A Low Rate VQ Speech Coding Algorithm with Variable Transmission Frame Length

가변 전송 Frame 길이를 갖는 저 전송속도 VQ 음성부호화 알고리즘에 대한 연구

Jeong Woo Jwa*, Sung Ro Lee*, Hwang Soo Lee*
좌 정 우*, 이 성 로*, 이 황 수*

ABSTRACT

In this paper, an efficient variable transmission frame length(VTFL) speech coding method is proposed and its performance is studied by computer simulation. The proposed speech coding method is based on the idea that the performance of speech coding would be improved by varying the transmission frame length according to the stationarity of input speech signal and vector quantizing the representative feature vector of the transmission frame. In the proposed speech coding method, the feature vector sequence consists of PARCOR coefficient vectors obtained by analyzing input speech signal sample-by-sample using the prewindowed recursive least square lattice algorithm. We take the segmentation from the input speech signal and then determine the representative PARCOR vector in each segment. By joining the consecutive segments of phonetically similar characteristics using the likelihood ratio distortion measure, we can obtain transmission frames of variable length. From the computer simulation of total transmission bit rate while maintaining the good reproduced speech quality.

요 약

본 논문에서는 저 전송속도의 음성 부호화기를 제안하였고 컴퓨터 시뮬레이션을 통하여 성능분석과 유연성을 입증하였다. 제안된 부호화 방식은 입력 음성신호의 Stationarity에 따라 전송 프레임의 길이를 가변하고, 전송 프레임의 대표적인 특징 벡터를 Vector Quatization으로 부호화하였다. 제안된 부호화 방식에서 특징 벡터열은 입력 음성신호를 샘플단위로 Prewindowed RLS Lattice 알고리즘을 통해 구한 PARCOR 계수로 구성된다. 입력 음성신호는 Subsegment로 분할되고, 각 Subsegment에서 대표적인 PARCOR 계수를 구한다. Likelihood Ratio Distortion Measure를 사용하여 유사도에 따라 Subsegment를 병합함으로써 전송프레임을 결정한다.

컴퓨터 시뮬레이션 결과로부터 제안된 VTEL 음성 부호화 방식은 좋은 음질을 유지하면서 전체 전송속도를 크게 줄일 수 있다.

I. INTRODUCTION

For efficient digital voice communication, many researches have been performed for decades to lower the transmission bit rate of speech signal while maintaining good reproduced speech quality. Among those researches on speech coding, variable-frame-rate coding is one of the promising coding techniques especially for speech storage applications because it exploits the nonstationari-

* 한국과학기술원 정보 및 통신공학과
Department of Information and Communication Engineering, KAIST
접수일자 : 1992. 12. 12.

ty property of speech signal and can reduce the coding bit rate significantly. Up to now, several variable-frame-rate coding methods are reported by Turner and Dickinsion[1], Viswanathan et al. [2], Papamichalis[3], Kuang and Chan[4], Fukabayashi and Chuang[5], and so on. Turner and Dickinsion proposed a variable frame length coder. In the coder, input speech signal is analyzed using the covariance lattice method. Transmission frame lengths are determined by joining consecutive frames by using the RMS log spectral difference measure with a certain threshold. Since the fast changing transition region and the steady phone region are encoded with different rates, this scheme can reduce total transmission bit rate. But the performance of this method is limited by the fixed frame length in transition and voiced regions. Fukabayashi and Chuang presented a variable rate coding method which analyzes input speech using a point-wise analysis algorithm. The input speech signal is segmented by summing the sum-of-squares of the difference between the preceding and the present lattice filter coefficients and comparing it to a pre-determined threshold value. They determined the representative feature vector by averaging the locally converged cluster of feature vectors in a segment. This method is very time-consuming in computing the sum-of-squares difference function at every sampling instant. Besides, the linear smoothing method to obtain the average of the locally converged cluster of feature vectors in a segment may degrade the reproduced speech quality.

To circumvent these problems, in this work, we first search pitch periods and make voiced/unvoiced/silence(V/UV/S) decisions from the input speech. Then the voiced region of speech signal is segmented into pitch period units. And the unvoiced region of speech signal is segmented into small sized(20-samples) intervals. Using this segmentation method, the characteristics of the transition region can be described more precisely. After the segmentation of input speech signal, we analyze the input speech using the prewindowed recursive least square(PRLS) lattice algorithm and select a PARCOR coefficient vector at a properly chosen sample point in the speech segment as a representative feature vector. By joining the consecutive speech segments of similar characteristics using the likelihood ratio(LR) distortion measure, we can determine transmission frames of variable length. In a transmission frame of speech signal, we select the PARCOR coefficient vector of the first speech segment as a representative model vector of the whole transmission frame. The model vector of the transmission frame is then vector quantized (VQ) for transmission. The proposed algorithm has moderate computational complexity and yields good reproduced speech quality. Following this introduction, we describe the proposed variable transmission frame length (VTFL) low rate speech coder in Section II, In Section III, we discuss the computer simulation results for evaluating the performance of the proposed speech coder. Finally, conclusions are given in Section IV.

II. THE PROPOSED VTEL LOW RATE SPEECH CODER

Fig. 1 shows the block diagram of the proposed VTFL speech coder. First, the speech signal is preemphasized and the V/UV/S decision is made. Then in the voiced region, pitch periods are extracted[6]. With the V/UV/S and pitch period information, the procedure for segmenting of speech signal is as follows. In a voiced region

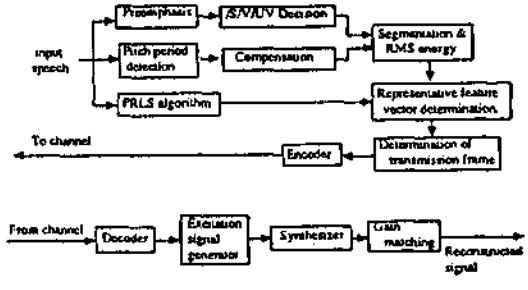

Fig.1. Block diagram of the proposed VTFL speech coder

of speech, we segment speech signal into the same length as the pitch period interval. By applying the PRLS algorithm to the voiced speech segment. This abrupt changes are due to the excitation of the speech signal. After the excitation signal location, however, the PARCOR coef ficient vectors converge to a certain value. The fluctuations around the converged value are rather small. In unvoiced regions, a 20-sample interval that is small compared with the pitch period interval is taken as a segment length because of the strong nonstationarity of the unvoiced sound. In that region, the PARCOR vectors change very rapidly.

We analyze the input speech signal using the PRLS algorithm and a PARCOR coefficient vec tor is extracted at each sample point by using the PRLS lattice algorithm. This point-wise analysis method is more suitable for tracking fast changing speech sound parameters than other block-wise analysis methods, thus results in the

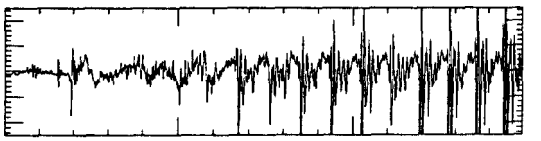

Fig.2. Speech waveform at the beginning portion of the sound /dal/

better output speech quality. Fig.2 shows begining part of speech waveform pronounced as /dal/ and Fig.3 shows segmentation in a voiced region. In the figure, the length of each segment interval is the same as the corresponding pitch period determined by the pitch period detector. For verifying the validity of the fore-mentioned

Fig.3. Segmentation results in the voiced region

speech segmentation method, we observe the variation of the first PARCOR coefficient, i.e., K_1, as shown in Fig.4. In the figure, one can notice that abrupt changes of the K_1 value occur at the speech segment boundaries as is expected.

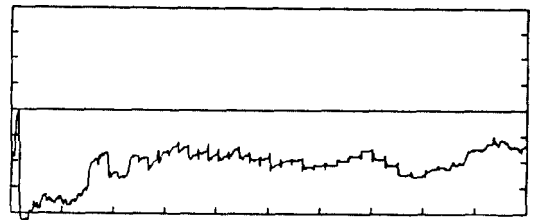

Fig.4. Variation of the 1st PARCOR coefficient K_1

A. Determination of a Representative Feature Vector in a Speech Segment

In order to improve the reproduced speech quality, it is very important how to determine the representative feature vector in a speech segment. In Fig.4, we observe the following facts. The PARCOR vectors in the unvoiced region change in a random fashion but not too much. In the voiced region, except for the region near the segment starting point where the excitation signal is located, the PARCOR coefficients change little over a pitch period. Also we notice that length of the transition region is short and abrupt changes of the feature vectors occur in the region. The transition region is usually included in the unvoiced region.

To accurately represent the speech signal in the transition regions or unvoiced regions, we choose shorter speech segments in these regions than those in voiced regions. Because the variation of the PARCOR vectors in this short segment of the 10th sample point of this speech segment as a representative feature vector for that sample point from the starting point of each speech segment is chosen as the representative feature vector of the segment because the PRLS algorithm converges sufficiently at about the 30th sample point (see Fig.4). As mentioned in section I, choosing the representative feature

vector in this way will yield the better synthesized speech quality than the feature vector obtained by the linear smoothing method which averages the PARCOR vectors in the speech segment.

B. Determination of Transmission Frame Length

The VTFL speech coding algorithm is a very efficient data compression method which can improve the reproduced speech quality. Since the correlations between consecutive speech segments in stable(stationary) speech signal regions are very high, the speech segments with phonetically similar characteristics can be joined to make a transmissioon frame by using the LR distortion measure. Fig.5 shows this joining procedure using

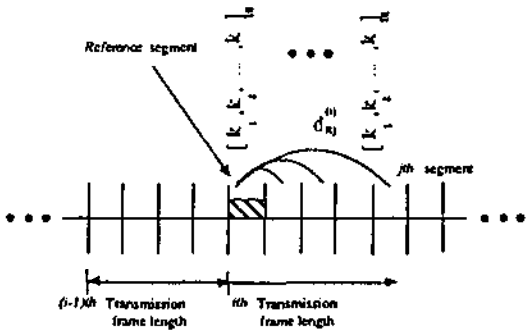

Fig.5. Determination of the transmission frame length

the LR distortion measure. We will call the first segment of the ith transmission frame as the reference segment. In order to determine the length of the ith transmission frame, the distortion between the reference segment and the jth segment, $d^{(i)}_{Rj}$, is computed as follows :

$$d^{(i)}_{Rj} = \frac{a_j^t R_{ref} a_j}{a_{ref}^t R_{ref} a_{ref}} - 1$$

where, a_{ref}, a_j are representative linear prediction coefficient(LPC) vectors in the reference segment and the jth segment, respectively, and R_{ref} is an autocorrelation matrix in the reference segment. If $d^{(i)}_{Rj}$ is less than a pre-determined threshold, d_{TH}, then the jth segment is included in the ith transmission frame and the above procedure is repeated for the (j+1)th segment. Otherwise, the iteration to determine the ith transmission frame is terminated. Then the resulting ith transmission frame consists of the segments from the reference to the (j−1)th segment. Next, the procedure for determining the (i+1)th transmission frame begins by taking the jth segment as the reference segment of the new transmission frame.

A representative feature vector for the transmission frame is chosen as the representative PARCOR vector of reference segment which is the first segment in that frame. When we determine transmission frames with this method, lengths of transmission frames will be varied. Since the amount of information of speech signal is larger in the abruptly changing transition region than in the steady-state region of a phoneme, we can obtain better performance if we make longer transmission frame lengths in the steady-state region of speech and shorter transmission frame lengths in the transition regions of speech signal. Since the average length of the voiced region is usually much longer than that of the unvoiced region and the average length of the transition region is very short, we can reduce the total bit rate while maintaining the output speech quality.

C. VQ Encoding

Here we consider encoding of the representative feature vectors in transmission frames using VQ[7][8]. In this work, we use the modified K-means(MKM) clustering algorithm[9] with splitting to design separate VQ codebook for voiced signal and unvoiced signal. Input training data set for VQ training consists of 100 words and 15 sentences spoken by three male speakers. The VQ input training vectors are obtained as follows :

i)The PARCOR coefficients are extracted sample by sample from the input training speech signal by using the PRLS lattice algorithm.

ii)The V/UV/S decision and the pitch period detection are performed.

iii)In an unvoiced region, the VQ input vectors are extracted every 10 sample-points.

iv)In a voiced region, the VQ input vectors are extracted every 10 sample-points with in the steady-state region of each speech segment, that is, from (segment starting point+15) to (segment ending point 10).

The size of the VQ codebook for unvoiced signal is 512 and that for voiced signal is 1024. The VQ encoding of the representative feature vector in a transmission frame is selecting the index of a codeword minimizing the LR distortion. The other parameters are encoded similar to the standard LPC-10. The encoded parameters are then transmitted to the receiver. At the receiver, the speech signal is reproduced by using the gain normalized synthesis method.

III. COMPUTER SIMULATION RESULTS

Now that the framework of proposed VTFL speech coding method has been established, computer simulation is done to obtain the performance of the proposed speech coding method.

First, the performance of the proposed VTFL speech coding algorithm is compared with that of the conventional LPC coding algorithm, which uses a block analysis method while the former uses a point-wise analysis method. Second, to investigate the effect of selecting a representative feature vector in a speech segment, we compare the performance of the proposed speech coding algorithm with that using averaged PARCOR vectors obtained by linear smoothing. Third, computer simulation is done to examine performance variations by varying the sample point at which we extract the representative PARCOR vector for the speech segment. For the computer simulation study, input speech is extracted from three male speakers. The speech signal is then bandlimited to 4.5kHz by a low pass filter and sampled at 10kHz.

A detailed discussion on computer simulation results for each case mentioned above is stated below. First, we compare the performance of the proposed speech coding algorithm with that of the conventional LPC-10 algorithm. We take the reproduced speech quality of the LPC-10 algorithm as the reference for comparing with that of the proposed VTFL coding method. We obtain comparable output speech quality of the proposed VTFL speech coding method by varying the threshold of the LR distortion measure(d_{TH}). From the informal listening test, the proposed VTFL speech coding method with VQ encoding of model vectors at transmission bit rate of about 650bits/s yields the performance comparable to the LPC-10 at 2400bits/s. For the VTFL speech coding method with scalar quantization of model parameters, the comparable speech quality to LPC-10 is obtained at about 1400bits/s. Since the point-wise analysis method used in the proposed speech coding algorithm is suitable for tracking fast changing speech sounds, the proposed speech coder results in smooth and natural sounding output speech quality.

Second, for comparing the proposed representative feature vector selection method with the linear smoothing method, we use the feature vector obtained by averaging the PARCOR vectors extracted from 5th sample point to 15th sample point of the speech segment within a voiced region. Informal listening test shows that the proposed method of extracting the representative feature vector yields slightly improved performance than the linear smoothing method.

Third, performances of the proposed VTFL coding algorithm are obtained by varying the sample point at which the representative PARCOR vector is extracted. This point should lie in the stable portion of each segment in the voiced region. As shown in Fig.4, fluctuation of the feature vector is very small in the stable region of each segment. The PARCOR vectors are extracted at the 20th, 30th, and 40th sample

points. Using each of these feature vectors as the representative model vector, we obtain the performance of the proposed coding algorithm. As expected, the listening test shows no difference in performance among them. So, we use the PARCOR vector extracted from the 30th sample point as the representative model vector in that speech segment. From the above simulation results, we find that the proposed speech coding method yields good performance at low bit rates.

Ⅳ. CONCLUSION

In many voice communication applications, it is desirable to compress the transmission rates as much as possible while retaining a reasonable level of output speech quality. In this paper, the VTFL speech coding method has been presented. The primary concerns for improving the performance of this speech coder are segmentation, representative feature vector selection, and determination of the transmission frame length. Speech signal is segmented according to the nonstationarity of the input speech. And the representative feature vector for each speech segment is chosen at a certain sample point considering the convergence behavior of the PRLS algorithm. Then, by joining the consecutive speech segments of phonetically similar characteristics, we obtain the transmission frames of variable length. The representative feature vectors of these transmission frames are VQ-encoded and transmitted to the receiver. From the computer simulation results, the proposed VTFL coding method results in reduced transmission bit rates while maintaining the good reproduced speech quality.

REFERENCES

1. J.M. Turner and B.W. Dickinson, "A variable frame length predictive coder," IEEE Int. Conf. ASSP, pp.454-457, 1978.
2. V. Viswanathan, J. Makhoul, R. Schwartz and A. W.F. Huggins, "Variable Frame Rate Transmission : A Review of Methodology and Application to Narrow-Band LPC Speech Coding," IEEE Trans. Commun., vol.COM-30, NO.4, Apr. 1982.
3. P. Papamichalis and T. Barnwell, "Variable Rate Speech Compression by Encoding Subsets of the PARCOR Coefficients," IEEE Trans. ASSP, vol. ASSP-31, NO.3, June 1983.
4. C.K. Chung and S.W. Chan, "Speech Recognition Using Variable Frame-Rate Coding," IEEE ICASSP pp.1033-1036, Apr. 1983.
5. T. Fukabayashi and C.K. Chuang, "Speech Segmentation and Recognition using Adaptive Linear Prediction Algorithm," IEEE Int. Conf. ASSP, pp. 17.12.1-17.12.4, 1984.
6. Han-Choon Park and Hwang-Soo Lee, "Pitch-Synchronous Waveform Vector Quantization," KAIST communication lab.
7. J. Markel and et al., "Vector Quantization in Speech Coding," Proc. IEEE, pp.1551-1587, Nov. 1985.
8. A.B. Buzo and A.H. Gray, "Speech Coding Based Upon Vector Quantization," IEEE Trans. ASSP, vol.28, NO.5, Oct. 1980.
9. J. G. Wilpon and L.R. Rabiner, "A modified K-means Clustering Algorithms for use in isolated Word Recognition," IEEE Trans. ASSP, vol.33, NO. 3, June 1985.

▲Jeong Woo Jwa

1985.2 : Department of Electronics Engineering, Hanyang Univerisity(B.S.)

1987.2 : Department of Electrical Engineering, KAIST(M.S.)

1987.3 : Research Engineer at Korea Telecom

1992.3 : Department of Information and Communication Engineering, KAIST(Ph.D. Course)

▲Sung Ro Lee

1987.2 : Department of Electronics Engineering, Korea University(B.S.)

1990.2 : Department of Electrical Engineering KAIST(M.S.)

1990.3 : Department of Electrical Engineering KAIST(Ph.D.Course)

▲Hwang Soo Lee (Vol.11 No.4)

Associate Professor, Department of Information and Communication Engineering
KAIST