Integrated Visual and Speech Parameters in Korean Numeral Speech Recognition

Sang-Won Lee, In-Jung Park, Chun-Woo Lee, Hyung-Bae Kim

Department of Electronics, Dankook University,
Anseo-dong, Chonan-si, Choongnam, Korea
ijp21ce@anseo.dankook.ac.kr

Tel/Fax : +82-417-550-3544 H.P : +82-11-327-8542

Field : Multimedia Signal Processing
Keyword : visual information, speech information

Abstract

In this paper, we used image information for the enhancement of Korean numeral speech recognition. First, a noisy environment was made by Gaussian generator at each 10 dB level and the generated signal was added to original Korean numeral speech.[1] And then, the speech was analyzed to recognize Korean numeral speech. Speech through micorphone was pre-emphasized with 0.95, Hamming window, autocorrelation and LPC analysis was used. Second, the image obtained by camera, was converted to gray level, autocorrelated, and analyzed using LPC algorithm, to which was applied in speech analysis.

Finally, the Korean numerial speech recognition with image information was more ehnanced than speech-only, especially in '3', '5' and '9'.

As the same LPC algorithm and simple image management was used, additional computation algorithm like a filtering was not used, a total speech recognition algorithm was made simple.

I. Introduction

Speech recognition at living environment is more difficult than laboratory, because of the environmental noise. A filtering method in pre-processor part within speech recognizer can remove it from noisy speech. But, some problems might be generated, which are the increase of removing time of noise from noisy speech, complexity of algorithm to process it.

So, for the solution of above problems and enhancement of speech recognition, in this paper, we suggest a method that speech and image are processed simultaneously and image is processed in the same algorithm as a speech recognition.

The reaseon for using image information is the fact that it is not affected by acoustic noise and has the same value through speech signal processing.[2]

II.Speech and Video Signal Processing

Mouth image is captured by video camera and converted to AVI file by capture board. And simultaneously, Korean numeral speechs are recored and stored in compter memory.

The start point of moving mouth and pronouncing speech should be synchronized.

1. Speech signal processing

Speech signal parameters are extracted with some processings such as speech capture, lowpass filtering, analog to digital converting, pre-emphasis, windowing, and LPC computations. [3]

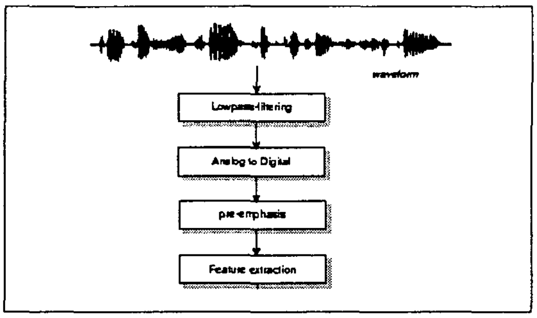


Fig 1. Flow diagram of speech parameter extraction

Pre-emphasis : $'s(n) = s(n) - 0.95 \cdot s(n-1)$

Windowing :

$$w(n) = 0.54 - 0.46 \cdot \cos\left(\frac{2n\pi}{N-1}\right)$$

Autocorrelation :

$$r(j) = \sum_{m=0}^{N-1-j} s(m)s(m+j) \quad (0 \le j \le P)$$

where P is the order of parameter.

LPC computation : Levinson-Durbin Algorithm

2. Video signal processing

(1) Capturing mouth information

The images are captured by a rate of 15 frame per a second. The mouth images are separated into frame by frame to extracted their visual information.

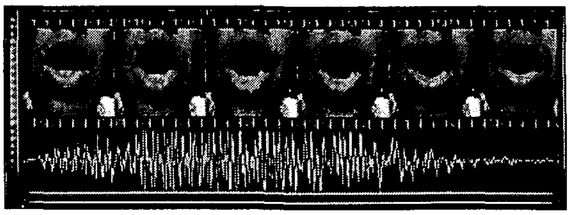


Fig 2 . Image sequence obtained from camera

(2) LPC method of mouth image

Before extracting their parameters, colorful mouth image is converted into gray image for less information. And then, 2-dimensional LPC analysis is used. Because the image is consist of data by raster scan, the prediction of pixel value is as follows.

	i	
	$g(i-1,j-1)$	$g(i,j-1)$
j	$g(i-1,j)$	$g(i,j)$

$g(i,j)$ value is estimated by following equation.

$$\hat{g}(i,j) = a_1 g(i-1,j-1) + a_2 g(i,j-1) + a_3 g(i-1,j)$$

$\hat{g}(i,j)$: estimated value at point (i,j)

a_1, a_2, a_3 : prediction coefficients

III. Recognizer for the experiments

A neural network is used as a recognizer in this experiment.[4][5]

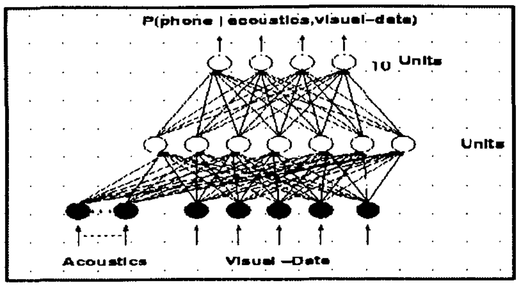


Fig 3. Multilayer nueral network

Ⅳ. Parameters for experiments

1. Configuration of the system

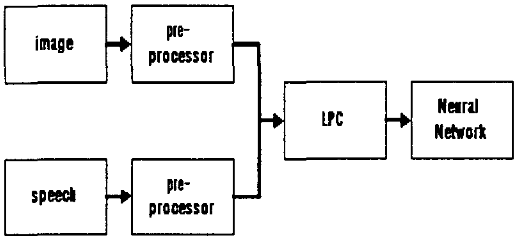
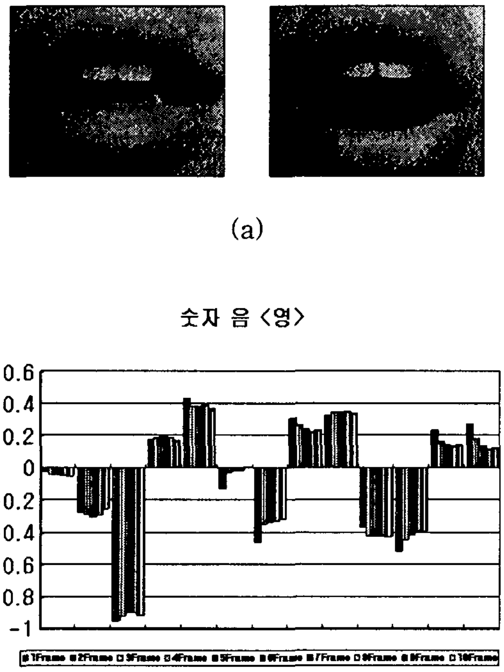


Fig 4 . Configuration of the system

2. LPC parameters of image



(a)

(b)

Fig 5. (a) mouth image of Korean numeral speech, '0'
(b) LPC coefficients of above image

3. Related parameters

To analyze acoustic information, the following parameters are used :

Sampling frequency : 11.05 KHz
Quantization : 8 Bits
Window function : Hamming window
the order of LPC parameters : 15th-order.

To analyze image information, the following parameters are used :

Color quality : gray level
Size of image : 100 × 80
Image feature : 2 dimensional LPC

Ⅴ. Discussion and Conclusion

In this paper, the inter-relation between the acoustic information and visual information through camera has been studied and a new method which is appropriate to the recognition of Korean numeral speech, was suggested. The acoustic information could be analyzed by high pass filter which has a parameter of 0.95, Hamming window, and autocorrelation analysis. It can compose the 15th-order linear prediction coefficients obtained by Levinson-Durbin algorithm.The image information could be converted gray from color image. And then, image autocorrelation was applied to get LPC coefficients. Like a acoustic analysis, the same Levinson-Durbin algorithm was applied(see Fig 4).

So, total flow of recognition algorithm was simple and additional complexity of algorithm also was lower.

To make a noisy situation, Gaussian noise generator was used, and the experiment was performed according to the noise levels. A table 1. shows that adding visual information to acoustic information is more effective and using 2 dimensional LPC is more efficient.

References

[1] PAUL M.EMBREE, BRUCE KIMBLE "C LANGUAGE ALGORITHMS FOR DIGITAL SIGNAL PROCESSING", Prentice-Hall Internati-ional, Inc. 1991

[2] T. Chen and Ram R. Rao, "Audio-Visual Integration in multimodal Communication," *Proc. of the IEEE*, vol. 86, no. 5, pp 837-852, May 1998.

[3] John R.Deller, Jr., John G.Proakis and John H.L. Hanson, "Discrete-Time Processing of Speech Signal", Macmillian Publishing Company, 1993.

[4] In-Jung Park, Chun-Woo Lee, Ho-Sung Chang, "A Fast Algorithm for Training Multilayer Perceptron Model of Neural Network", NNASP, pp.311 ~ 317, 17-20, August, 1993.

[5] Patrick K.Simpson, "Artificial Neural Systems" PERGAMON PRESS, 1990.

Table 1. The result of speech recognition without/with image information

	'3'		'5'		'9'	
dB	speech-only	image add	speech-only	image add	speech-only	image add
0	0.9962	0.9960	0.9963	0.9964	0.9958	0.9959
10	0.9930	0.9947	0.9900	0.9918	0.9916	0.9950
20	0.9835	0.9919	0.9322	0.9799	0.9563	0.9833
30	0.8898	0.9836	0.12211	0.9049	0.5877	0.8061
40	0.2092	0.9443	0.0006	0.3821	0.0495	0.3523