DESIGN OF A TOLL-QUALITY 4-KBIT/S SPEECH CODER BASED ON PHASE-ADAPTIVE PSI-CELP

Kazunori Mano

NTT Human Interface Laboratories,
3-9-11, Midori-cho, Musashino-shi, Tokyo 180, Japan
E-mail: mano@splab.hil.ntt.co.jp

ABSTRACT

This paper describes the design of a toll-quality 4-kbit/s speech coder based on phase-adaptive PSI-CELP. This adaptation method not only gives pitch periodicity to the random excitation but also synchronizes the basic point of the stored random vector with the pitch phase. We further improve the proposed coder by introducing a backward gain prediction scheme. In subjective evaluation experiment, there is no significant difference between the quality of ITU-T G.726 32-kbit/s coder and that of the proposed 4-kbit/s coder under the conditions of normal and low input levels, tandem connection for clean speech. In noisy environment, there are also no significant differences between G.726 and 4-kbit/s coders from MOS results of ACR test.

1. INTRODUCTION

The ITU-T recently began the standardization process for a new 4-kbit/s coder for low-bitrate visual telephony, personal communication, and mobile communication.

In the conventional CELP[1], speech signal is synthesized by filtering and gain-scaling excitation signals through a linear prediction (LP) synthesis filter frame by frame. The excitation signals are composed of a codevector selected from an adaptive codebook and a codevector from a random codebook. The synthesis filter is coded by an open-loop method, while the adaptive excitation code, random excitation code, and gain code are determined by using a closed method referred to Analysis-by-Synthesis (A-b-S) method, which finds best codes to minimize a perceptually weighted distortion between input speech and synthesized speech.

At bit rates at or below 4 kbit/s, speech coded by the conventional CELP coder badly degrades due to the lack of pitch periodicity in the random codevectors.

We previously proposed Pitch Synchronous Innovation CELP (PSI-CELP) coding[2][3], in which random codevectors are adaptively converted to have pitch periodicity for voiced frames. Similar pitch synchronous techniques for excitation signals are introduced in ITU-T Recommendations G.723.1(MP-MLQ/ACELP) and G.729 (CS-ACELP).

We have also proposed a new excitation coding method called Phase-Adaptive PSI-CELP[2][4]. The subjective evaluation results show that the coder has potential abilities to be a candidate for the ITU-T standard coder.

2. OVERVIEW OF THE PROPOSED CODER

Figure 1 shows the basic encoder structure of phase-adaptive PSI-CELP. The coder sequentially processes speech in 20-ms frames, and the subframe length is 10 ms.

From the encoder, four transmission codes of LPC parameter code (A), codebook 1 excitation index (L), codebook 2 excitation index (C), and gain parameter code (G) are transmitted to the decoder.

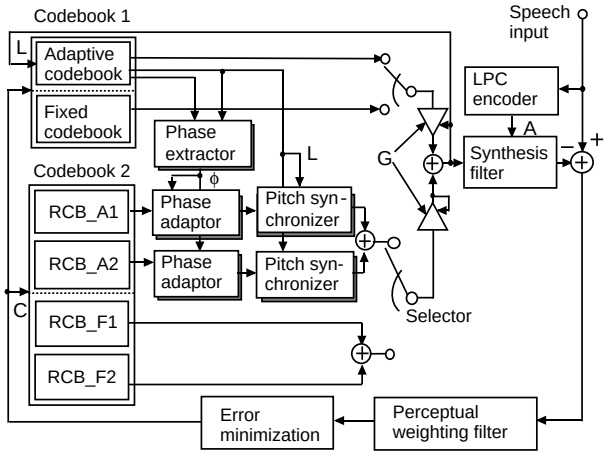

Figure 1. The block diagram of the phase-adaptive PSI-CELP

The LPC encoder quantizes the LPC parameters from LPC analysis. This quantized LPC parameter is used as synthesis filter coefficients. The synthesis filter is excited by the excitation signals consisting of the codebook 1 vector and the codebook 2 vector. These excitation vectors are selected by using an Analysis-by-Synthesis (A-b-S) search method which minimizes the perceptually weighted error between the input speech and the synthesized speech.

The excitation codebook 1 has an adaptive codebook and a fixed codebook. The adaptive codevectors stored in the adaptive codebook are updated from previous excitation signals in each subframe. This codebook encodes the pitch interval. The fixed codebook is a set of random vectors prepared for silent, unvoiced, and transient parts of speech.

In the codebook 2 search, the phase extractor, the phase adaptor, and the pitch synchronizer are activated if the adaptive codebook is selected in the codebook 1 search. If the fixed codebook is selected in the codebook 1 search, the random codevector in the codebook 2 is not phase-adapted and not pitch-synchronized. This switching function mainly corresponds to voiced frames and unvoiced frames. Trained random codebooks RCB_A1 and RCB_A2 are used when the adaptive codebook is selected, and RCB_F1 and RCB_F2 are used when the fixed codebook is selected.

A PSI-CELP random codevector, that is synchronized with the adaptive codevector, is generated by repeating elements of pitch interval length L that are obtained from the adaptive codebook search. The pitch repetition scheme improves the quality of voiced parts of speech. In addition to the pitch repetition scheme, the phase adaptation method described in section 4.2. further reduces the quantization

distortion at low bitrate.

The gain parameters for the codebook 1 and 2 vectors are vector-quantized with a backward gain prediction.

Decoding process consists of the synthesis part and the postfiltering part.

Table 1 shows the bit allocation for the proposed coder.

Table 1. Bit allocation in 20-ms frame.

Parameter	bits/frame
LPC parameters	20
Adaptive CB/Fixed CB (Codebook 1)	8 x 2
Random CB (Codebook 2)	(7+7) x 2
Gain	8 x 2
Total	80

3. LPC PARAMETER ENCODING

The LPC parameters are quantized in the Line Spectrum Pair (LSP) domain. As LSP parameters between successive frames are highly correlated, an inter-frame moving-average prediction and a multistage vector quantization are used.

$$\boldsymbol{\Omega}_n^{(2)} = \overline{\boldsymbol{\Omega}} + \sum_{i=1}^{M} \boldsymbol{B}_i \boldsymbol{\omega}_{n-i} + \boldsymbol{B}_0 \boldsymbol{\omega}_n, \tag{1}$$

$$\boldsymbol{\Omega}_n^{(1)} = (\boldsymbol{\Omega}_{n-1}^{(2)} + \boldsymbol{\Omega}_n^{(2)})/2, \tag{2}$$

where $\boldsymbol{\Omega}_n^{(2)}$ and $\boldsymbol{\Omega}_n^{(1)}$ are

respectively reconstructed LSP parameters at the second and first subframes at time n. The order of the LSP parameter is 10. $\overline{\boldsymbol{\Omega}}$ is an average offset vector. $\boldsymbol{\omega}_n$ is quantized by a multistage vector quantization (VQ) and transmitted as a LPC code at time n. In the first stage quantization, the 10 dimensional VQ is performed. In the second state, the vector is split into a lower dimensional part and a higher dimensional part, and these subvectors are quantized. $\boldsymbol{B}_i, i = 1, \ldots, M$ are matrices of prediction coefficients, which have only non-zero elements on the diagonal. In the current coder, the prediction order M is 2.

4. ENCODING OF EXCITATION

4.1. Adaptive and fixed excitation

Codebook 1 has an adaptive codebook the same as in the conventional CELP, and it also has a fixed codebook. 8 bits (256) codes are assigned to codebook 1. 192 codes are used to represent the adaptive code and 64 codes for fixed codevectors. The adaptive codebook encodes non-integer pitch delays. The output excitation of the codebook 1 is generated by repeating an excitation of pitch length of the previous subframe. The fixed codebook consists of trained vectors and Gaussian noise vectors. This fixed codebook allows the coder to improve the unpredictable parts of speech, such as silent, unvoiced, or transient parts of speech. This codebook also makes the coder robust against additive non-periodic speech and background noise.

4.2. Phase-adaptive random excitation

The distortion measure d of CELP coding is represented as follows.

$$d = \|\boldsymbol{W}(\boldsymbol{x} - \boldsymbol{x}_{zir} - g_{adp}\boldsymbol{H}\boldsymbol{e}_{adp} - g\boldsymbol{H}\boldsymbol{e})\|^2 \tag{3}$$

where $\boldsymbol{x}$: input speech signal, $\boldsymbol{x}_{zir}$: a zero-input response of the frame which is a ringing from a previous subframe, $\boldsymbol{H}$: a lower triangular matrix of impulse response of the LP synthesis filter, $\boldsymbol{e}_{adp}$: an adaptive/fixed codevector, $\boldsymbol{e}$: a random codevector, $\boldsymbol{W}$: matrix representation of the perceptual weighting filter, g_{adp}: a gain for the adaptive/fixed codebook, and g: a gain for the random codevector.

If a random codevector consists of a stored vector $\boldsymbol{c} = (c_0, c_1, \cdots, c_{N-1})$ and a sign $s(= +1, -1)$ for polarity of the vector, $\boldsymbol{e}$ is represented as

$$\boldsymbol{e} = s\boldsymbol{c}. \tag{4}$$

In CELP coding, residual components after short-term prediction of linear prediction filtering and long-term prediction of pitch filtering are coded by random codevector selected from random codebook. Then the performance of the CELP coder heavily depends on the random codebook structure and methods of the codebook search.

On the other hand, at the bitrate below 4 kbit/s, if the dimension of the excitation vector must be more than 5 or 10 ms, the pitch periodicity still remains in the predicted residuals. So PSI-CELP gives pitch periodicity to the random codebook as well as the adaptive codebook. A subvector of pitch length L is extracted from the stored random vector of dimension N, and the subvector is duplicated and concatenated to the length N [2].

The pitch synchronization method is formulated by the matrices of pitch synchronization matrix $\boldsymbol{P}$, stored random vector $\boldsymbol{c}$ and polarity s as follows.

$$\boldsymbol{e} = s\boldsymbol{P}\boldsymbol{c}, \tag{5}$$

$$\boldsymbol{P} = \left(\begin{array}{cccc|cc} \overleftarrow{1} & \lfloor L \rfloor & \cdots & \overrightarrow{0} & \cdots & 0 \\ 0 & 1 & \cdots & 0 & \cdots & 0 \\ & & \ddots & & & \\ \cdots & \cdots & & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & 1 & \cdots & 0 \\ \hline f_0 & f_1 & \cdots & 0 & \cdots & 0 \\ & & \ddots & & & \\ f_{-1} & f_0 & & 0 & \cdots & 0 \\ & \ddots & \ddots & & & \\ \cdots & & & \cdots & \cdots & \cdots \\ 0 & \cdots & f_{-1} & f_0 & \cdots & 0 \\ \hline \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \end{array}\right). \tag{6}$$

The elements $f_i, (i = \ldots, -1, 0, 1, \ldots)$ in $\boldsymbol{P}$ are interpolation coefficients obtained from the sampling function for non-integer pitch repetition.

The waveform of the prediction residuals has not only pitch periodicity but also power localization synchronized with the phase of the peak location of the pitch waveform. In CELP-type coding, as the dimension of each excitation vector is fixed to a subframe length, there is no relation between the pitch peak position and the start point of the excitation vector.

The phase-adaptive PSI-CELP utilizes the information of the power localization at the pitch peak by synchronizing the start position of the excitation vector with the pitch peak which is obtained from the adaptive codevector by a backward processing in section 4.3..

The dimension of the stored codevector is $N + N_s$, where N is the length of the subframe and N_s is the maximum phase-shift length. In the figure, the start point of the codevector varies among $[-N_s, 0]$ so that the pitch peak of the input speech corresponds to the basic point 0. The pitch peak point can be located at any samples in a subframe, the maximum phase-shift length N_s is $N - 1$. The difference of the start point and the basic point is defined as a phase-shift length ϕ. After the phase-adaptation, a subvector of pitch length L is extracted from the stored vector with phase-shift length ϕ, and the subvector is repeated and concatenated to the random codevector length N in the pitch synchronization part.

The processes of the phase adaptation and the pitch synchronization are represented by using a pitch synchronization matrix $\boldsymbol{P}$, random stored vector $\boldsymbol{c}$, phase-shift matrix $\boldsymbol{F}$, and polarity s as follows.

$$\boldsymbol{e} = s\boldsymbol{P}\boldsymbol{F}\boldsymbol{c}, \tag{7}$$

$$\boldsymbol{F} = \left(\begin{array}{ccc|cccc|ccc} \overleftarrow{0} & \overleftarrow{\phi} & \overrightarrow{0} & \overleftarrow{1} & N & \cdots & \overrightarrow{0} & 0 & \cdots & 0 \\ 0 & \cdots & 0 & 0 & 1 & 0 & 0 & 0 & \cdots & 0 \\ & \cdots & & 0 & \cdots & \cdots & & & \cdots & \\ 0 & \cdots & 0 & 0 & \cdots & 0 & 1 & 0 & \cdots & 0 \end{array}\right), \tag{8}$$

where $\boldsymbol{P} = \boldsymbol{P}(L)$ and $\boldsymbol{F} = \boldsymbol{F}(\phi)$ since $\boldsymbol{P}$ is a function of pitch interval L, and F is a function of the phase-shift length ϕ.

The actual codevector of the codebook 2 is then made in the pitch synchronizer by extracting a fraction of the stored vector. Here, the start point is moved to $-\phi$ and the length is the same as the pitch interval. The extracted fractions are concatenated repeatedly until the total length is N. The random codebook for the phase-adaptive PSI-CELP is generated by codebook training.

4.3. Backward phase adaptation

There are variations which calculate the pitch phase ϕ. We use a method described in [5]. First, impulse sequences of candidate phase ϕ and interval $\lfloor L \rfloor$ are generated. Then, the best pitch phase is defined as the phase which gives the minimum waveform distortion between the waveform excited by adaptive/fixed codevector and that excited by the impulse sequence. This scheme is a backward processing which has advantage of taking no additional transmission bits. Furthermore, the filtering operation of the impulse sequence is small, since the sequence is sparse. This filtering operation is implemented as a matrix operation using impulse response matrix of the synthesis filter [6].

If the backward phase adaptation is well operated to extract the correct pitch phase ϕ, the impulse sequences are matched to the pitch peak locations. However, in the actual processing, the obtained impulse sequences do not match to the pitch peaks due to incorrect extraction of pitch intervals. Since the conventional PSI-CELP coder has the pitch phase information in random excitation codebook, various phase codevectors must be contained. The phase-adaptive PSI-CELP can make effective codebooks based on the backward phase adaptation.

4.4. Training of the excitation codebook

The excitation codebooks are trained by using a generalized Lloyd algorithm. The codebook structure is a two-channel conjugate structure. A distortion measure d and a total distortion D for the codevector k are represented as follows.

$$d = \|\boldsymbol{W}(\boldsymbol{z} - g\boldsymbol{H}\boldsymbol{e}_1)\|^2, \tag{9}$$

$$\boldsymbol{z} = \boldsymbol{x} - \boldsymbol{x}_{zir} - \boldsymbol{x}_{adp} - g\boldsymbol{W}\boldsymbol{H}\boldsymbol{e_2}, \tag{10}$$

$$\tag{11}$$

$$D = \sum_k d_k, \tag{12}$$

where $\boldsymbol{x}_{adp}$ is the synthesized wave excited by codebook 1, g is a gain parameter, $\boldsymbol{e}_1$, $\boldsymbol{e_2}$ are the phase-adaptive and pitch-synchronized excitation vectors.

Training codevectors $\boldsymbol{c}$ in each channel is a stored vector, which is not phase-adapted and not pitch-synchronized. The codevectors in each channel are trained alternatively [7]. The codevector $\boldsymbol{c}$ should meet the condition.

$$\frac{\partial D}{\partial \boldsymbol{c}} = 0. \tag{13}$$

From the equations (7) and (13), the $\boldsymbol{c}$ for phase-adaptive PSI-CELP coding is given by

$$\boldsymbol{c} = \boldsymbol{\Lambda}^{-1} \sum_k (sg\boldsymbol{W}\boldsymbol{H}\boldsymbol{P}\boldsymbol{F})^t \boldsymbol{W}\boldsymbol{z}, \tag{14}$$

$$\boldsymbol{\Lambda} = \sum_k (g\boldsymbol{W}\boldsymbol{H}\boldsymbol{P}\boldsymbol{F})^t (g\boldsymbol{W}\boldsymbol{H}\boldsymbol{P}\boldsymbol{F}). \tag{15}$$

5. GAIN PARAMETER ENCODING

In the previous research[2, 3, 4], two-stage quantization method consisting of a power quantization and a gain quantization was used. The gain quantizer quantizes the residual power obtained from a total power and a prediction gain of an LPC filter.

Since gain parameters have high correlation between adjacent frames, a quantization method by using a backward prediction obtained higher quantization performance. This backward prediction method has been implemented in ITU-T G.729. This method is also expected to be robust against channel errors when no protection bits are available. In this research, a gain quantizer with MA-type backward prediction in logarithmic domain.

The gain $g_{adp}^{(m)}$ for the adaptive/fixed codevector $\boldsymbol{e}_{adp}^{(m)}$ at time m is represented as follows.

$$g_{adp}^{(m)} = \gamma_{adp}^{(m)} g'_{adp}, \tag{16}$$

$$\begin{aligned} 20\log g_{adp}^{(m)} &= 20\log \gamma_{adp}^{(m)} \\ &+ 10\log(\|g_{adp}^{(m-1)}\boldsymbol{e}_{adp}^{(m-1)} + g^{(m-1)}\boldsymbol{e}^{(m-1)}\|^2/N) \\ &- 10\log(\|\boldsymbol{e}_{adp}^{(m)}\|^2/N), \end{aligned} \tag{17}$$

where g'_{adp} is the predicted gain from previous subframe excitation, and $\gamma_{adp}^{(m)}$ is the fine scaling factor which is quantized.

The gain $g^{(m)}$ for RCB excitation $g^{(m)}$ at time m, is represented as follows.

$$g^{(m)} = \gamma^{(m)} g', \tag{18}$$

$$\begin{aligned} 20\log g^{(m)} &= 20\log \gamma^{(m)} + \sum_{i=1}^{Q} b_i U^{(m-i)} + \overline{E} \\ &- 10\log(\|\boldsymbol{e}^{(m)}\|^2/N), \end{aligned} \tag{19}$$

$$U^{(m)} = 20\log \gamma^{(m)}. \tag{20}$$

where g' is the predicted gain from preceding subframe excitations, $\gamma^{(m)}$ is the fine scaling factor which is quantized, and $\overline{E}$ is an offset value removed in the prediction process.

The best combination of $(\gamma_{adp}^{(m)}, \gamma^{(m)})$ in the gain codebook is selected to minimize the equation(3).

6. SUBJECTIVE QUALITY EVALUATION

A subjective quality evaluation was performed for two 4-kbit/s coders, which have different gain quantization schemes. One 4-kbit/s coder called method 1 is previously presented version in [4], which quantizes the gain parameters in the form of a power factor and fine scaling factors without the backward gain prediction. Another 4-kbit/coder called method 2 has the quantization scheme described in the section 5..

Figure 2 shows the results of mean opinion scores (MOS) in the absolute category rating (ACR) experiment. The subjects are 16 Japanese non-experts in speech research. A

set of eight modified-IRS filtered speech of Japanese sentence pairs is tested. First it is confirmed that the quality of method 2 coder, which has the backward gain prediction scheme, is better than that of method 1 coder.

The MOS results of the method 2 coder with backward gain prediction are comparable to those of ITU-T G.726 32-kbit/s under the conditions of normal and low input levels, and tandem connection.

Figure 3 shows the results of an ACR test for noisy environment. The original is the speech with background noise. In the ACR test for noisy environment, the quality performances of all the coders of G.726, method 1, and 2 are not significantly different to that of the original.

7. CONCLUSION

We have designed a 4-kbit/s phase-adaptive PSI-CELP coder for ITU-T standard. The main features of the proposed coder are pitch synchronous and phase-adaptive excitation structures. MOS experiments for both clean and noisy environments have been performed for 4-kbit/s coders with/without the backward gain prediction. It has been confirmed that the quality of the 4-kbit/s coder with the backward gain prediction is better than that of the coder without the prediction. It has been also confirmed that the MOS results of the proposed coder with backward gain prediction are comparable to those of ITU-T G.726 32-kbit/s under the conditions of normal and low input levels, and tandem connection. The MOS results under noisy environments show that the quality performances of all the coders of G.726, and proposed coders are not significantly different to that of the original.

ACKNOWLEDGMENTS

The author would like to thank Dr. Nobuhiko Kitawaki and Takao Kaneko for their guidance in this research. The author also thank the members of the speech coding group for their valuable discussions.

REFERENCES

[1] M. R. Schroeder and B. S. Atal: "Code-excited linear prediction (CELP): high-quality speech at very low bit rates," *Proc. IEEE ICASSP-85,* 25.1, pp.937–940, (1985).

[2] S. Miki, K. Mano, T. Moriya, K. Oguchi, and H. Ohmuro: " A pitch synchronous innovation CELP (PSI-CELP) coder for 2–4kbit/s," *Proc. IEEE ICASSP-94*, pp.II-113–II-116 (1994).

[3] K. Mano, T. Moriya, S. Miki, H. Ohmuro, K. Ikeda, and K. Ikedo: "Design of a Pitch Synchronous Innovation CELP coder for mobile communications," *IEEE J. on Select. Areas Commun.,* Vol. 13, No. 1, pp.31-41 (1995).

[4] K. Mano, and T. Moriya: " Improved 4-kbit/s PSI-CELP coding using pitch and phase adaptation," *Proc. IEEE Workshop on speech coding for telecommunications*, pp.41-42 (1995).

[5] Yamaura: "Improving the quality of CELP coder at 2.4kbps," *Proc. Fall Meeting of Acoust. Soc. of Japan,* pp.269-270, (in Japanese)(1992).

[6] C. Laflamme, J-P. Adoul, H. Y. Su, and S. Morissette: "On reducing computational complexity of codebook search in CELP coder through the use of algebraic codes," *Proc. IEEE ICASSP-90,* pp. 177-180 (1990).

[7] T. Moriya: "Two-channel conjugate vector quantizer for noisy channel speech coding," *IEEE J. on Select. Areas Commun.,* Vol. 10, No. 5, pp.866-874, (1992).

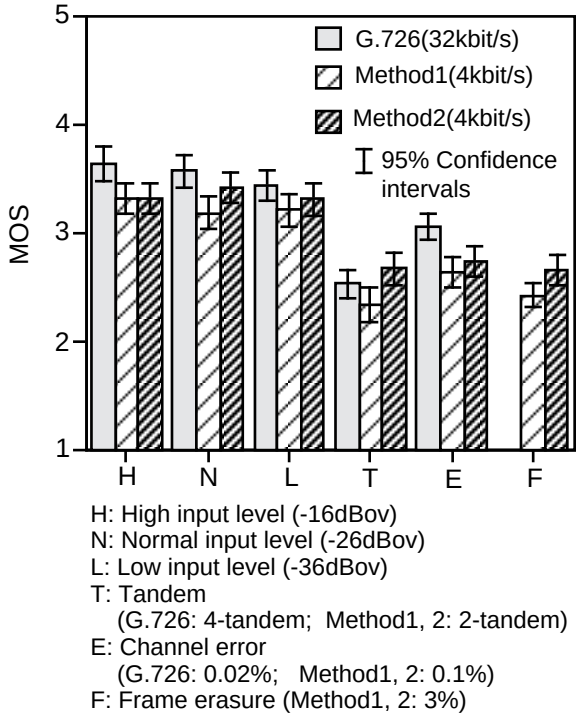

Figure 2. Result of MOS experiment

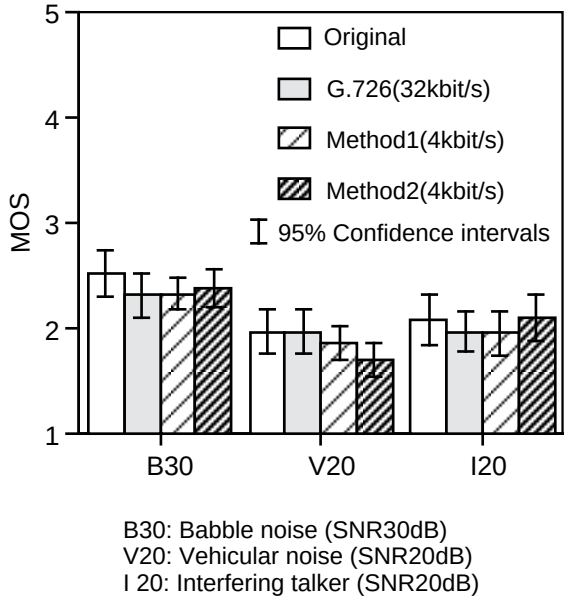

Figure 3. Result of MOS experiment in background noise condition