論 文 論文 81-18-1-6

Analysis and Synthesis of Korean Vowels by LP Method

(LP方法에 依한 한국모음의 분석과 합성)

손 호 인*, 신 동 진*, 안 수 길**
(SON, Hoin SHIN, Tongjin and ANN, Souguil)

要 約

사람의 목소리는 넓은 폭의 주파수 대역을 차지하지만 많은 redundancy를 포함하고 있다. 通信線路를 効率좋게 使用하고 작은 memory 용량으로 computer로서 소리를 합성하려면 data를 축소할 必要가 있다. 사람의 発声機関을 8次의 dynamic system으로 modelling 각 parameters를 求하였다. 이 parameter는 発声과정에서 変하기 때문에 20msec마다 交替하여 준다.

20msec analyzer와 synthesizer를 만들어 한국어 에서 /아/, /어/, /오/, /우/, /으/, /이/, /애/, 및 /외/ 等, 母音을 分析하고 parameter를 얻고 이 parameter를 利用하여 音声을 合成하였고 合成된 音声이 原発声者의 特色을 지녔음을 관찰하였다.

Abstract

The human speech contains many redundancies. To economize communication channel or memory size for a computerized synthesis of human voices, it is necessary to compress the data before sending.

We have treated human speech organ as an eighth order dynamic system which is time varying as the person speaks.

Using an anaylyzer of our design, each eight parameters are obtained for the vowels [아], [어], [오], [우], [으], [이], [애] and [외] of Korean language with considerable discrepancies between persons. Supplying those parameters to a synthesizer which we have made, we have succeeded in the simulation of human speech for the above mentioned vowels of Korean language and observed that they bear all the features of the original speakers.

I. Introduction

To increase the efficiency of communication utilities, we have every interest to diminish the band requirement for the transmitted information, that is, the narrower the necessary bandwith, the better the utility officiency. For the human voice, it is well known to a communication engineer that we need to send up to 3.4 Khz to permit a clear understanding.

Human speech spans a spectrum showing several peaks and valleys up to 10,000 Hz range, if we admit down to 60dB decrease from the maximum value.

The peaks on the curve are called formants and beginning with low frequency, we name the first, second, and third peak, the first formant, second formant, and third formant respectively. The organization CCITT (Comité Consultatif International de Téléphonie et Télégaphie) recommands to treat the voice frequency up to 4 Khz, thus obliging us to use 8 kbps as sampling frequency, because by the sampling theorem we must use the double of the maximum frequency of the orginal signal.

In that case, sometimes even the fourth formant

* 準会員, ** 正会員, 서울大学校 工科大学 電子工学科
(Dept. of Electronics Eng. Seoul Univ.)
接受日字 : 1980年 9月 5日

can be included when it is a voiced speech, for example the vowels. As for the unvoiced speech, the spectrum is single mode, the peak of which being around 3 Khz. The process is random and almost Gaussian.

Though its spectrum occupies large bandwidth, the human speech organ is rather slow one, when viewed as dynamic system. It will suffice to send the elements which distinguish each pronounciation-shape of vocal tract, place of the tong and some parameters of the (voice producing) system. This means that though its appearance is very complicated, the spectrum does not contain much information. The main part of the spectrum is redundancy and the number of linearly independent variables is quite limited. But in this case we can not put them as discrete variables.

It is rather represented as a continuity of variables forming a continuum of frequency of nonzero width. This width if we succeed in reducing the original signal to the minimum without any redundancy from which we might reconstruct the original voice signal with minimum complication. But that is an ideal to which we strive. The rather slow organ of human voice production is easily represented by a model of eighth order dynamic system. Since late sixties several researchers including FANT[4] studied the modelling of the speech organ to be followed by PETERSON,[18] SHOUP, FLANAGAN,[1-2] SAITO,[10] ITAKURA,[10] ATAL,[8] SCROEDER[17] etc. In the seventies, they found that it can be perfectly represented by eight parameters and the pitch frequency when voiced, and that those parameters do not vary much if observed each fiftyth of a second or less, resulting 3200 bps ($64 \times \frac{1}{50}$) when A-D converted with 8 bit resolution.

It means that we need less than 4000 bps when transmitting the pitch frequency. It is a great reduction of the necessary bandwidth when compared with the bandwidth of 64 Khz (8000x8) of normally sampled speech.

On the other hand, the progress and the wide diffusion of computers begin to demonstrate the diffusion of computers begin to demonstrate the necessity of better and easier interface between machine and human. And for most applications computers are asked to recongnize and compose human speech. If we do not use a magnetic tape recorder, we must store a 8K time-per-second sampled and A-D converted digital data.

Then we will consume the totality of memory of any 16bit address bussed microcomputer in one second and it will be quite a burden to even minicomputers.

We must call for any method of data compression, for example the linear prediction of speech which is possible because the movement of each part of speech organ is slow and the parameters which represent every features of human vocal tract, do not change rapidly, which permits us to send only eight parameters and pitch frequency each twenty milliseconds. This data reduction gives us any possibilitiy of letting a computer speak to human being economically in view point of memory capacity. If we succeed to communicate with computer by human speech, it will represent an easier interface with computers and this will open the new era of the genuine computer civilization. This branch is actually under sever competition and we will have our portion for the particularlity of our language, that is, we must put our efforts to synthesize the Korean language to cope with the researchers of other countries.

II. ARMA Process

In a random discrete process, if the observed value St at the time t is expressed by the weighted sum of the passed values S_{t-1}, S_{t-2}, S_{t-3},... S_{t-p}, and the value of present input at, then the process is classified as Auto-Regressive process of order p. And if the observed value St can be obtained by present and weighted sum of passed random inputs a_{t-1}, a_{t-2}, ... a_{t-q} etc. then the proces is classified as Moving-Average process of oreder q. A process in which we need P precedent observed values S_{t-1}, S_{t-2}, ... S_{t-p} and q precedent random shocks a_{t-1}, a_{t-2}, ... a_{t-q} as well as the present value a_t at input, we call it a mixed, Auto-Reqressive and Moving Average process of order (p,q).

Thus the present observation value St can be

given by the following expression.[5]

$$S_t = \phi_1 S_{t-1} + \phi_2 S_{t-2} + \cdots\cdots + \phi_p S_{t-p} + a_t - \theta a_{t-1} - \theta_1 a_{t-1} - \theta_2 a_{t-2} \cdots\cdots - \theta_q a_{t-q} \quad (1)$$

We can of course get the expression for the AR process of order P or MR process of order q by equating q=0, and p=0 each in turn in the equation(1) and the observed values $S_t, S_{t-1}, \cdots S_{t-p}$ are zero average values, that is, the deviations from the level (average value of the observed data) sum up to zero. This equation concerning the time series S_t can be solved easily by using the Z transform which can be defined as follows.[6]

$$Z \text{ transform of } \{S\} = Z\{S_t\} = \sum_{n=0}^{P} S_{t-n} Z^{-n} \quad (2)$$

and these equation can be extended to the extreme case of deterministic process. On the other hand we define a backward shift operator B as

$$B S_t = S_{t-1}$$
$$B S_{t-p} = S_{t-p-1}$$
$$B a_t = a_{t-1}$$
$$B^2 S_t = S_{t-2} \quad (3)[5]$$

Then the expression (1) will be[5]

$$(1 - \phi_1 B - \phi_2 B^2 \cdots\cdots - \phi_p B^p) S_t = (1 - \theta_1 B - \theta_2 B^2 \cdots - \theta_q B^q) a_t \quad (4)$$

III. Vocal Tract Transfer Function

For the human speech, the characteristic of the vocal tract is composed of those of[6]

(1) glottal shaping model; G(z)
(2) vocal tract model (which can be approximated as; V(z), all pole model)
(3) lip radiation model; L(z)

If we express the driving function which is the infinite train of pulses caused by the open and shut motion of the vocal chord by E(z) and the final articulated voice by S(z). We obtain[6]

$$S(z) = E(z)\, G(z)\, V(z)\, L(z) \quad (5)$$

The transfer function of the speech production mechanism $\frac{1}{A(z)}$ is equated as

$$\frac{1}{A(z)} \doteqdot G(z)\, V(z)\, L(z) \quad (6)$$

In this expression each factor has the following characteristics

1. $G(z) = \dfrac{1}{(1-e^{-CT} z^{-1})^2}$ characteristics of the glottal shaping model

2. $V(z) = \dfrac{1}{\prod_{i=1}^{k} [1-2e^{-C_i T}\cos(b_i T) z^{-1} + e^{-2C_i T} z^{-2}]}$

3. $L(z) = 1 - z^{-1}$ characteristics of lip radiation model

Final expression for the transfer function $\frac{1}{A(z)}$ will be[6]

$$G(z)V(z)L(z) = \frac{1-Z^{-1}}{(1-e^{-CT}z^{-1})^2 \left\{ \prod_{i=1}^{k} [1-2\, e^{-C_i T} \cos (b_i T) Z^{-1} + e^{-2C_i T} z^{-2}] \right\}} \quad (7)$$

The only factor $1-Z^{-1}$ in the numerator will be almost canceled by the $(1-e^{-CT}z^{-1})$ factor in the denominator, the real value of CT being very small and e^{-CT} being almost one.

The resultant transfer function is an all pole model.

IV. Linear Prediction

Changing a_t to E_t in (4) and limiting the case to the all pole model it will become[5]

$$S_t - \phi_1 S_{t-1} - \phi_2 S_{t-2} - \cdots - \phi_p S_{t-p} = E_t \quad (8)$$

Transforming into the Z domain

$$S(z)\left\{1 - \phi_1 z^{-1} - \phi_2 z^{-2} - \cdots - \phi_p z^{-p}\right\} = E(z) \quad (9)$$

To characterize a portion of human speech it will suffice to get the transfer function of the vocal tract $\frac{1}{A(z)}$ which is

$$\frac{1}{A(z)} = G(z)V(z)L(z) = \frac{1}{(1 - \sum_{k=1}^{P} \phi_k Z^{-k})} \quad (10)$$

A factor of G(z) being canceled by L(z), the order p is twice the number of the formants considered, augmented by one for the remainder of G(z).

If we estimate the present value of S_t by the weighted sum of the precedent values S_{t-1}, S_{t-2}, etc. then the estimate S_t will be[6]

$$\hat{S}_t = \phi_1 S_{t-1} + \phi_2 S_{t-2} + \cdots + \phi_p S_{t-p} \quad (11)$$

The error between the observed value at t and the estimation $\hat{S}t$ will be E_t[6]

$$S_t - \hat{S}_t = E_t \quad (12)$$

This equation mans that we can estimate the present observation value S_t by the weighted sum of the passed observation values St-1, St-2, St-3, ... St-p and the error in this case is the excitation value E_t, that is, we can estimate the value which is to occur within the excitation itself. Thus we estimate the future value when it does not occur yet. Simply we get the most natural state (expectable value of system) guessable from the latent system dynamics.

If we prepare two (or more) estimators to use one each at the transmitting end and the receiving end, then we can save the channel capacity drastically because the most part of the quantity to be transmitted will be autonomously prepared at both ends. It will suffice to send the discrepancy which may occur between the real value and the estimated value. In the eq.(12) the discrepancy does not bear any information. And we can call it "achromatic" or "bleached." It will be enough to send the pitch frequency and p prameter values of the system. To obtain exact values of parameters we must calculate them from as many as possible observed values and need a criterion to optimize the parameters. For our case we use the MMSE (Minimum Mean Square Error) criterion to get the following equation.

Total squared error P_E is expressed as

$$P_E = \sum_{\text{total n}} (S_t - \hat{S}_t)^2$$

$$P_E = \sum_{\text{total n}} E_t^2(n) = \sum_{\text{total n}} (S_t - \hat{S}_t)^2$$

$$= \sum_{\text{total n}} \left\{ \sum_{\substack{k=0 \\ z=0}}^{P} \phi_k S_{t-k} \phi_j S_{t-j} \right\} \quad (13)$$

This calculation is done over n sets of p observed valued St and if n is large enough $\sum_{\text{total n}} S_{t-k} S_{t-j}$

will give the autocorrelation R (k-j). Thus

$$R_{(k-j)} = \sum_{\text{total n}} S_{t-k} S_{t-j} = R_{(k-j)} = R_{(j-k)} \quad (14)$$

The total squared error will be[6]

$$P_E = \sum_{\substack{k=0 \\ j=0}}^{P} \phi_k R_{(k-j)} \phi_j \quad (15)$$

This is a quadratic form and the matrix R (k-j) is positive semidefinite as it is an autocorrelation function of a real time series $\{S_t\}$

The parameter ϕ_k will be obtained by differenciating the total squared error P_E by j, and equating them (j=1 ... p) to zero. The result is

$$2 \sum_{k=0} \phi_k R_{(k-j)} = 0$$

$$\sum_{k=1} \phi_k R_{(k-j)} = -\phi_o R(-j) = -\phi_o R_{(j)} = -R_{(j)} \quad (16)$$

$$\because \phi_o = 1 \qquad j = 1,2, \cdots, P$$

That is[5]

$$\begin{aligned} R(0)\phi_1 + R(1)\phi_2 + \cdots\cdots + R(p-1)\phi_p &= -R(1) \\ R(1)\phi_1 + R(0)\phi_2 + \cdots\cdots + R(p-2)\phi_p &= -R(2) \\ R(p-1)\phi_1 + R(p-2)\phi_2 + \cdots\cdots + R(0)\phi_p &= -R(p) \end{aligned} \quad (17)$$

This is Yule-Walker's equation and the matrix formed by the coefficient is a Toeplitz matrix. This simultaneous equation means that, given the autocorrelation function of the time series, we can calculate the best parameters that identify the voice producing mechanism at that moment. We are going to update each 20 ms the set of parameters as human speech is a

time varying process.

V. Correlation Cancellation Loop

As this analysis is a procedure to obtain each component in a Cartesian coordinate system, we need to find the values of correlations between the signal and the axis (a set of orthogonal functions) from (10) we see that[6]

$$E(z)\cdot\frac{1}{A(z)} = E(z)G(z)V(z)L(z) = S(z) \qquad (18)$$

$$S(z)A(z) = E(z) \qquad (19)$$

The latter tells that in passing the (information bearing) voice into a system with transfer function A(z) we obtain the original driving force E(z) which is train of pulses.

We introduce a voice signal to a system which adjust itself to render the signal "achromatic". If any component lingers in the achromaticed or "bleached" signal, we will readjust the system parameters so that there is virtually nothing except the train of pulses without any information. We are going to measure the components which are linearly independent each other if the coordinate is complete and orthonormal (CON). Each delayed signal has its own indapendance to form the necessary CON coordinate. We extract p components of the speech signal projected to the p axis which are $\phi_1, \phi_2, \phi_3, \ldots, \phi_p$.

Eq. (8) shows that in extracting those components from the signal S_t we obtain E_t. All that we do is amplifying the almost achromatic (treated) signal to remultiply to get the correlation of the treated signal and the p (axis forming) function set as in Fig. 1.

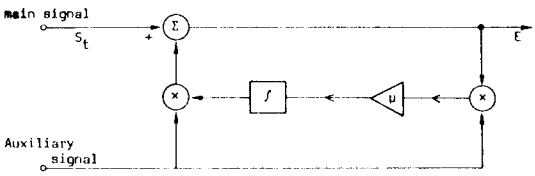

Fig. 1. CCL.

This circuit is well known as LMS algorithm. In Fig. 2 the input signal S_t is delayed by p time units to be used as the coordinate and each are multiplied by the output signal which may have still some information to obtain the correlation values with them. If there are some, they are amplified and summed to get better cancelation. The signals at the output of the intergrators (low pass filters) correspond to the parameters (ϕ_1-ϕ_p) when well compensated as they are the indication of the signal quantities (projected component values to each axis) to substract to render the input signal "achromatic", that is, they are the components of the signal in the function space. The value of p in our case is eight which permits us to cover up to fourth formant. For better articulation, this number seems to fail to be sufficient.

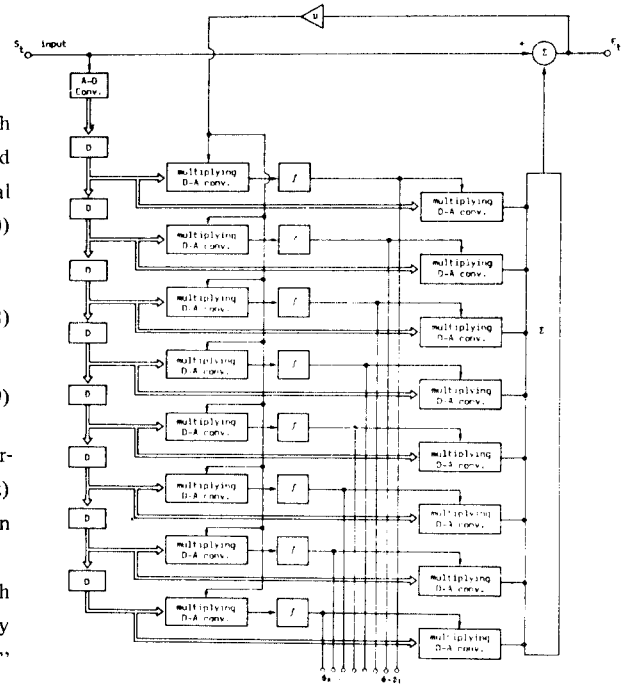

Fig. 2. Human speech analyzer.

VI. Speech Synthesizer

The parameters for a given vowel known, it is quite easy to synthesize the vowel. For the voiced consonants, the situation is the same. Both cases need the synthesizer to be excited by a pulse generator with the pitch frequency of the sound. The circuit is shown in Fig. 3.

On the reference voltage terminals, we feed the parameter voltages (ϕ_1, ϕ_2, ... ϕ_p). On the schema it is indicated by potentiometers but in reality we use

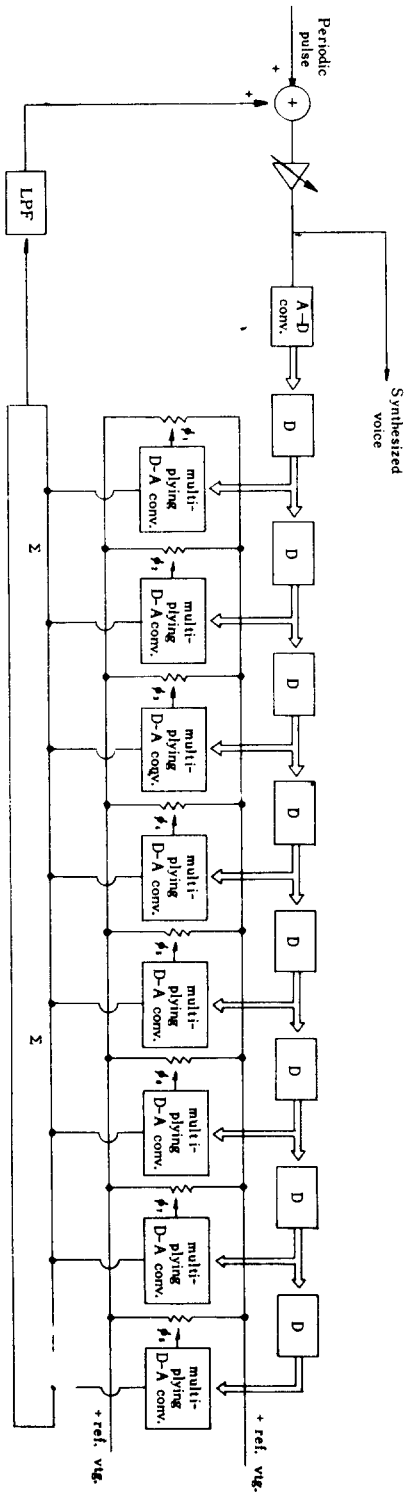

Fig. 3. Voice synthesizer.

another D – A converters as we have the necessity to change the parameters each 20 m sec. The delay elements are TTL 74198 shift registers. Integrators and adders are realized with the compensated operational amplifier μA 741 S. The circuit for A – D converter is shown in Fig. 4. We show also in photos the shapes and spectrum of each vowels.

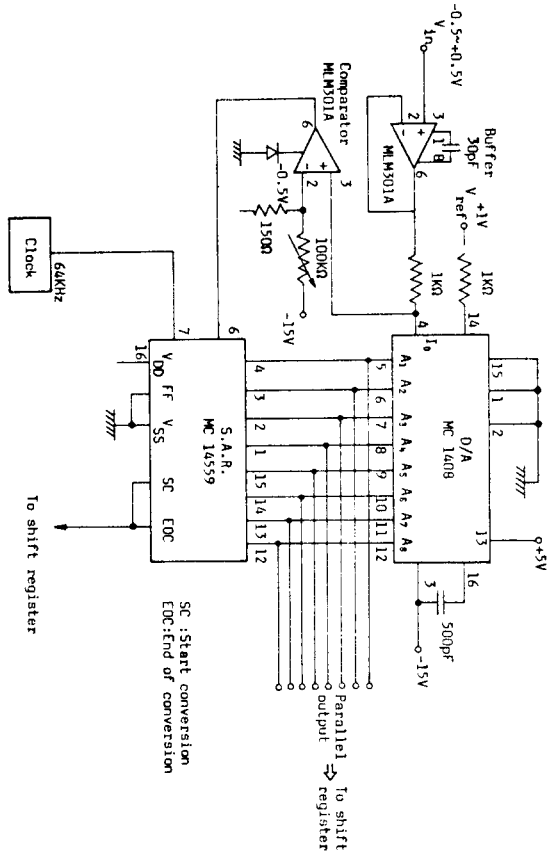

Fig. 4. A-D Converter circuit.

VII. Results and Discussion

We have measured the parameters of [아], [어], [오], [우], [으], [이], [애] and [의] by the circuit on Fig. 2. The results are shown in Table 1. Adjusting the reference voltages of the multiplying D – A converters after the values of parameters from the table and applying a pulse generator of 125 pps on the input terminals, we have obtained the expected synthesized voices.

Table 1. Parameters for Korean vowels which are obtained by the above mentioned analyser. The entries of each row indicate the parameters in volt of the vowel indicated in the 1st column. The general tendency is that the parameters in the left side column are bigger in absolute values.

	a_1	a_2	a_3	a_4	a_5	a_6	a_7	a_8
아	+1.56	+0.30	−0.50	−0.20	−0.30	−0.05	+0.24	+0.25
어	+1.91	+0.40	−0.42	−0.22	−0.45	−0.45	−0.02	−0.25
오	+1.73	+1.08	−0.27	−0.22	−0.50	−0.26	−0.17	−0.05
우	+1.80	+0.82	−0.31	−0.46	−0.17	−0.13	−0.04	−0.04
으	+1.44	+0.33	+1.03	−0.50	−0.05	−0.20	−0.40	−0.20
이	+1.75	+0.70	+0.51	+0.48	−0.02	−0.20	−0.30	−0.26
애	+1.87	+0.30	−0.42	−0.45	+0.02	−0.19	−0.17	−0.07
의	+1.56	+0.30	−0.50	−0.20	−0.30	−0.05	+0.24	+0.25

Photo 1. Waveform of excitation source.

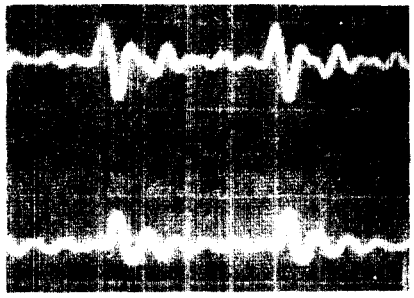

Photo 2. Waveforms of "아" original and synthesized.

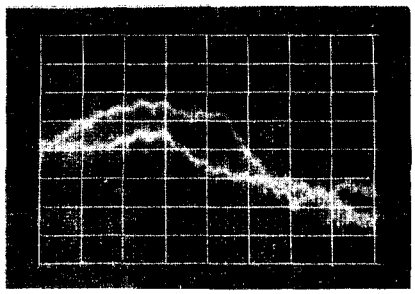

Photo 3. Spectrum of "아".

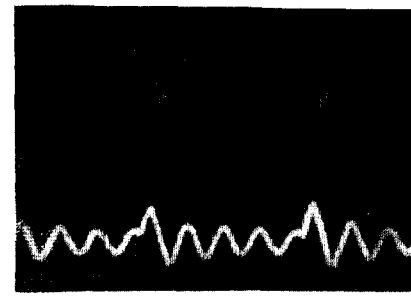

Photo 4. Waveforms of "어" original and synthesized.

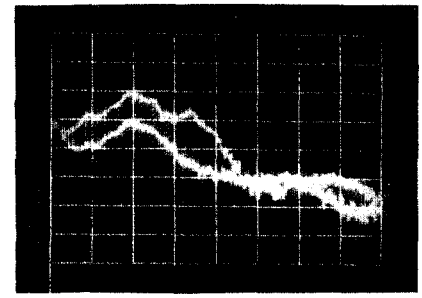

Photo 5. Spectrum of "어".

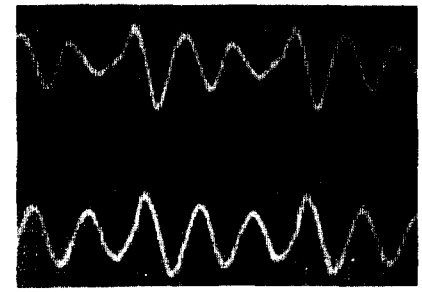

Photo 6. Waveforms of "오" original and synthesized.

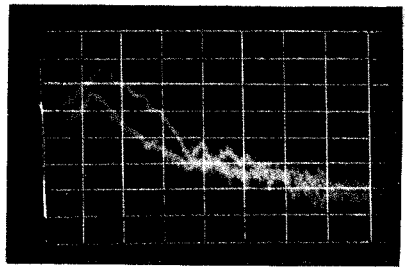

Photo 7. Spectrum of "오".

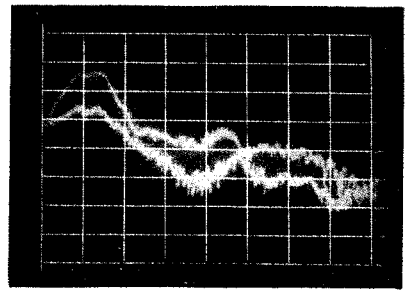

Photo 11. Spectrum of "으".

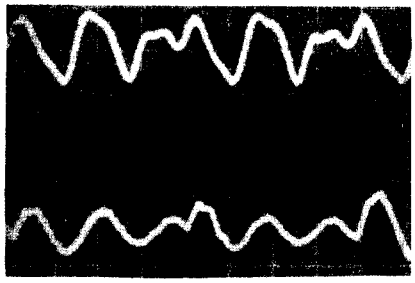

Photo 8. Waveforms of "우" original and synthesized.

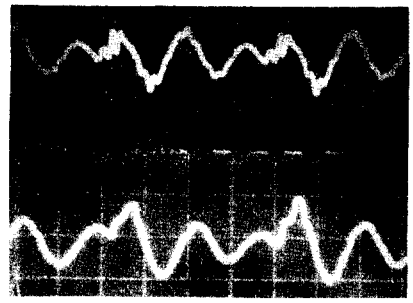

Photo 12. Waveforms of "ㅇ|" original and synthesized.

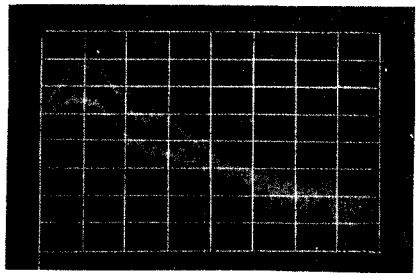

Photo 9. Spectrum of "우".

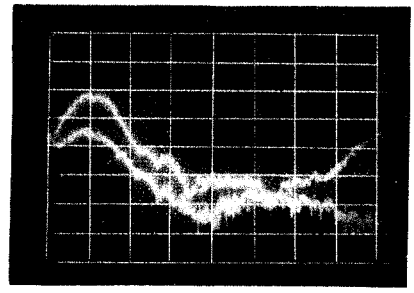

Photo 13. Spectrum of "ㅇ|".

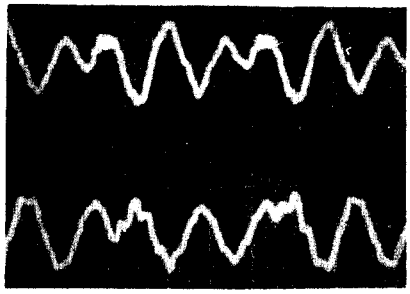

Photo 10. Waveforms of "으" original and synthesized.

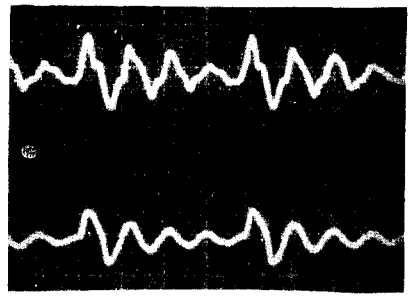

Photo 14. Waveforms of "애" original and synthesized.

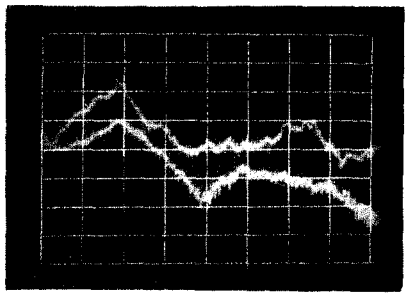

Photo 15. Spectrum of "ㅐ".

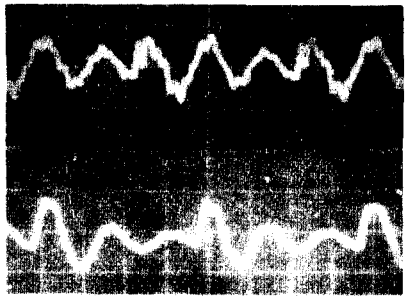

Photo 16. Waveforms of "ㅚ" original and synthesized.

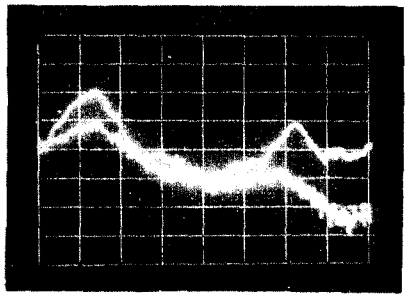

Photo 17. Spectrum of "ㅚ".

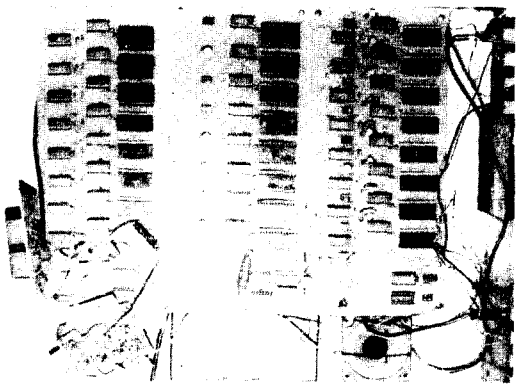

Photo 18. Synthesizer.

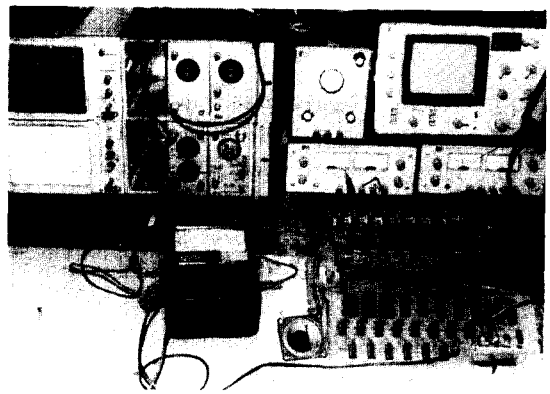

Photo 19. Setup of the experiment.

Analyzing the voice of a twenty seven year old boy with a peculiar feature (deep baritone), obtaining the parameters and synthesizing by the apparatus, we have suceeded to get very much him-like sound. But also we have observed that simply continuing a voice sound tends to be perceived as noise - kind of buzzing or humming. The not-human like impression can occur from a not-human like breath length and the human-like uniformness of the pitch frequency. On the other hand, we have studied the parameters of another twenty seven year old male to find them quite different from those of first. These discrepancies seem quite normal on account of the difference in voice features of two persons.

VIII. Conclusion

The human speech signal spans quite large spectrum. But it does not contain much information judging from the comparatively slow motion of the vocal organ. This vocal organ is modeled as an eighth order dynamic system for a limited duration of time, namely twenty milli second.

We have extracted the parameters of the vowels [아], [어], [오], [우], [으], [이], [애], and [외] of Korean language, and observed some discrepancies between persons which are quite normal. We have made an analyser and a synthesizer using a digital and analog hybrid technique to circumvent the difficulty of obtaining multitapped analog delay and multipliers. And we have succeeded to synthesize Korean vowels with the synthesizer and we could recongnize the speaker because of this voice features.

References

1. J.L. Flanagan, "Digital representation of speech signals," presented at the Bell Telephone Laboratory Symp. Digital Techniques in Communication, Nov. 12-13, 1970.
2. J.L. Flanagan, speech Analysis Synthesis and perception (and Edition). New York; Springer 1972.
3. J. Makhoul., "Linear Prediction; A tutorial review," Proc. IEEE, Vol. 63, pp. 561-580, Apr. 1975.
4. C.G.M. Fant. Acoustic Theory of Speech Production S-Gravenhase, Mouton and Co., 1960.
5. G.E.P. Box and G.M. Jekins, "Time series analysis: forcasting and control," 1970.
6. J.D. Markel, A.H. Gray, "Linear prediction of speech."
7. L.R. Rabiner, R.W. Shafer, "Digital processing of speech signals"
8. B.S. ATAL, S.L. Hanauer, "Speech analysis and Synthesis Soc. Am. Vol. 50, pp. 637-655, 1971.
9. J. Burg, "A New analysis Techniques for time series Data," Proc. NATO advanced Study Institute on signal proc. Vol. ASSP-25, No.5, pp.423-428, Oct. 1977.
10. F.I. Itakura and S. Saito, "Analysis-Synthesis Telephony Based upon the Maximim likelihood Method," proc. 6th Int. congress on Acoustics, pp. C17-20, Tokyo, 1968.
11. S.L.S. JACOBY, J.S. KOWALIK, J.T. PIZZO, "Iterative Methods for Nonlinear Optimization Problems," Prentice Hall, 1972.
12. B. Widrow, "Adaptive antenna systems," proc. IEEE Vol. 55, pp. 2143-2159, Dec. 1967.
13. "Adaptive filters," in Aspects of Network and system Theory, Part IV, R.E.Kalman and N. Declaris, Eds. N.Y. HOLT, Rinehart, and Winston, 1971, pp. 563-587.
14. L.E. Brennan and I.S.Reed, "Theory of adaptive radar," IEEE Aerosp. Electron. Syst. Vol. AE S-9, pp. 237-252. Mar. 1973.
15. R. Viswanathan and J. Makhoul, "Quantization Properties of Transmission parameters in linear predictive systems," IEEE Trans. Acoust. speech and signal proc. Vol. ASSP-23, No.3, June 1975.
16. C.K. UN and D.T. Magill, "The Residual-Excited linear Prediction Vocoder with Transmission Rate Below 9.6 Kbps," IEEE Trans. on Comm. pp. 1466-1473, Dec. 1975.
17. M.R. Schroeder, "Vocoders: Analysis and Synthesis of speech," proc. IEEE, vol. 54, pp. 720-734 May 1966.
18. G.E. Perterson and H.L. Barney, "Control methods used in a study of the vowels," J. Acoust. Soc. Amer., Vol. 24, pp. 175-184, 1952.