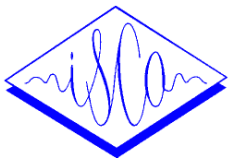

First International Conference on
Spoken Language Processing (ICSLP 90)
Kobe, Japan
November 18-22, 1990

SPEAKER IDENTIFICATION BASED ON MULTIPULSE EXCITATION AND LPC VOCAL-TRACT MODEL

Seiichiro HANGAI and Kazuhiro MIYAUCHI

Department of Electrical Engineering
SCIENCE UNIVERSITY OF TOKYO
1-3 Kagurazaka Shinjuku-ku Tokyo 162 JAPAN

ABSTRACT

In automatic speaker identification, the reduction of dimensions of template is a key to realize a quick identification and to save storage. In this study, we extract some glottal wave parameters which does not seem to be susceptible to mimicry and combine them with some LPC vocal tract parameters to make a smaller sized template. In the extraction of feature parameters of glottal wave, the multipulse excitation model is adopted under modifying the pulse search algorithm.

According to experimental results obtained from 30 speakers' 5 vowels, 94% identification rate with a template of 10 feature parameters (8 vocal tract and 2 glottal parameters) and 99% with a template of 36 feature parameters (14 vocal tract and 22 glottal parameters) are obtained respectively. In comparison with the number of parameters to get same identification rate only with a formants' template, the number of feature parameters is reduced by 5 in case of getting 94% identification rate and 7 in case of getting 99% identification rate.

1. INTRODUCTION

As feature parameters of template which represents speaker's characteristics, long time power spectrum[1], linear prediction coefficients[2], pitch contours[3] and Cepstral analysis [4] are reported. To get high identification rate with these parameters, a large number of parameters is required. In automatic speaker identification within large speaker population, the reduction of dimensions of template is required to realize quick response and to save storage. This means that parameters in the template should be orthogonal each other. Furthermore, parameters should be stable over time and not be susceptible to mimicry.

In the investigation of speech synthesis, it is reported that the multipulse glottal model[5] makes synthesized voice natural and gives individuality. This means that the locations and sizes of pulses involves individuality. This glottal parameter also seems to be orthogonal to the vocal tract parameter and be susceptible to mimicry.

In this paper, we describe the pulse search algorithm which is modified to match our aim, i.e., searching pulses not in a frame but in a pitch, and the reduction of dimensions of template. And, we also show the relation between identification rate and the number of parameters based on multipulse excitation and LPC vocal-tract model.

2. PROCEDURE FOR EXTRACTING FEATURE PARAMETERS

2.1 EXTRACTION OF VOCAL TRACT PARAMETERS

Fig.1 shows a flow of extracting individual feature parameters from speech signal. In order to get vocal tract parameters, speech signal is pre-emphasized and weighted by Hamming window. The emphasized and weighted signal is framed in 25.6ms and analyzed to get 12 LPC coefficients. When the signal is vowel, we can regard the transfer function H(z) of vocal tract as an all pole model and write H(z) as

$$H(z)=\frac{G}{1+a_1z^{-1}+a_2z^{-2}+\cdots\cdots+a_{12}z^{-12}} \qquad (1)$$

where G is constant and a_i is the i-th LPC coefficient. From this equation, the formant

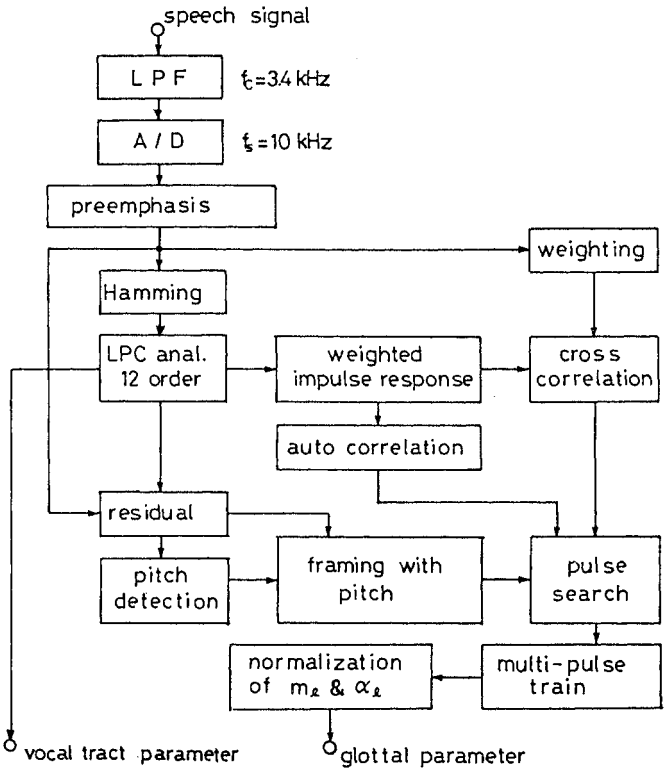

Fig.1 flow of extracting feature parameters

frequencies F_j and bandwidths B_j are given by complex roots as follows,

$$F_j=\frac{1}{2\pi T_s}\tan^{-1}[Im(z_j)/Re(z_j)] \quad (2.1)$$

$$B_j=\frac{1}{\pi T_s}\ln|z_j| \quad (2.2)$$

where z_j is one of complex roots and T_s is sampling period. We define 5 pairs of F_j and B_j as the vocal tract parameters.

2.2 EXTRACTION OF GLOTTAL WAVE PARAMETERS

To get glottal wave parameters, we adopt the multipulse excitation glottal model[5]. In this model, pulse train is determined to minimize squared error between original speech and synthesized speech by A-b-S method. This method, however, requires much calculation time. So, we use correlation technique to search pulses successively[6]. As shown in Fig.1, the emphasized speech and the LPC filter's impulse response which is derived from eq.(1) are perceptually weighted respectively. By obtaining cross correlation function of the weighted speech and the weighted response, the amplitude α_l of l-th pulse is successively calculated as follows ($0\leq l\leq L-1$, L:number of pulses),

$$\alpha_l(m_l)=\frac{\sum_{n=1}^{256}s_w(n)h_w(n-m_l)-\sum_{k=1}^{l-1}[a_k\sum_{n=1}^{256}h_w(n-m_k)h_w(n-m_l)]}{\sum_{n=1}^{256}h_w(n-m_l)h_w(n-m_l)} \quad (3)$$

where m_l is location of pulse α_l, $s_w(n)$ is the weighted speech and $h_w(n)$ is the weighted impulse response. In the original pulse search algorithm, pulses are determined in one frame. However, this is not suitable for our application, because we do not require natural sounding speech but need the information of locations and sizes of pulses per pitch. Therefore, we modify the algorithm to search pulses in one pitch period.

Fig.2 shows the procedure to get 10 pulses including main pulse (pitch pulse) per one pitch and the waveform, when the speech is vowel /a/. After getting 10 pulses, locations and sizes of pulses are normalized by those of main pulse and plotted on the polar-plane as shown in Fig.3 with cause of treating the pulse at rT_p ($0\leq r\leq 1$, T_p:pitch period) and the pulse at $(1-r)T_p$ as the same pulse.

We define 9 pairs of abscissa x_l and ordinate y_l of plots of subpulses in Fig.3 as the glottal wave parameters ($l=1,2,..,9$).

3. EVALUATION OF FEATURE PARAMETERS

In order to effectively reduce the amount of feature parameter, we measure the F-ratio ,i.e., inter-speaker variance to intra-speaker variance ratio for each parameter[7].

Table 1 shows the F-ratio of vocal tract and glottal wave parameters of 5 vowels (/a/,/i/,/u/, /e/,/o/), when 30 persons' 20 times utterances are used as sample. Parameters are sorted in order of F-ratio and 10 significant parameters in each vowel are listed. From Table 1(a), it is found

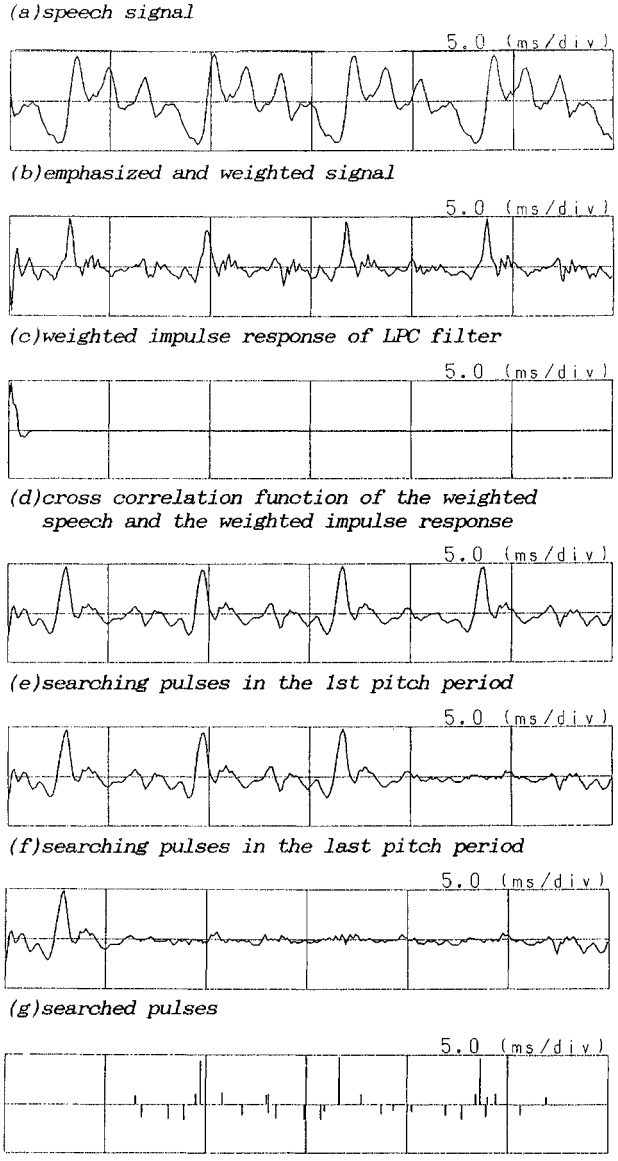

Fig.2 procedure to get 10 pulses per one pitch

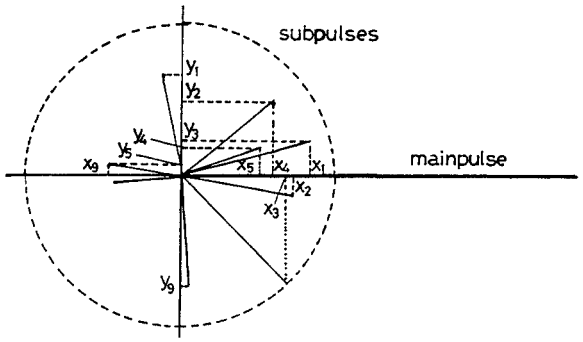

Fig.3 9 sub-pulses plotted on polar-plane

Table 1 F-ratio of vocal tract and glottal wave parameters of 5 vowels

(a) vowel /a/		(b) vowel /i/	
paramtrs.	F-ratio	paramtrs.	F-ratio
F_1	7.2	F_1	7.6
F_2	3.8	B_2	2.5
x_8	3.3	F_3	2.3
x_7	3.2	y_1	1.8
x_9	2.6	x_1	1.7
x_6	2.4	F_4	1.6
x_5	1.9	x_2	1.6
B_3	1.9	x_3	1.5
x_1	1.6	x_4	1.1
x_4	1.6	y_2	1.1

(c) vowel /u/		(d) vowel /e/	
paramtrs.	F-ratio	paramtrs.	F-ratio
F_3	4.4	x_7	2.8
B_1	3.7	x_8	2.7
y_9	3.5	x_6	2.6
F_2	3.4	x_9	2.3
x_9	2.8	x_5	2.1
F_1	2.2	F_1	2.0
x_1	2.1	y_1	1.6
y_1	2.1	x_4	1.5
x_7	1.8	F_4	1.4
x_8	1.7	y_5	1.3

(e) vowel /o/	
paramtrs.	F-ratio
F_2	5.9
F_1	4.4
x_9	3.0
x_8	2.3
y_9	2.2
F_3	1.8
x_1	1.6
x_7	1.6
B_1	1.6
B_3	1.5

that vocal tract parameters F_1, F_2 and B_3 of vowel /a/ have high F-ratio but that other 7 parameters are glottal wave parameters. From Table 1(b) to Table 1(e), 4 vocal tract and 6 glottal wave parameters in case of vowel /i/, 4 vocal tract and 6 glottal wave parameters in case of vowel /u/, 2 vocal tract and 8 glottal wave parameters in case of vowel /e/ and 5 vocal tract and 5 glottal wave parameters in case of vowel /o/ is significant parameters. The fact that many glottal wave parameters rank relatively high means the availability of the parameters in speaker identification.

4. SPEAKER IDENTIFICATION

By combining glottal wave parameters and vocal tract parameters, we attempt to reduce dimensions of template for speaker identification. The reference template is made from 5 times utterance of each vowel. In the identification, every Mahalanobis' distance between a template obtained from unknown speaker's 5 vowels and 30 reference templates is calculated and an unknown speaker is identified as the speaker who has a most similar template in the Mahalanobis space.

Table 2 Feature parameters in reference template

num. of paramtrs	feature paramtrs (combined)	feature paramtrs (vocal tract only)
1	/a/ F_1	/a/ F_1
2	/i/ F_1	/i/ F_1
3	/o/ F_2	/o/ F_2
4	/o/ F_1	/o/ F_1
5	/u/ F_3	/u/ F_3
6	/a/ F_2	/a/ F_2
7	/u/ B_1	/u/ B_1
8	/u/ y_9	/u/ F_2
9	/u/ F_2	/i/ B_2
10	/a/ x_8	/i/ F_3
11	/a/ x_7	/u/ F_1
12	/o/ x_9	/e/ F_1
13	/u/ x_9	/a/ B_3
14	/e/ x_7	/o/ F_3
15	/e/ x_8	/i/ F_4
16	/e/ x_6	/o/ B_1
17	/a/ x_9	/u/ B_4
18	/i/ B_2	/u/ F_5
19	/a/ x_6	/o/ B_3
20	/i/ F_3	/e/ F_4
21	/o/ x_8	/u/ F_4
22	/e/ x_9	/e/ F_5
23	/o/ y_9	/a/ F_4
24	/u/ F_1	/e/ B_1
25	/e/ x_5	/o/ B_4
26	/u/ x_1	/e/ F_3
27	/u/ y_1	/i/ F_2
28	/e/ F_1	/a/ F_3
29	/a/ x_5	/e/ F_2
30	/a/ B_3	/u/ B_2
31	/i/ y_1	/e/ B_3
32	/o/ F_3	/i/ B_3
33	/u/ x_7	/u/ B_3
34	/i/ x_1	/e/ B_4
35	/u/ x_8	/e/ B_2
36	/u/ x_6	/a/ B_4
37	/o/ x_1	/i/ B_4
38	/e/ y_1	/u/ B_5
39	/o/ x_7	/o/ F_4
40	/i/ F_4	/a/ B_1
41	/a/ x_1	/o/ F_5
42	/i/ x_2	/o/ B_5
43	/o/ B_1	/e/ B_5
44	/u/ x_2	/o/ B_2
45	/a/ x_4	/a/ F_5
46	/e/ x_4	/a/ B_5
47	/i/ x_3	/i/ B_5
48	/a/ y_9	/a/ B_2
49	/u/ B_4	/i/ B_1
50	/u/ F_5	/i/ F_5

Table 2 shows what kind of feature parameters are selected in the reference template as number of parameters is increased. The left table consists of glottal wave and vocal tract parameters and the right table consists of vocal tract parameters only. The table shows that no difference is found in both templates when the number of parameters is 7 or less. According to identification test, the identification rate is 16% by 1 parameter, 35% by 2 parameters, 49% by 3 parameters, 67% by 4 parameters, 77% by 5 parameters, 81% by 6 parameters and 83% by 7 parameters.

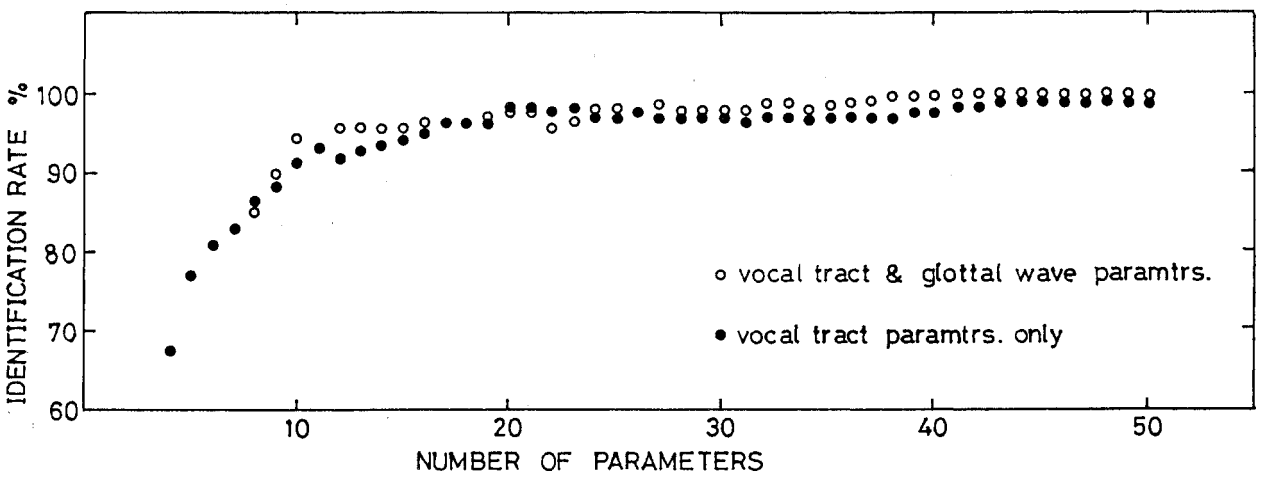

Fig.4 Relation between identification rate and number of parameters

Fig.4 shows the relation between the identification rate and number of parameters with the combined parameters and only with the vocal tract parameters. From the figure, the identification rate using combined parameters exceeds the identification rate using only vocal tract parameters by 0.5% to 2% when number of parameters is 24 or more. When number of parameters is between 17 and 23, however, the identification rate using only vocal tract parameters is better. This is caused by sorting parameters not in optimum order but in F-ratio order.

It is also found, from Table 2 and Fig.4, that 94% identification rate with a template of 10 feature parameters (8 vocal tract and 2 glottal parameters), 96% with a template of 16 feature parameters (8 vocal tract and 8 glottal parameters) and 99% with a template of 36 feature parameters (14 vocal tract and 22 glottal parameters) are obtained. In comparison with the number of parameters to get same identification rate only with formants' template, the number of feature parameters is reduced by 5 in case of getting 94% identification rate, 1 in case of getting 96% identification rate and 7 in case of getting 99% identification rate.

5.CONCLUSION

For the purpose of making a smaller sized template, we extract some glottal wave parameters which does not seem to be susceptible to mimicry and combine them with some LPC vocal tract parameters. As glottal parameters, abscissa and ordinate of subpulses plotted on a polar plane are used.

According to experimental results obtained from 30 speakers' 5 vowels, glottal parameters show relatively high F-ratio. In comparison with the number of parameters to get same identification rate only with a formants' template, the number of feature parameters is reduced by 5 in case of getting 94% identification rate, 1 in case of getting 96% identification rate and 7 in case of getting 99% identification rate.

Further improvement of identification rate might be done by adaptive subpulse treatment.

The authors wish to thank Masayuki Abe, a graduate student of the Science University of Tokyo, for his technical assistance.

REFERENCES

[1]S.Furui et al,"Talker recognition by longtime averaged speech spectrum," Trans. on IECEJ, 55-A, No.10, pp549-pp556, 1972

[2]M.R.Sambur,"Speaker recognition using orthogonal linear prediction," IEEE Trans. on ASSP ASSP-24, No.4, pp.283-pp289, 1976

[3]B.S.Atal,"Automatic speaker recognition based on pitch contours," J. Acoust.Soc.Amer., vol.52, pp.1687-1697, Dec., 1972.

[4]S.Furui,"Cepstral Analysis Technique for Automatic Speaker Verification," IEEE Trans. on ASSP, ASSP-29, No.2, pp254-pp272, 1981

[5]B.S.Atal et al,"A New Model of LPC Excitation for Producing Natural-Sounding Speech at Low Bit Rates," Proc. ICASSP, 82, pp614-617, 1982

[6]K.Ozawa et al,"Speech Coding Based on Multi-pulse Excitation Method," Tech. Rep of IEICEJ, CS82-161, pp115-pp122, 1982

[7]S.Furui,"Digital Speech Processing, Synthesis, and Recognition," pp302, Dekker, 1989