A New Structural Preprocessor for Low-Bit Rate Speech Coding

Joon-Hyuk Chang[1], *Jong-Won Shin*[2], *Seung Yeol Lee*[2] *and Nam Soo Kim*[2]

[1]Department of Electrical and Computer Engineering
University of California, Santa Barbara,
Santa Barbara, 93117, USA
[2]School of Electrical Engineering and INMC
Seoul National University Seoul, Korea.
Kwanak P.O.Box 34, Seoul 151-742, Korea,
jhchang@ece.ucsb.edu, {jwshin, sylee}@hi.snu.ac.kr, nkim@snu.ac.kr

Abstract

In this paper, we apply a new structural approach to generalized analysis-by-synthesis (GAbS) for system identification as a preprocessor of a low-bit-rate speech coder. In our approach, the coder-decoder (CODEC) system is separately estimated and then applied to modify the current input signal. This is different from that originally proposed where the CODEC system is sequentially estimated and then applied to the next input signal. The proposed estimation scheme is compared to the conventional method in terms of the signal modification approach under the various noise data and in several SNR conditions, and shows better performance.

1. Introduction

Low-bit-rate speech coding technologies have been an important part in today's mobile communication systems [1]. In a low-bit-rate speech coder, the speech signal is represented by vectors of parameters which is usually vector quantized. An effective method for quantization is to employ the parameter value to resynthesize the original signal and to select the quantized value which results in the most accurate reconstruction. This procedure is known both as *closed-loop* quantization and *analysis-by-synthesis* (AbS). Almost all the low-bit-rate speech coders that are in use today are based on the AbS [2], [3]. Recently, Kim and Chang introduced a preprocessor that modifies the signal applied to a low-bit-rate speech coders as an input [4], [5]. Specifically, a criterion which compromises the quantization error with the distortion incurred due to the modification was introduced. Also, the coder-decoder (CODEC) characteristics were described in terms of a transfer matrix, and it was estimated according to the recursive least squares (RLS) criterion for system identification.

In this paper, we propose a new structural approach to generalized AbS (GAbS) for the robust system identification. Based on the generalized AbS mechanism, a *try*-and-*catch* (TC) structure, which is referred to as TC-GAbS, is designed for the estimating transfer matrix of the CODEC system. A major advantage of the proposed structure is that more accurate estimation of the CODEC transfer matrix is obtained by the *try* stage compared to the sequential RLS estimation since we can see that how the CODEC system responds to input signal ahead of the actual transmission. It has been shown that a more precise estimation of the CODEC characteristics is much more helpful in enhancing the subjective speech quality.

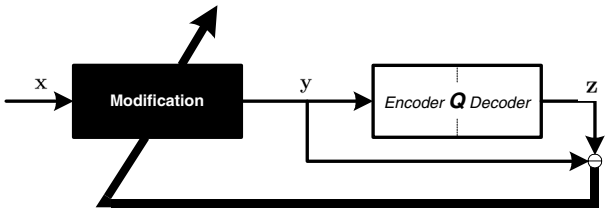

Figure 1: Overall structure of the originally proposed approach (RLS-GAbS)

2. Review of Signal Modification

We briefly review the notion of the signal modification based on GAbS which was proposed in [4], [5]. The basic structure of the corresponding overall system is depicted in Fig. 1. As shown, in the previous approach, prior to applying to the encoder, we modify $\mathbf{x}$ such that the modified vector can better fit to the speech coder. Let $\mathbf{y} = [y(0), y(1), \cdots, y(M-1)]^T$ be the signal samples obtained by modifying $\mathbf{x} = [x(0), x(1), \cdots, x(M-1)]^T$ and $\mathbf{z} = [z(0), z(1), \cdots, z(M-1)]^T$ be the output vector which is produced when $\mathbf{y}$ is applied to the coder and then re-synthesized in the decoder. Also let $\mathbf{X} = [X(0), X(1), \cdots, = X(N-1)]^T$, $\mathbf{Y} = [Y(0), Y(1), \cdots, Y(N-1)]^T$ and $\mathbf{Z} = [Z(0), Z(1), \cdots, Z(N-1)]^T$ be the transform domain representation of $\mathbf{x}$, $\mathbf{y}$ and $\mathbf{z}$, respectively.

In order to develop a mathematically tractable approach, we assume that the CODEC system with the input and output pair is approximated by the linear system model in terms of $\mathbf{Q}$ which represents transfer matrix as follows:

$$\mathbf{Z} = \mathbf{Q}\mathbf{Y} \tag{1}$$

where $\mathbf{Q}$ is assumed to be a diagonal matrix. It gives us an efficient way to derive an optimal solution, and it works quite well in the signal modification scheme. Given $\mathbf{Q}$, modification of the input vector, $\mathbf{X}$ is achieved according to the following criterion:

$$\hat{\mathbf{Y}} = \arg\min_{\mathbf{Y}} J(\mathbf{Y}) \tag{2}$$

where $\hat{\mathbf{Y}}$ denotes the desired modified vector. Here, the objective function $J(\mathbf{Y})$ is given by

$$J(\mathbf{Y}) = ||\mathbf{X}-\mathbf{Y}||_W^2 + K||\mathbf{Y}-\mathbf{Q}\mathbf{Y}||_W^2 \tag{3}$$

in which

$$||\mathbf{a}||_W^2 = \mathbf{a}^H W \mathbf{a} \tag{4}$$

with W representing a perceptual weighting filter; H means the Hermitian operation; and K is a positive constant for the signal modification. The definition of the proposed objective function, $J(\mathbf{Y})$ is motivated by the following assumptions which are considered reasonable: If the input signal is extracted from a pure speech sound, the difference between the CODEC input and output will be small and very little modification is desired. On the other hand, if the input signal deviates much from the pure speech characteristics, the quantization error will become larger and it would be better modify the input signal such that the CODEC quantization error could be reduced [5].

Since $J(\mathbf{Y})$ is a quadratic function of $\mathbf{Y}$, differentiating it with respect to $\mathbf{Y}$ and then equating to zero leads us to

$$\begin{aligned}\frac{\partial J(\mathbf{Y})}{\partial \mathbf{Y}} &= -2\mathbf{W}(\mathbf{X}-\mathbf{Y}) + 2K\tilde{\mathbf{Q}}^H\mathbf{W}\tilde{\mathbf{Q}}\mathbf{Y} \\ &= 0\end{aligned} \tag{5}$$

with $\tilde{\mathbf{Q}}$ being $\mathbf{I}-\mathbf{Q}$; $\mathbf{I}$ means the $M \times M$ identity matrix. From (5), it is easy to show that

$$\hat{\mathbf{Y}} = \left[\mathbf{W} + K\tilde{\mathbf{Q}}^H\mathbf{W}\tilde{\mathbf{Q}}\right]^{-1}\mathbf{W}\mathbf{X}. \tag{6}$$

The system transfer matrix $\mathbf{Q}$ is an important factor in the signal modification and it was estimated on-line by means of the generalized AbS using the recursive least squares (RLS) in [4], [5]. Assume that we apply T input vectors, $\mathbf{Y}_1$, $\mathbf{Y}_2$, $\cdots$, $\mathbf{Y}_T$ to the speech CODEC, and obtain the corresponding output vectors, $\mathbf{Z}_1, \mathbf{Z}_2, \cdots, \mathbf{Z}_T$. According to the least square estimation (LSE) criterion, $\hat{\mathbf{Q}}$ is obtained as follows:

$$\begin{aligned}\hat{\mathbf{Q}} &= \arg\min_{\mathbf{Q}} \sum_{t=1}^{T} w_t ||\mathbf{Z}_t - \mathbf{Q}\mathbf{Y}_t||^2 \\ &= \arg\min_{\mathbf{Q}} \sum_{t=1}^{T} w_t \left(\sum_{k=0}^{N-1} |Z_t(k) - Q(k)Y_t(k)|^2 \right)\end{aligned} \tag{7}$$

where $\hat{\mathbf{Q}}$ is the least squares estimate for $\mathbf{Q}$ and w_t indicates the weighting factor assigned to tth frame input-output pairs, $(\mathbf{Y}_t, \mathbf{Z}_t)$[1].

The function $E(\mathbf{Q})$ denoting error term is decomposed into N separate functions $E_1(Q(1))$, $E_2(Q(2))$, $\cdots$, $E_{N-1}(Q(N-1))$ where

$$E_k(Q(k)) = \sum_{t=1}^{T} w_t |Z_t(k) - Q(k)Y_t(k)|^2. \tag{8}$$

Differentiating $E(\mathbf{Q})$ with respect to $\mathbf{Q}$, we obtain

$$\begin{aligned}\frac{\partial E(\mathbf{Q})}{\partial Q(k)} &= \frac{\partial E_k(Q(k))}{\partial Q(k)} \\ &= -\sum_{t=1}^{T} w_t \left(Z_t^*(k) - Y_t^*(k)Q^*(k)\right) Y_t(k)\end{aligned} \tag{9}$$

for $k = 0, 1, \cdots, N-1$ and equating each differential to zero results in

$$\hat{Q}(k) = \frac{\sum_{t=1}^{T} w_t Z_t(k) Y_t^*(k)}{\sum_{t=1}^{T} w_t |Y_t(k)|^2}. \tag{10}$$

In our previous work [4], [5], estimation of $\mathbf{Q}$ was implemented using the RLS based GAbS method. See [4], [5] for the detailed review of the RLS for $\mathbf{Q}$.

3. New Structure with Generalized Analysis-by-Synthesis

As we mentioned in the previous section, the transfer matrix, $\mathbf{Q}$ is estimated prior to input modification by the sequential manner with an appropriate forgetting factor in [4], [5], which is based on the assumption that the transfer characteristics of a CODEC evolve slowly. Since, however, the CODEC input-output characteristics are usually time varying depending on the given speech frame, it has been frequently observed that the RLS based GAbS does not work well for the rapidly evolving speech frames. For that reason, we propose a new structural approach for system identification in order to take into account the time-varying characteristics of the CODEC system and to improve the performance of the estimation of $\mathbf{Q}$. Since the actual CODEC input-output characteristics are unknown until the current input signal is applied into the CODEC, all that we propose is to analyze the response of the real

[1] w_t is not employed in this work

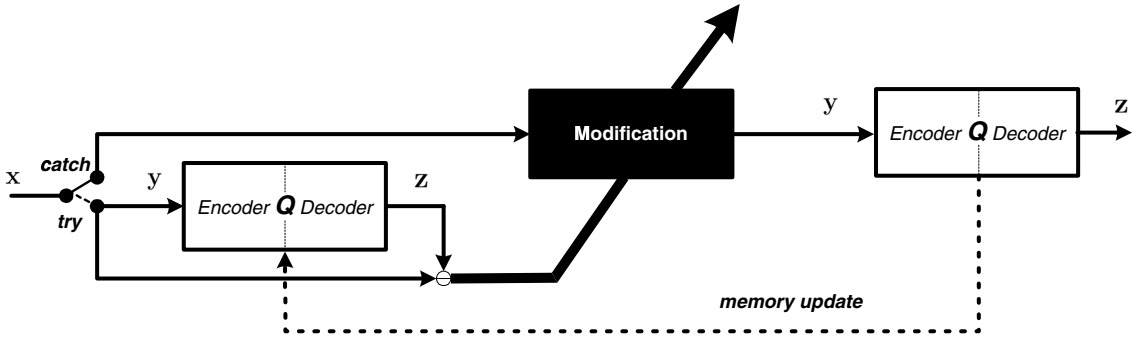

Figure 2: Proposed structure for the generalized AbS (TC-GAbS)

CODEC system for the current signal rather than the previous signal. For this reason, we adopt the TC-GAbS, in which CODEC transfer matrix is estimated on a part of *try* and then the *catch* process modifies the input signal using the estimated $\mathbf{Q}$ of the *try* part. Here, it should be noted that all memory of the catch process after the modification process is copied to the *try* process for the next frame. Memory in the CODEC refers to all the static variables and arrays that must be kept between each call of the main encoder or decoder procedure, and can not be used for other purposes until the coder is switched off.

The proposed structure is given in Fig. 2. Based on the figure, we summarize the TC-GAbS scheme in terms of the several steps for signal modification as follows:

- *Try* Routine
  1. $\mathbf{X}_t$ is directly assigned to $\mathbf{Y}_t$.
  2. Update memory from *Catch* process in the previous $t-1$ frame
  3. Obtain the CODEC output $\mathbf{Z}_t$ by applying the modified input vector $\mathbf{Y}_t$
  4. Estimate $\hat{\mathbf{Q}}_t$ based on (6) by using $\mathbf{Y}_t$ and observed $\mathbf{Z}_t$ as follows: $\hat{Q}_t(k) = \frac{w_t Z_t(k) Y_t^*(k)}{w_t |Y_t(k)|^2}$.
- *Catch* Routine
  1. Obtain $\mathbf{Y}_t$ by modifying $\mathbf{X}_t$ based on the estimated $\mathbf{Q}_t$ according to (6)
  2. Speech coder compresses $\mathbf{Y}_t$ and transmits it.
  3. Quantized parameters also pass through the decoder at *Catch* routine[2].

While the superiority of the proposed approach is expected, their computational complexity comparing to that of conventional method should be assessed for a fair comparison. The original and proposed system were implemented in real-time using the PC with Intel Inc. Pentium IV 2.8 GHz. Based on this hardware, we used the profiler in the Visual C++ of the Microsoft Inc. to calculate the overall computation times, where we could check the separate and overall computation time for each function of the reference C code (floating point). For the overall system, it is discovered that the computational time of the proposed TC-GAbS scheme from the processed speech, on average, was 39 % larger than that of the RLS-GAbS.

[2]It is necessary for updating the memory of the *Try* Routine.

Table 1: PESQ for various noises and SNR (dB); no modification (G.729A), RLS-GAbS, and TC-GAbS

noise	SNR	G.729A	RLS-GAbS	TC-GAbS
clean	∞	3.25	3.25	3.25
white	0	1.48	1.58	1.61
	5	1.74	1.84	1.88
	10	2.15	2.24	2.27
	15	2.45	2.53	2.55
	20	2.78	2.85	2.87
babble	0	1.80	1.95	1.99
	5	2.14	2.31	2.36
	10	2.56	2.72	2.75
	15	2.81	2.93	2.96
	20	3.08	3.15	3.18
office	0	1.77	1.91	1.95
	5	2.09	2.26	2.30
	10	2.54	2.69	2.74
	15	2.78	2.90	2.94
	20	3.04	3.12	3.15

4. Experimental Results

We carried out a number of experiments on the conventional RLS-GAbS and presented TC-GAbS technique. As a target speech coder, the G. 729 CS-ACELP was employed [6]. For the speech modification, each frame of data was transformed into a vector consisting of the corresponding discrete Fourier transform (DFT) coefficients, and the modification was done in the DFT domain. In all cases, the speech modification was conducted with K = 3 and an identity matrix was employed for the weight-

Table 2: MOS Test Results For the Signal Modification; No Modification (G.729A), RLS-GAbS, and TC-GAbS and (with 95 % Confidence Interval)

noise	SNR	G.729A	RLS-GAbS	TC-GAbS
clean	∞	4.30 ± 0.03	4.30 ± 0.03	4.31 ± 0.03
white	0	1.37 ± 0.05	1.60 ± 0.05	1.66 ± 0.05
	5	1.66 ± 0.07	1.89 ± 0.07	1.95 ± 0.08
	10	1.91 ± 0.08	2.17 ± 0.09	2.26 ± 0.09
	15	2.10 ± 0.10	2.48 ± 0.11	2.53 ± 0.12
	20	2.33 ± 0.14	2.76 ± 0.14	2.85 ± 0.14
babble	0	1.85 ± 0.06	2.20 ± 0.06	2.30 ± 0.06
	5	2.20 ± 0.08	2.58 ± 0.08	2.66 ± 0.08
	10	2.49 ± 0.09	2.82 ± 0.10	2.86 ± 0.09
	15	2.79 ± 0.11	3.10 ± 0.11	3.18 ± 0.12
	20	2.98 ± 0.15	3.31 ± 0.15	3.43 ± 0.15
office	0	1.80 ± 0.06	2.15 ± 0.05	2.25 ± 0.06
	5	2.17 ± 0.07	2.55 ± 0.07	2.65 ± 0.07
	10	2.46 ± 0.08	2.81 ± 0.07	2.88 ± 0.07
	15	2.73 ± 0.09	3.07 ± 0.10	3.17 ± 0.10
	20	2.98 ± 0.13	3.28 ± 0.13	3.40 ± 0.13

ing matrix $\mathbf{W}$ in (3). 24 sentences spoken by three male and three female speakers were used for the experiment data. The speech modification was done for each frame of 10 ms with the sampling rate of 8 kHz. Three type of noise sources: the white and the babble noises from the NOISEX-92 database and the office noise from the Dynastat database, were added to the clean speech waveform at SNRs of 0, 5, 10, 15 and 20 dB [7]. While the Gaussian noise is completely stationary, the babble noise is nonstationary and the office noise is highly nonstationary.

For the purpose of evaluating the performance of the presented method, we first measured a perceptual evaluation of speech quality (PESQ) produced by the ITU-T P.862 tests. The PESQ results are given in Table I. Table I illustrates that the TC-GAbS approach outperforms the RLS-GAbS methods under the given noise conditions, where we can see that incorporation of the TC-GAbS has definitely a positive effect in terms of the PESQ scores. The performance of the TC-GAbS approach is the same to that of the RLS-GAbS at the clean condition.

Secondly, for the purpose of evaluating the subjective quality of the presented TC-GAbS scheme, we carried out a set of informal listening tests. Subjective opinion scores were decided by a group of 12 listeners and each listener gave for each test sentence a score such that 5 (Excellent), 4 (Good), 3 (Fair), 2 (Poor), 1 (Bad). The scores were then averaged to yield the mean opinion score (MOS) results [8]. The MOS results are shown in Table II. From Table II, the results obtained with the employment of the TC-GAbS scheme were consistently better than that of the original RLS-GAbS. Performance improvement was found greater for the babble and office noise environments compared to the white noise case. The improvement using the PESQ score is somewhat small but, as MOS test confirm again, significantly less quantization noise is audible.

5. Conclusions

We have presented an approach which incorporates a new structure for system identification with GAbS, which is called the TC-GAbS. With a higher computational burden, the TC-GAbS gives us more precise modeling of the CODEC characteristics and improved speech quality. Further improvement is expected if we combine the comparing methods with an appropriate trade-off between the computational load and an accurate modeling of the CODEC system.

6. References

[1] W. B. Kleijn and K. K. Paliwal, *Speech Coding and Synthesis*. Elsevier Science B. V., 1995.

[2] W. B. Kleijn, R. P. Ramachandran and P. Kroon, "Interpolation of the pitch-predictor parameters in analysis-by-synthesis coders," *IEEE Speech, Audio Processing*, vol. 2, no. 1, pp. 42-54, Jan. 1994.

[3] B. Atal and M. Schroeder, "Predictive coding of speech signals and subjective error criteria," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-27, no. 3, pp. 247-254, March 1979.

[4] N. S. Kim and J. -H. Chang, "A preprocessor for low-bit-rate speech coding," *IEEE Signal Processing Letters*, vol. 9, no. 10, pp. 318-321, Oct. 2002.

[5] N. S. Kim and J. -H. Chang, "Signal modification for robust speech coding," *IEEE Trans. Speech and Audio Processing*, vol. 12, no. 1, pp. 9-18, Jan. 2004.

[6] R. Salami et al., "Design and description of CS-ACELP: a toll quality 8 kb/s speech coder," *IEEE Speech, Audio Processing*, vol. 6, no. 2, pp. 116-130, Mar. 1998.

[7] A. Varga and H. J. M. Steeneken, "Assessment for automatic speech recognition: II. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems," *Speech Communication*, vol 12, no. 3, pp. 247-251, July 1993.

[8] ITU-T, "Subjective quality test plan for the ITU-T 4-kbit/s speech coding algorithm," July 1999.

[9] ITU-T P.800, "Methods for subjective determination of transmission quality," Aug. 1996.