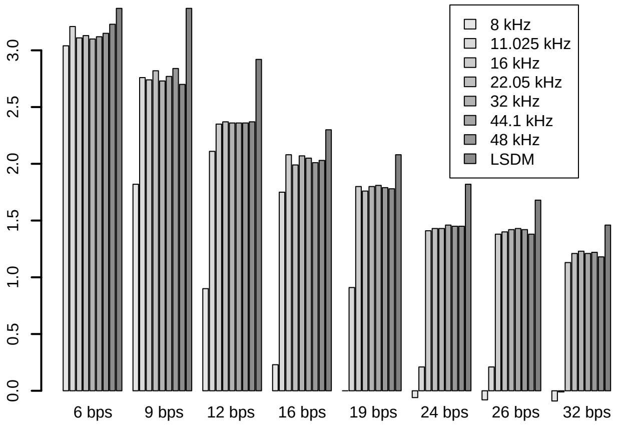

Fig. 5. The average bit reduction (in percent) for both SSNR-based (rst 8 bars in each group) and LSDM-based (the right most bars in each group) adaptive prediction order selection, for 8 different sampling frequencies (SSNR-based only) and 8 different bit rates.

From Fig. 5 it can be observed that for audio signal with sampling frequency larger than 16 kHz, the bit reduction (thus percentage of lower order frame) is similar. As explained previously the bit rate increase for 8 and 11.025 kHz at higher bit rates (the negative bit reductions) appeared because of low percentage of lower order use – if lower frames is used less than 3.33% of the total number of frames, the bit rate is

TABLE IV
AVERAGE SNR REDUCTION

Sampling rate	8 kHz	11.025 kHz	16 kHz	22.05 kHz
SSNR-based	0.5488	0.5058	0.7365	0.7697
LSDM-based	0.3799	0.4007	0.5049	0.4047
Sampling rate	24 kHz	32 kHz	44.1 kHz	48 kHz
SSNR-based	0.4300	8.9556	8.3199	10.1908
LSDM-based	0.1424	0.2768	0.2654	0.2049

TABLE V
AVERAGE PEAQ REDUCTION

Sampling rate	8 kHz	11.025 kHz	16 kHz	22.05 kHz
SSNR-based	0.0970	0.1285	0.2433	0.2985
LSDM-based	0.1274	0.1050	0.2406	0.2850
Sampling rate	24 kHz	32 kHz	44.1 kHz	48 kHz
SSNR-based	0.3854	0.9188	0.9185	1.1683
LSDM-based	0.3530	0.3857	0.3820	0.3855

going to increase – the bits used to signal the prediction order will outnumber the bit reduction from lower order prediction coefficients quantization and cause the bit rate to increase.

Table IV shows the average SNR reduction as a result of the bit reduction, which demonstrates the audio quality. As is discussed previously, the average SNR is used because the order selection method is based on SNR comparison. As a reference the SNR reduction resulted from the LSDM-based method is also given. The average SNR of higher frequencies tests SSNR-based results drop significantly. This is due to the fact that with a combination of high sampling and bit rates, the use of medium and long TCX encoding mode increases dramatically, and the parameter mismatch problem in earlier discussion becomes severe, which means that for a good portion of the frames the LP coefficients can only be estimated from parameters of a different order.

B. Quality Assessment with PEAQ

Table V shows the average PEAQ reduction for different approaches. The average bit reduction is the same as Fig. 5 shows. Table V shows that the LSDM-based prediction order selection is almost always better than the SSNR-based approach, which gives more bit reduction and better audio quality.

VI. CONCLUSION

Aiming at reducing bit rates of speech and audio coders for Internet transmission, we have presented an adaptive prediction order scheme to the state-of-the-art audio codec AMR-WB+ in order to introduce variable rate coding feature to the codec that was originally designed for fixed rate mobile transmissions. Evaluation results suggest that by making the prediction order adaptive, the bit rate can be reduced around 2%, depending on the selected bit rate, with a near transparent audio quality compare to the original codec with the LP order fixed to $r = 16$.

Our current concern is to enhance the open-loop order prediction, lower order LP parameter estimation, and bit reallocation module. With the aid of bit reservoir and sophisticated adaptive bit allocation within and among frames, reduced bits can be expected to contribute quality enhancement under the same average bit rates. The result are planned to be reported shortly.

REFERENCES

[1] 3GPP TS 26.190: *Adaptive Multi-Rate Wideband speech codec; transcoding functions.*

[2] B. Bessette, R. Salami, R. Lefebvre, M. Jelinek, J. Rotola-Pukkila, J. Vainio, H. Mikkola, and K. Jarvinen, “The Adaptive Multirate Wideband Speech Codec (AMR-WB)”, *IEEE Trans. On Speech and Audio Processing,* vol. 10, no. 8, November 2002.

[3] R. Salami, R. Lefebvre, A. Lakaniemi, K. Kontola, S. Bruhn, and A. Taleb, “Extended AMR-WB for high-quality audio on mobile devices,” *IEEE Communication Magazine,* vol. 44, no. 5, pp. 90–97, May 2006.

[4] 3GPP TS 26.290: *Extended Adaptive Multi-Rate Wideband codec; Transcoding functions.*

[5] L. Rabiner and R. Schafer, *Digital Processing of Speech Signals,* Prentice-Hall, Inc.: Englewood Cliffs, NJ, 1978.

[6] P. Vary and R. Martin, *Digital Speech Transmission.* John Wiley & Sons, Ltd, 2006.

[7] B. Kleijn and K. Paliwal, *Speech Coding and Synthesis,* Elsevier Science Inc. New York, NY, USA, 1995.

[8] S. So and K. Paliwal, “A Comparative Study of LPC Parameter Representations and Quantization Schemes for Wideband Speech Coding,” *Digital Signal Processing,* vol. 17, issue 1, pp. 114–137, January 2007.

[9] Y. Bistritz and S. Pellerm, “Immittance Spectral Pairs (ISP) for Speech Encoding,” *IEEE Int. Conf. Acoustics, Speech, Signal Processing (ICASSP),* pp. II-9–II-12, 1993.

[10] A. Gray and J. Markel, “Quantization and bit allocation in speech processing,” *IEEE Trans. Acoust., Speech, Signal Process.,* vol. ASSP-24, pp. 459–473, 1976.

[11] ITU-R BS. 1387.1, *Method for Objective Measurements of Perceived Audio Quality.*