SPOKEN WORD RECOGNITION WITH DIGITAL COCHLEA USING 32 DSP-BOARDS

Masao Namiki, Takayuki Hamamoto, Seiichiro Hangai

Department of Electrical Engineering,
Science University of Tokyo,
1-3 Kagurazaka, Shinjuku-ku, Tokyo, 162-8601, JAPAN

ABSTRACT

A digital cochlea, which has a cascade of 16 filter sections, is realized by 32 commercially available DSP-boards. Each section consists of travelling waves filter, velocity transformation filter and second filter. The artificial cochlea is also applied to spoken word recognition by feeding 16 output signals through a multi-channel A/D converter on PC From experimental results, it is found that 50 Japanese words uttered by three speakers are recognized with 3% error. This means the cochlea extracts feature parameters for speech recognition and shows the possibility of the signal processor for the cochlear implants.

1. INTRODUCTION

For auditory handicapped person, the realization of an artificial cochlea with many electrodes to extract spectrum envelope in detail is desired. However, commercially available cochlea has 6 to 22 electrodes and gets features by SMSP or SPEAK strategies[1]. Even under such a condition, many handicapped persons improve hearing abilities and get high speech perception scores month by month[1] Therefore it is well expected for the person to improve the ability with short term, if the number of electrodes increases.

Digital cochlear model, which was suggested by Kates in 1991, had 112 outputs[2]. It had the advantage of computational efficiency and numerical stability[3]. In the previous research, we have constructed the model with 87 outputs on PC and applied it to 50 Japanese words recognition under noisy environment. The model extracts temporal-spectral features and gives 83% recognition rate at 0dB-SNR. In the experiment, however, it was difficult to get results in real time, because the huge calculation is required to convolution process.

In this research, we construct the digital cochlea by using 32 DSP-boards, TMS320C3xDSK, and find the possibility of realization of the artificial cochlea from experimental results. We also discuss the extension of the model and the performance of spoken word recognition under noisy environment.

2. INSTALLATION OF DIGITAL COCHLEA

2.1 Reduction of Sections of the Time-Domain Digital Cochlear Model

Digital Cochlear Model, which was developed by Kates, consisted of a cascade of digital filter sections. It extracted spectrum in detail and fed pulse trains from each section. However, 112 sections covering the frequency range of 100Hz-16kHz with 40kHz sampling is difficult to realize by DSPs and seems to be over specification for speech recognition. So, we check the influence of reducing the number of sections on speech recognition rate.

Fig.1 shows the relationship between the speech recognition rate and the number of sections from experimental results. In the experiment, 50 kinds of Japanese words uttered by 60 speakers are used. From this figure, in noisy environment such as SNR=10dB or less, many filter sections are required. However, in noise free environment, it is found that the reduction does not affect the recognition rate, if the number is more than 10. So, it is good for us to try to make the digital cochlea with 32 DSP-boards which are available by our responsibility.

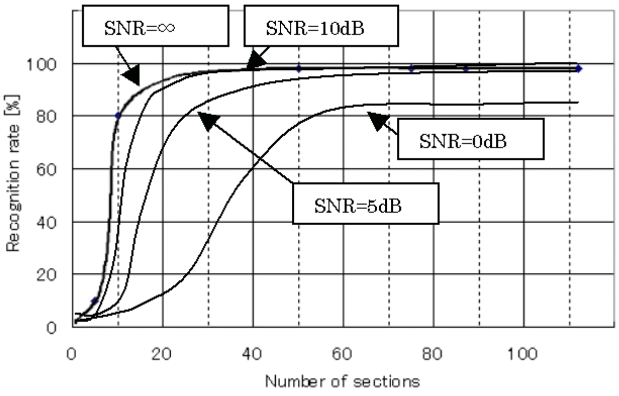


Fig.1 Speech recognition rate VS Number of sections

Fig.2 shows a cascade of digital filter sections. Each section has a traveling-wave filter H(z), a velocity transformation filter T(z) and a second filter F(z). The filter is not same as the digital cochlear model proposed by Kates, because there is no feed back path for adjusting Q for simplification.

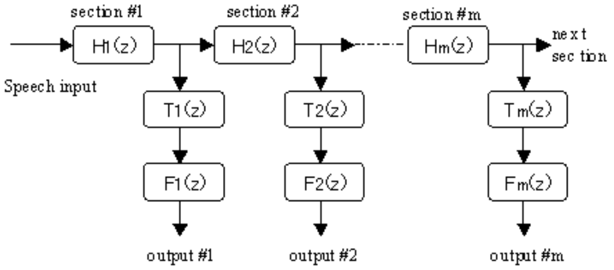


Fig.2 A cascade of digital filter section

2.1.1 Traveling-Wave Filter

Speech signal travels a cascade of traveling-wave filters from left to right in Fig. 2. Each filter shows low pass filtering characteristics and has different cut off frequency.

The transfer function $H_m(z)$ of the m-th traveling-wave filter given by

$$H_m(z)=\frac{a_m(1+z^{-1})}{(a_m+\mu)+(a_m-\mu)z^{-1}}$$

$$\times\frac{(a_m{}^2Q_m+a_m(Q_m\mu+1)+b\mu)+2(a_m{}^2Q_m-b\mu)z^{-1}+(a_m{}^2Q_m-a_m(Q_m\mu+1)+b\mu)z^{-2}}{(a_m{}^2Q_m+a_m+Q_m)+2(a_m{}^2Q_m-Q_m)z^{-1}+(a_m{}^2Q_m-a_m-Q_m)z^{-2}} \tag{1}$$

$$a_m=\tan(\tfrac{\pi f_m}{f_s}) \quad f_r\ \text{:sampling frequency}$$

$$Q_m=0.10\log_{10}({}^{f_m}\!/\!_{160}+0.8)+0.26 \quad \mu=0.5,\ \ b=0.5$$

The transfer function gives the third-order low pass filtering characteristics.

2.1.2 Velocity Transformation Filter

The output of traveling-wave filter is fed through the second filter after low frequency components removed by the velocity transformation filter. Its transfer function is given by a one-pole high pass filter with the cut off frequency of two octaves below the center frequency of each segment. The transfer function $T_m(z)$ of the m-th velocity transformation filter is given by

$$T_m(z)=\frac{1-z^{-1}}{(\frac{a_m}{4}+1)+(\frac{a_m}{4}-1)z^{-1}} \tag{2}$$

2.1.3 Second filter

In order to simulate the behavior of a cochlea, we make the second filter, which has a notch at one octave below of the center frequency of each section. The transfer function $F_m(z)$ of the m-th second filter is given by

$$F_m(z)=\frac{4(b_{m0}{}^2+\frac{b_{m0}}{Q_{m0}}+1)+8(b_{m0}{}^2-1)z^{-1}+4(b_{m0}{}^2-\frac{b_{m0}}{Q_{m0}}+1)z^{-2}}{(b_{mp}{}^2+\frac{b_{mp}}{Q_{mp}}+1)+2(b_{mp}{}^2-1)z^{-1}+(b_{mp}{}^2-\frac{b_{mp}}{Q_{mp}}+1)z^{-2}} \tag{3}$$

$$b_{mp}=\tan(\tfrac{\pi f_m}{f_s}) \quad f_r\ \text{:sampling rate}$$

$$b_{m0}={}^{b_{mp}}\!/\!_2\ ,Q_{m0}=2Q_{mp}\ ,Q_{mp}=1.5(1+\tfrac{fm}{1000})$$

2.2 The Realization of Digital Cochlea filter using 32 DSP-boards

The cochlear model proposed by Kates gives temporal-spectral features of speech signal in detail by 112 sections of the model. On the other hand, from the point of view of hardware installation, it is desired to reduce the number of sections as small as possible. As described above, it is expected that more than 10 sections of the cochlear filter give sufficient recognition rate if the environmental noise is free. Therefore, we have decided to use 32 DSP-boards to realize the digital cochlea with 16 sections. Table 1 shows the section number and the corresponding center frequency, when the sampling frequency is 16kHz. Frequency response of each section to the speech input is shown in Fig. 3.

Table 1 Center frequency of each section

section	frequency[Hz]	section	frequency[Hz]
1	7531.20	9	979.41
2	5886.50	10	741.61
3	4594.97	11	554.87
4	3580.78	12	408.24
5	2784.37	13	293.09
6	2158.98	14	202.66
7	1667.88	15	131.66
8	1282.24	16	75.90

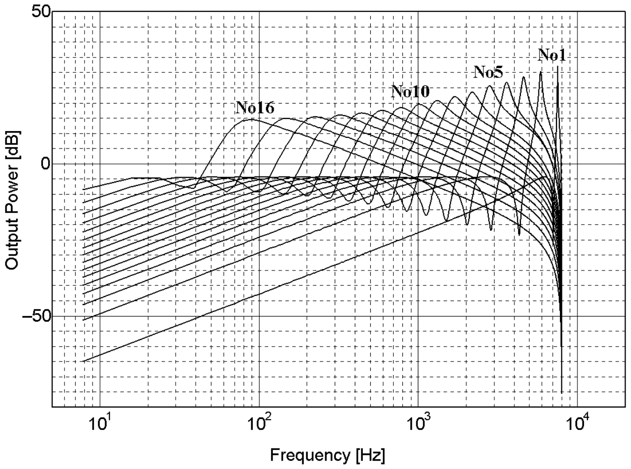


Fig.3 Frequency response of each section

From the figure, each section represents band pass characteristics with notch at one octave below the center frequency. Fig.4 shows the schematic diagram of the realized cochlea filter. In each section, we allocate one DSP-board for the traveling wave filter and another DSP-board for the velocity transformation filter and the second filter including post processing, i.e., mean square processing over 160 samples. After the post processing, 16 signals are multiplexed and acquired into the PC by the A/D converter with 200Hz sampling.

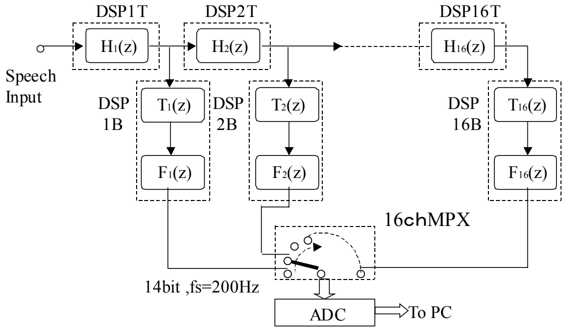


Fig.4 Schematic diagram of the realized cochlear filter

The overview of the realized digital cochlea is shown in Fig.5.

The utilized dsp-board is TMS320C3xDSK board, which includes TMS320C31 (50MFLOPS, 25 MIPS) and TLC32040 Analog Interface Circuit. Therefore, all signals between boards are easily checked by an oscilloscope. Code for each DSP is sent via printer port from the PC.

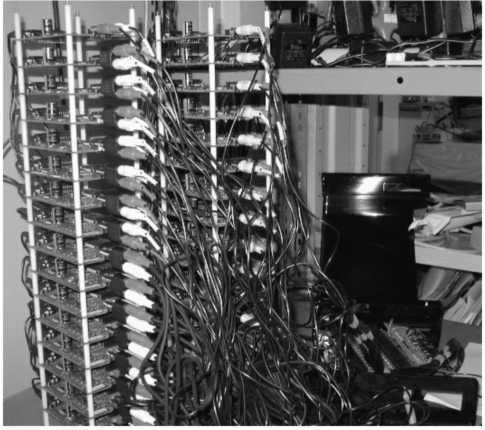


Fig.5 Overview of Digital Cochlea

Table 2 shows the processing time in the DSP to realize a travelling wave filter. As the sampling period is about 60us, there is no problem in real time processing.

Table 2 Processing times in a traveling-wave filter

Category of the process	Processing times [μs]
Initializing	6.36
Filtering	1.80
Sending data to AIC	1.36
Interrupt	0.36
Total (except for Initializing)	3.52

Table 3 shows the processing time in the DSP to realize a velocity transformation filter, a second filter and mean square calculation. In this DSP, we cannot find any problem in real time processing.

Table 3 Processing times for concatenated filtering and mean square calculation.

Category of the process	Processing times [μs]
Initialize	13.00
Filtering and MS calculation	9.60
Sending data to AIC	1.08
Interrupt	0.36
Total (except for Initializing)	11.04

From these tables, we find that all process in respective section can be done in one DSP board. Therefore, if the DSP board has two analog ports, we can realize a digital cochlea with 32 sections.

3. EXPERIMENTS OF SPOKEN WORD RECOGNITION

3.1 Frequency Response

Before recognition, we check the actual frequency response of each section in the digital cochlea. Fig. 6 shows the experimental results and theoretical results of section #8.

From results, experimental results agree with the theoretical curve in 100Hz-7kHz. Actual dynamic range is about 35dB. We have certified that other sections also show the similar frequency characteristics.

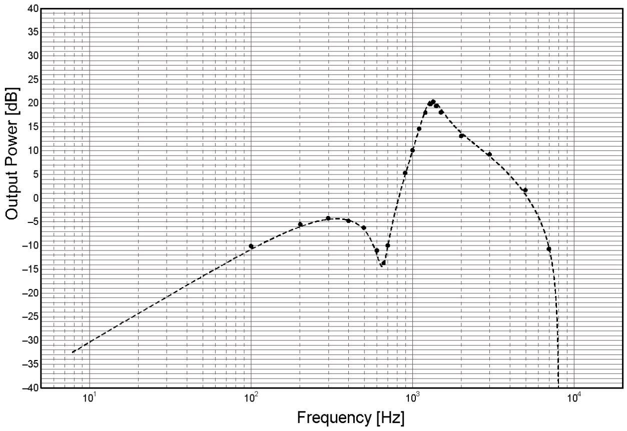


Fig.6 Experimental and Theoretical Frequency Response of section #8

3.2 Spoken Word Recognition

In order to examine the word recognition performance, we feed spoken words through the digital cochlea, and recognize them on the PC using the feature signal from 16 sections. Fig.7 shows the output pattern, when a Japanese word "Genzaich" is uttered. 16 signals are obtained from the digital cochlea every 5ms. Although the frequency resolution is not fine, we can find features from the pattern. Outputs are normalized by the peak power and the time is also normalized for matching in registration process and recognition process.

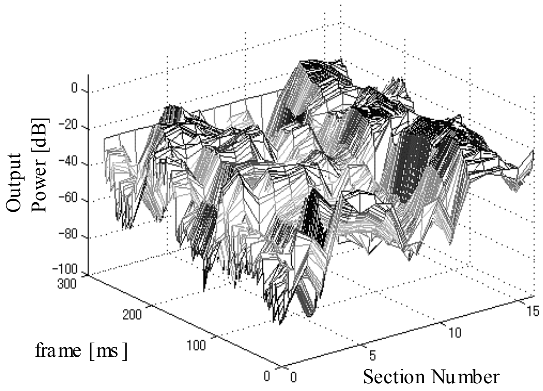


Fig.7 Output pattern of “Genzaichi"

Table 4 Word Recognition Rate

Word & Recog. Rate in %		Word & Recognition Rate in %	
Genzaichi	100	Jidousya	100
Mokutekichi	100	Gasorin Stand	100
Ru-to Tansaku	100	Cyuusyajyou	100
Ru-to Sai Tansaku	100	Konbiniensusutoa	100
Syuuhennkensaku	100	Famiri resutoran	100
Hukki	100	Byouin	100
Kakudai	100	Ginko	100
Syukusyo	100	Yuubinkyoku	93.3
Yarinaoshi	100	Eki	96.6
Mouichido	100	Syukuhakusisetu	100
Chyuusi	100	Keisatu	90.0
Syuuryou	100	Yakusyo	100
Jyuutaijyouhou	100	Hokkaido Sapporoshi	90.0
Ryoukin	100	Miyagiken Sendaishi	100
Ichiranhyou	100	Toukyouto Meguroku	96.6
Hai	100	Aichiken Nagoyashi	100
Iie	100	Hyougoken Koubeshi	100
Keiyuchi	100	Hirosimaken Fukuyamashi	100
Gaidopoint	100		
Tugiha	100	Fukuokaken Fukuokashi	96.6
Jyuusyo	100		
Jitaku	100	Fukushimaken Fukushimashi	90.0
Kaisya	100		
Kousaten	83.3	Wakayamaken Wkayamashi	86.6
Kousokudouro	100		
Inter change	100	Okayamaken Okayamashi	86.6
Parking area	100		
Service area	100	Average	97.6

In registration process, we make the database of 50 words uttered by 3 speakers. The data is made by averaging 5 patterns(each speaker uttered 5 times) after normalization. In word recognition process, we calculate the similarity between the tested patternP1 and all of registered patternsP2 by the following equation,

$$r(P1,P2)=\left[1+\frac{1}{MN}\sum_{n=0}^{N-1}\sum_{m=0}^{M-1}\{P1(m,n)-P2(m,n)\}^2\right]^{-1} \quad (4)$$

where M is the number of frames, and N is the number of sections.

Table 4 shows the average recognition rate of 50 spoken words uttered by 3 speakers with 10 times under noise free environment. From the table, it is found that the average recognition rate is good enough to recognize 50 words.

3.3 Recognition Rate under Noisy Environment

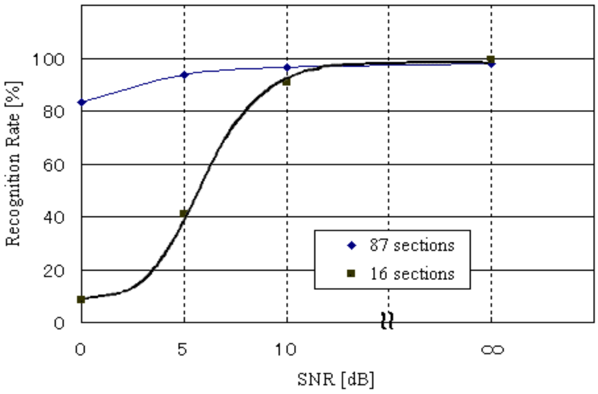
Our ears have an ability to hear and recognize spoken words under noisy environment. So we expect the cochlear filter to suppress disturbing components. Actually, the cochlear filter with 86 sections composed on a PC has shown a good performance even when the SNR is 0dB[4]. However, by the hardware limitation, we reduce the number of sections and examine the recognition performance against noise.

Table5 and Fig.8 show the relationship between the mean recognition rate of 50 words uttered by 60 speakers and SNR.

Table 5 Comparison of Recognition Rate

SNR [dB]	87 sections [4]	16 sections on DSP by 1speaker
0	83.27%	8.6%
5	93.58%	41.0%
10	96.48%	90.6%
∞	98.00%	99.2%

Fig. 8 Comparison of Recognition Rate



Although results of 87 sections processed on a PC shows a good performance, the cochlear filter with 16 sections shows undesirable degradation especially in low SNR such as 5dB or less. We think that this is simply caused by the reduction of numbers of sections. So, in near future, we can achieve the low recognition error by developing the cochlear filter with many sections. This means an increasing of sections and the number of electrodes which enables handicapped persons to improve hearing abilities and get high speech perception scores with short term.

4. CONCLUSION

For the purpose of realizing a cochlear filter for auditory handicapped person, the digital cochlear model, which was suggested by Kates, is evaluated on the PC and installed by 32 DSP-boards, TMS320C3xDSK. The cochlear model with 87 sections shows good performance in recognizing 50 words even when the SNR is 0dB. However, the installed cochlear filter is affected by the noise drastically, because the number of sections is 16, one seventh of the Kates model.

In near future, we can achieve the low recognition error by increasing the number of sections and the number of electrodes which enables handicapped persons to improve hearing abilities and get high speech perception scores with short term.

References

[1] P.C.Loizou: "Mimicking the Human Ear", IEEE Signal Processing Magazine, vol.15, No.5, pp101-130, 1998

[2] J.M.Kates: "A Time-Domain Digital Cochlear Model", IEEE Trans. on Signal Proccessing, vol.39, No.12, pp2573-2592, 1991

[3] J.M.Kates: "Accurate Tuning Curves in a Cochlear Model", IEEE Trans. on Speech and Audio Processing, vol.1, No4, pp453-462, 1993

[4] T. Harada, T.Hamamoto, S.Hangai: "An Improvement of Speech Recognition using Digital Cochlear Model under the Noisy Environments", Proc. of National Conf. of IEICE, D-14-32, pp253, 1999 (in Japanese)