Voice Expression using a Cochlear Filter Model

Soon Suck Jarng*

Abstract

Speech sounds were practically applied to a cochlear filter which was simulated by an electrical transmission line. The amplitude of the basilar membrane displacement was calculated along the length of the cochlea in temporal response. And the envelope of the amplitude according to the length was arranged for each discrete time interval. The resulting time response of the speech sound was then displayed as a color image. Five vowels such as a, e, i, o, u were applied and their results were compared. The whole procedure of the visualization method of the speech sound using the cochlear filter is described in detail. The filter model response to voice is visualized by passing the voice through the cochlear filter model.

I. Introduction

A main part of speech sound processing for voice recognition may be the extraction of speech sound characteristics. Most of research works for the characteristic parameter extraction of the speech sound have been done by means of statistical or random signal processing techniques [1]. Recently, spectral analysis methods that are physiologically based was introduced, and it was shown that auditory-based signal processing could be more robust to noise and reverberation than alternative spectral analysis procedures [2]. According to Ghitza's EIM (Ensemble Interval Histogram) model, an input sound source was 'simultaneously' applied to a bank of filters that modeled the frequency selectivity at various points along a simulated basilar membrane [2]. The phase characteristics of the corresponding cochlear filters was minimum phase. It means his model ignored the phase information of the applied signal processing. In physiological mechanisms, the mechanical vibrations impinging on the oval window are transmitted to helicotrema through the cochlear fluid which causes the basilar membrane (BM) to vibrate at a place associated with the input acoustic wave frequency. Because adjacent places along the length of the BM vibrate in phase with the transmission of the fluid, the information of the phase as well as the amplitude of the BM displacement are both transmitted to the brain [3, 4]. A cochlear filter which includes the phase processing of the signal can be realized by a transmission line model of the cochlea [5, 6, 7, 8].

The main aim of this paper is to apply speech sounds to a cochlear filter and to show how the speech sound could be visualized through the cochlear filter. The motive of the research is to develop software tools for voice visualization.

II. Methods

The one-dimensional linear and active model of the cochlea suggested by Neely and Kim [6] is used as a cochlear filter. Their frequency model is transformed to an electrical transmission line model for time responses [9]. Fig. 1(a) shows the circuit diagram of the transmission line model. An input voltage source, V_o, represents the sound pressure onto the ear drum, and Zm represents a middle ear impedance. The longitudinal length of the cochlea is 2.5cm (in x axis) and the length is divided into 500 sections. Each sectional impedance, $Z_i(x)$, as a function of x is described in detail in fig. 1(b), and their quantitative values are derived from the Neely and Kim's model (see Appendix). Each vertical impedance section is connected by an inductor, L_i, which represents the inductance of the cochlear fluid. Each loop current, $I_i(t)$, is a variable of the cochlear model. The current

*Dept. of Control & Instrumentation Engineering, Chosun University
Manuscript Received: September 18, 1995.

Acknowledgement:
The work has been supported by a research grant (NSF) from the Regional Research Center (Factory Automation Research Centre for Parts of Vehicle) of the Korea Science and Engineering Foundation.

which flows through $Z_i(x)$, that is $(I_{i-1}(t)-I_i(t))$, represents the displacement velocity of the BM at its corresponding point. The dependent voltage, $P_i(t)$, represents the inside pressure of the outer hair cell (OHC), and it is proportional to the displacement velocity of the OHC $(I_{i-1}(t)-J_i(t))$ as follows :

$$P_i(t) = \gamma \cdot R4_i \cdot (I_{i-1}(t) - J_i(t)) \frac{\gamma}{C4_i} \int_{-\infty}^{t} (I_{i-1}(\tau) - J_i(\tau))d\tau. \qquad (1)$$

$R4_i$ and $C4_i$ represent characteristic impedances of the OHC. γ is an amplifying gain constant and the value of γ is 1.0 for the present study. If γ is increased, the model becomes unstable, while smaller γ produces less sensitivity of the BM displacement [6, 9]. The spatial tunning resolution of the BM displacement for different γ is well published by Neely and Kim [6]. Extra 350 impedance sections are added to the end of the transmission line. It is because of the echos from the helicotrema. The expanded sections improve the spatial resolution of lower freqeuncies below 100Hz.

The temporal solution of the transmission line model for the unknown variables, $I_i(t)$, is carried out using the Gaussian Elimination technique [10] for every instant time of an input signal. Computation is done on a PC with the Intel Pentium CPU (Clock speed 60MHz) using the fortran language. Speech sounds are used as input signal sources (Fig. 2). A commercially available DSP board is used for analog-to-digital conversion and for the data storage of speech sounds (Ariel DSP-96). Speech sounds are sampled at 44100 Hz through a 16 bit channel. Since the model input signal has a sampling interval of 2 sec(= dt) at the entrance of the transmission line, every speech sound has to have the same time interval. It is done by the curve fitting technique [11].

The amplitude of the BM displacement is integrally calculated along the length of the cochlea :

$$\int_{-\infty}^{t} (I_{i-1}(\tau) - I_i(\tau))d\tau \qquad (2)$$

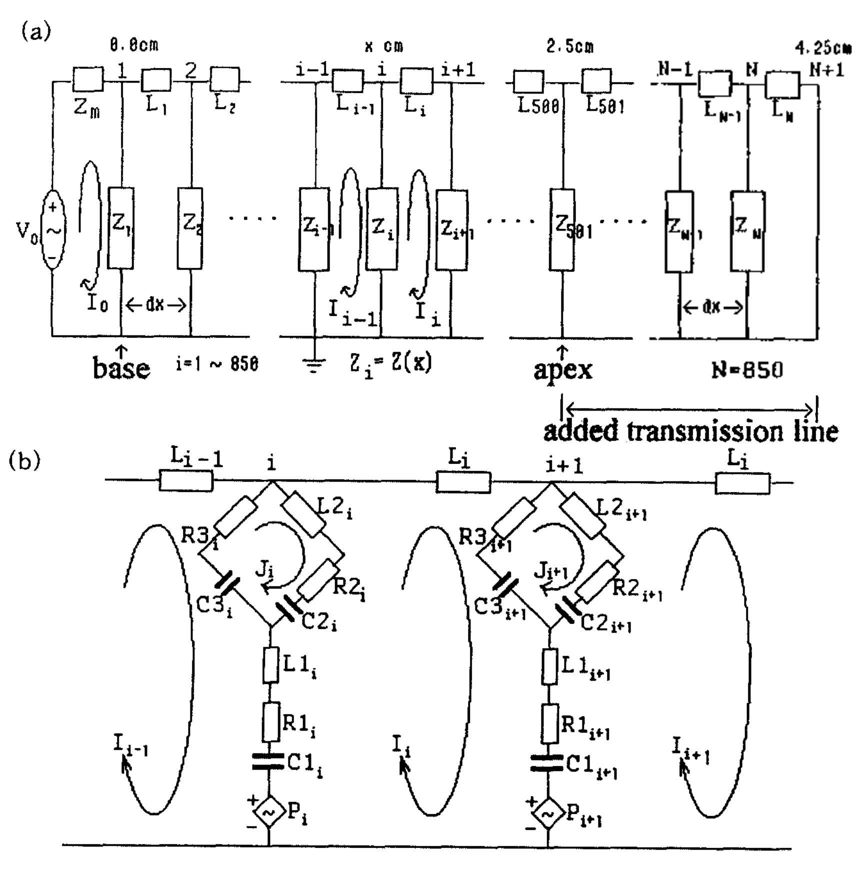

Fig 1. (a) One dimesional cochlear filter(equivalent electrical circuit model)
(b) Each section impedance, Zi, is described in detail

And the envelope of the displacement amplitude for each section is kept with maximum peaks during one cycle of the characteristic frequency (CF) for each different position along the length. Since the CF of the base (0.0cm) corresponds to about 54429 Hz and the CF of the apex (2.5cm) corresponds to about 113Hz, the following equation is derived as a number of enveloping constant as a function of x :

$$e^{2.471x} \div 54429dt \qquad (3)$$

The resulting envelope of the BM displacement amplitude of the speech sound is then saved in a disk storage for every constant time interval (12*dt). When the discrete computation of the cochlear filter model is arrived at the end of each input speech

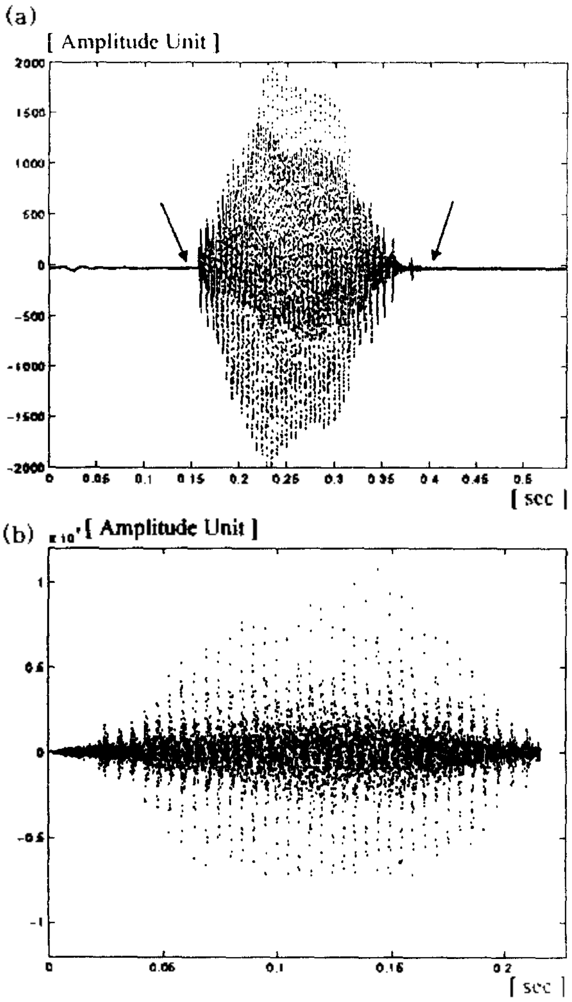

Fig 2. (a) speech sound of 'a' is digitized, and the figure shows the speech waveform
(b) Every speech signal is pre-processed before it is delivered to the cochlear filter.

sound data, the calculation finishes. Next, the complete storage data of the envelope BM displacement amplitude is displayed as a color image as follows. The size of the complete envelope data is 500 by 10000 in matrix. 500 is the number of sections from the base to the apex, and 10000 is the number of time interval from 0[sec] to (10000*12*dt) [sec]. If the input speech sound is longer than this total interval time, the column number could be further increased. The magnitude of each element value of the matrix is arranged to be put between 0 to 255 by logarithmic scaling for each value. Regarding each element value as a pixel number, the envelope matrix of the BM displacement is directly displayed as a color image. Coloring of the image is arbitrarily done by changing R, G, B, components for each pixel. For the whole image, total coloring number is fixed to 256 in a look-up table. The computation time for each speech is taken about 8 hours for the 500 by 10000 image matrix.

Ⅲ. Results and Discussion

A vowel, 'a', is initially examined. A typical waveform of 'a' is shown in fig. 2(a). The speech sound of fig. 2(a) is pre-processed before it is delivered to the cochlear filter. Firstly, the total sound data are cut leaving only for speech waveform (between two arrows). Secondly, the speech waveform is then pre-emphasized by passing through a high pass filter as follows [12] ;

$$V_n(t) = V_o(t) - V_o(t-dt), \textit{ where } V_o(-dt) = 0 \qquad (4)$$

The pre-emphasis procedure is for emphasizing the characteristics of the vocal tract. The high pass filter has 6dB/octave in its gain. Thirdly, the pre-emphasized speech waveform is differentiated for convenience, because the matrix equation which simulates the transmission line has been differentiated for the temporal solution. Lastly, the pre-processed speech waveform is masked in its front part by multiplying with an window as described :

$$0.5(1-\cos(\omega_o t)),\ 0 \le t \le \frac{\pi}{\omega_o} \qquad (5)$$

This windowing is for reducing the echoes caused by the impulse response of the cochlear filter. Fig. 2 (b) shows the speech waveform of 'a' passed through

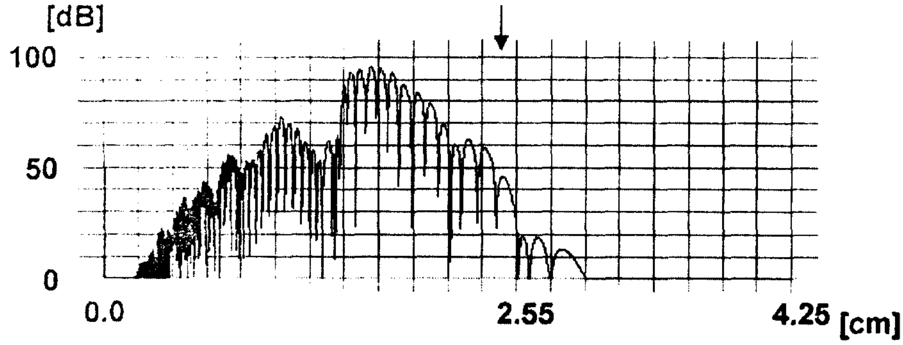

Fig 3. The amplitude of the basilar membrane displacement along the length of the BM in logarithmic scale The vertical arrow indicates the point of the apex.

the above four pre-processes.

When the speech waveform of fig. 2(b) is computationally processed by the cochlear filter as described in methods, the visualized image of the speech sound shows both the spectral and the temporal information of the speech in two-dimension. Fig. 3 shows the magnitude of the BM displacement along the length of the BM in logarithmic scale. This figure is instantly captured at 36msec after input onset. The vertical arrow indicates the position of the apex (2.5cm). Fig. 4(a) shows only a part of the pre-processed waveform between 108msec and 120msec of fig. 2(b). Fig. 4(b) shows the visualized image of the selected waveform for the exactly same time interval of fig. 4(a). In fig. 4(b), its horizontal axis represents a time interval while the vertical axis corresponds to the distance of the BM from the base to the apex (from bottom to top). The magnitude of the image pixel value is increased from cold colors to hot colors (blue<green<yellow<red). In both fig. 4(a) and fig. 4(b), the quasi-periodic pattern of the signal waveform is well matched each other.

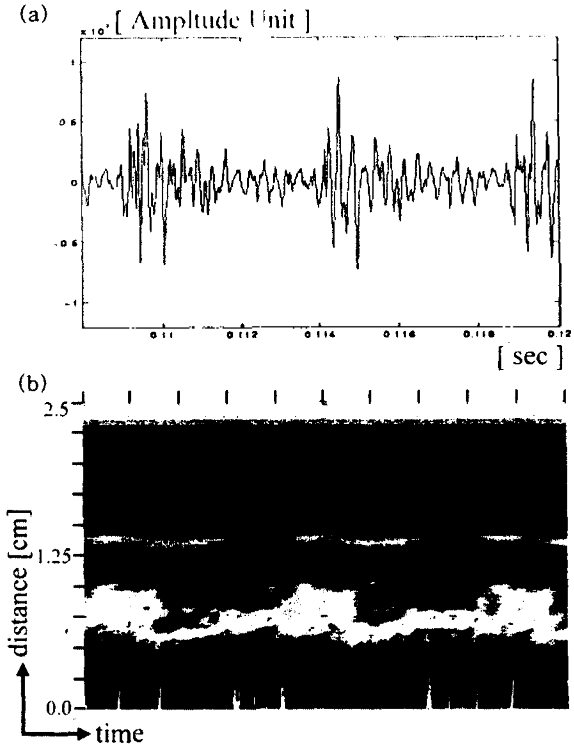

Fig 4. (a) The pre-processed waveform of the speech sound, 'a', between 108msec and 120msec of fig. 2(b) (b) The visualized image of the speech sound, 'a', at the same time interval of (a)

Fig. 5 shows the total image of the speech sound, 'a', processed by the cochlear filter. The quasi-periodic spot of the image pattern clearly appears between 0.85cm and 1.275cm (equivalent to 6663Hz and 2331Hz respectively) in vertical axis. The magnitude of the BM displacement is kept high constantly above 1.4875cm (equivalent to less than 1379Hz) in vertical axis. It results from the quasi-periodicity of the speech sound waveform. These places or their equivalent CFs in fig. 5 could be comparatively understood if the instant magnitude of the BM displacement of fig. 3 is compared with the spectrum of the same speech waveform produced by the fast fourier transformation (FFT) (fig. 6). Fig. 6(a) is the same as fig. 3 in its magnitude but the x-axis is swapped in opposite order. The left-hand side corresponds to the apex of which the CF is about 100Hz, while the right-hand side is for the base of which the CF is about 54429Hz. Two arrows of fig. 6(a) indicate the corresponding places of their CFs respectively. Fig. 6(b) shows the spectrum of the same speech sound, 'a', produced by the FFT. Both

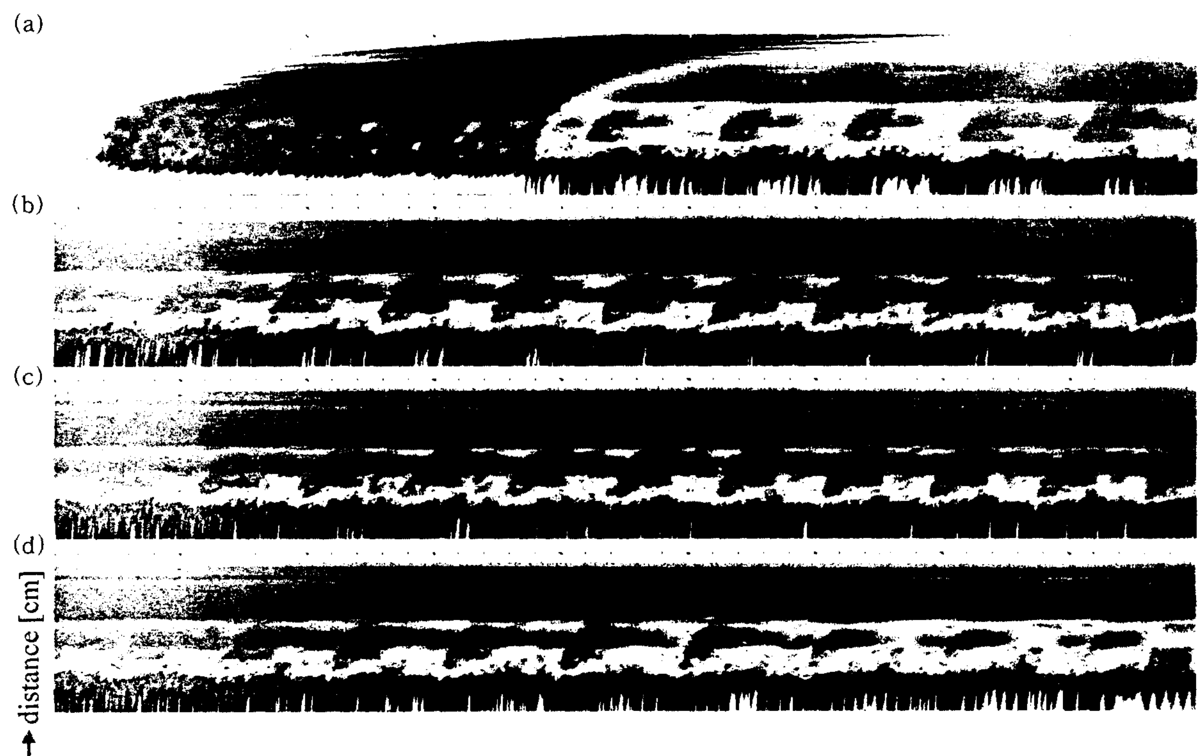

Fig 5. The total image of the speech sound, 'a', processded by the cochlear filter (a), (b), (c), (d) are all connected in series in the time axis.

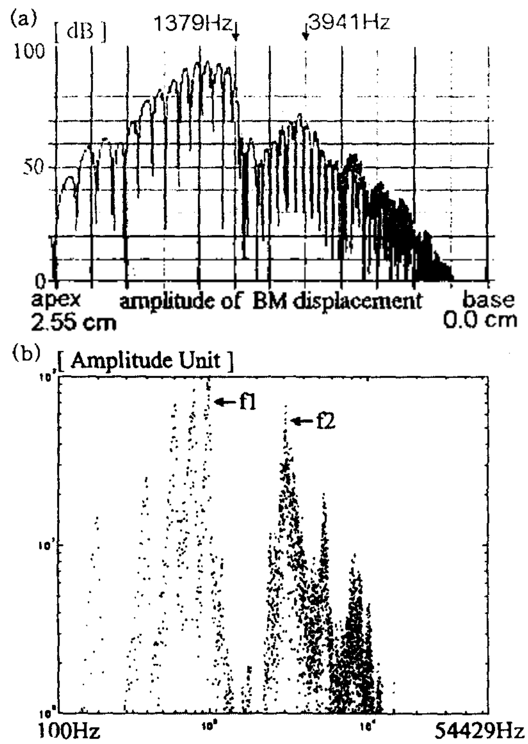

Fig 6. (a) The amplitute of the BM displacement Two arrows of (a) indicate the corresponding places of the CFs.
(b) The spectrum of the speech sound, 'a', produced by the FFT

of x and y axes are in logarithmic scale, so that the spectrum (b) can be directly compared with the displacement magnitude (a). Two significant formants are expressed by f1 and f2. f1 and f2 are 958Hz and 3038Hz respectively. Since the speech signal has these two significant formants, the magnitude of the BM displacement has also similar peaks around their corresponding CF places. The resolution of fig. 6(b) is much narrower than that of fig. 6(a). It is partly because the present cochlear filter has a moderate value of the amplifying gain. However, since fig. 6 (a) is only a temporal magnitude, the quantitative comparison between fig. 6(a) and fig. 6(b) is meaningless. The idea of extracting characteristic parameters of speech sounds in the scheme of the cochlear filter is based upon the pattern analysis of the visualized image as shown in fig. 5.

Fig. 7(a) and fig. 7(b) show speech images of two vowels : One is 'a' with medium pitch and the other is 'a' with higher pitch. One clear difference is the frequency of the red spot along the horizontal time axis, both occurring between 0.75cm and 1.75cm from the base (equivalent to 8530Hz and 721Hz in CF) in the vertical axis. The higher pitch vowel has more frequent appearance of red spots. Dotted line of

Fig. 7(c) indicates the FFT spectrum of the same medium pitch 'a' while that of Fig. 7(d) is for the higher pitch 'a'. Fig. 7(c) and Fig. 7(d) show that the FFT spectra of the two speech sounds do not show the characteristic of the quasi-periodic difference.

Other different five vowels such as 'a', 'e', 'i', 'o', 'u', have different patterns for each speech image (from fig. 8 to fig. 12 respectively). These figures are selectively chosen because of their prominent patterns throughout whole time intervals. Like fig. 7, prominent changes of the image patterns happen between 0.75cm and 1.75cm from the base (equivalent to 8530Hz and 721Hz in CF) in vertical axes. More precise analysis of the speech image requires more advanced image processing techniques.

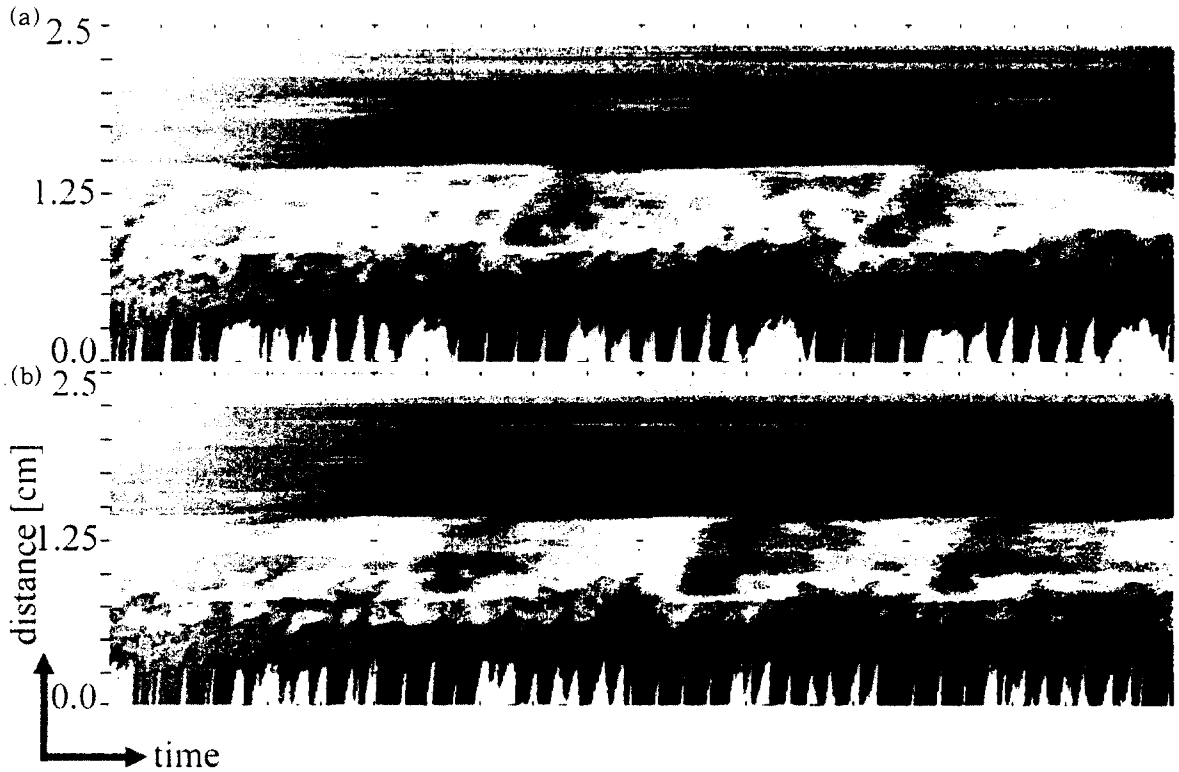

Fig 7. Visualized images of speech sounds of 'a', with medium pitch (a) and with high pitch (b)

Fig 8. Visualized image of speech sound of 'a', with medium pitch

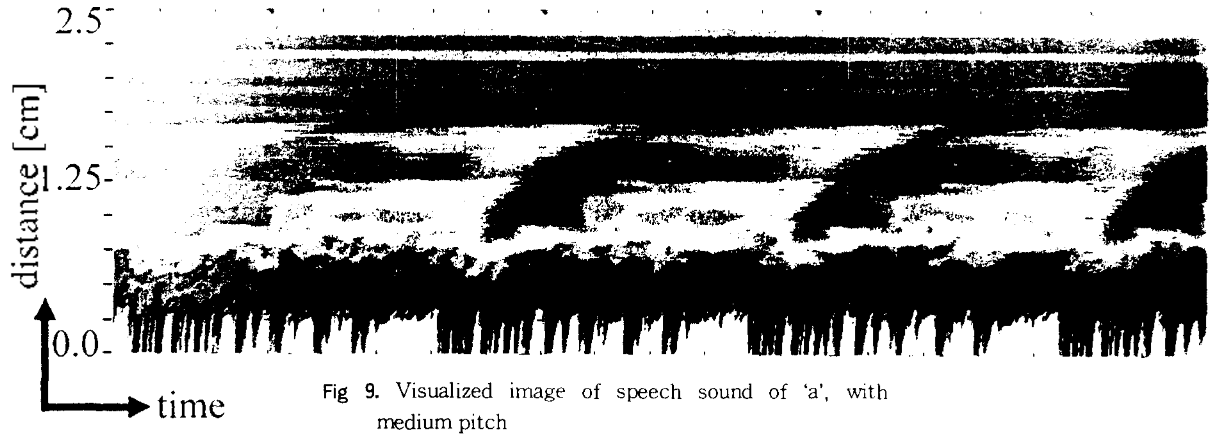

Fig 9. Visualized image of speech sound of 'a', with medium pitch

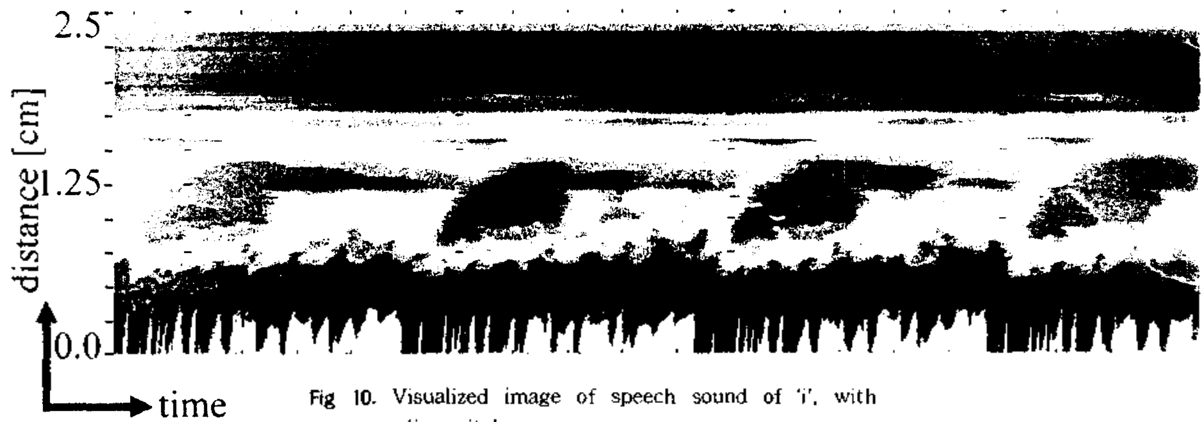

Fig 10. Visualized image of speech sound of 'i', with medium pitch

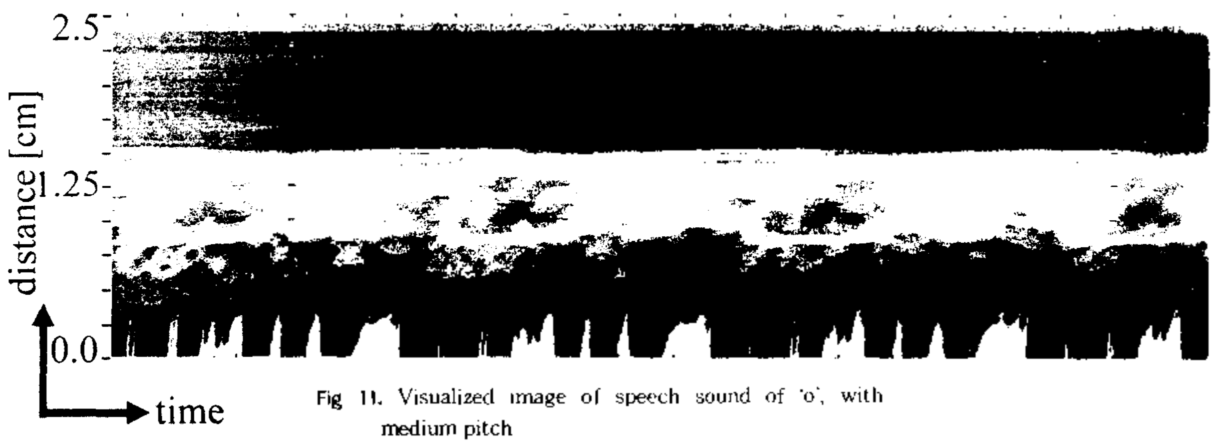

Fig 11. Visualized image of speech sound of 'o', with medium pitch

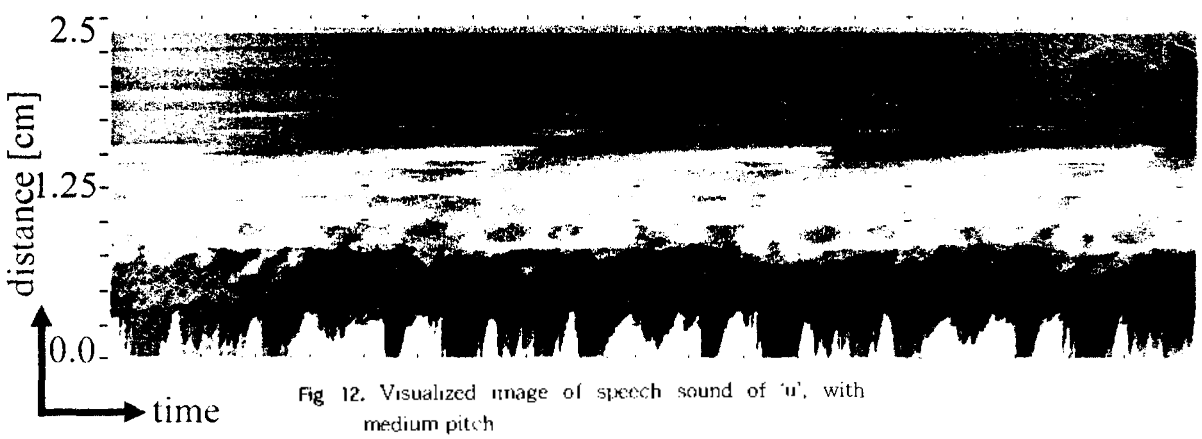

Fig 12. Visualized image of speech sound of 'u', with medium pitch

Ⅳ. Conclusion

A cochlear filter, which is simulated by an electrical transmission line, can be used to analyze speech sounds. For this purpose, visualizing processes of speech sounds are introduced using the cochlear filter. As shown in figures from fig. 8 to fig. 12, each different vowel has prominently different patterns of the speech sound image. It is important to have a similar amplitude of the input signal for each different speech sound, so as to compensate the color of the speech image. It is manually controlled by examining the signal amplitude for every speech input. Particular interest is that the characteristic parameters of the speech sound may be extracted from only a local part of the sound image. It is because the whole picture of the sound image repeats similar patterns for the total range of time interval. However, the cochlear filtering process produces the prominent patterns of the sound image after a certain amount of time, that is, after the onset of the speech sound input until the desired time of interest. More precise and qualitative analysis of the speech image pattern should be found out in the near future.

Appendix

θ = phase [cycles, radian]

f = frequency [Hz]

t = time [sec]

x = distance from the base [cm]

ω = angular frequency ($=2\Pi f$)

Pi(t) = dependant voltage [V]

Zi(x) = section impedance of the transmission line

N = total section number of the transmission line = 850

γ = active amplifying gain = 1.0

dx = distance between two adjacent sections = 5E-3 [cm]

dt = sampling time interval of discrete input signal = 2E-6 [sec]

Vo(t) = input voltage source equivalent to sound pressure onto the ear drum [V]

Rm = middle ear characteristic resistance = 400 [Ω]

Lm = middle ear characteristic inductance = 4.5E-2 [H]

Cm = middle ear characteristic capacitance = 4.76E-6 [F]

L1 = cochlear fluid inductance = 0.01 [H]

Li = cochlear fluid inductance = 5E-4 (i = 2 ~ 849) [H]

L850 = cochlear fluid inductance = 0.01 [H]

R1(x) = BM characteristic resistance = 20.0 + 1500.0 •exp(-2.0•x) [Ω]

L1(x) = BM characteristic inductance = 3.0E-3 [H]

C1(x) = BM characteristic capacitance = 0.9E-9•exp (4.0 x) [F]

R2(x) = TM characteristic resistance = 10.0•exp(-2.2•x) [Ω]

L2(x) = TM characteristic inductance = 0.5E-3•exp (x) [H]

C2(x) = TM characteristic capacitance = 1.43E-7•exp (4.4•x) [F]

R3(x) = TM-RL coupled characteristic resistance = 2.0•exp(-0.8•x) [Ω]

C3(x) = TM-RL coupled characteristic capacitance = 1.0E-7•exp(4.0•x) [F]

R4(x) = OHC characteristic resistance = 1040.0•exp (-2.0•x) [Ω]

C4(x) = OHC characteristic capacitance = 1.63E-9•exp (4.0•x) [F]

BM = basilar membrane

CF = characteristic frequency

OHC = Outer Hair Cell

RL = reticular lamina

TM = tectorial membrane

Zm = middle ear characteristic impedance

References

1. Rabiner L. R. and Schafer R. W., "Digital processing of speech signals", Prentice-Hall. Inc., PP : 396-455, 1978.
2. Ghitza O., "Auditory Nerve Representation as a Basis for Speech Processing", in Advanced in Speech Signal Processing, S. Furui and M. Sondhi, Eds., Marcel Dekker, NY, PP : 453-485, 1991.
3. Evans E. F., Wilson J. P., "Psychophysics and Physiology of Hearing", Academic Press, Lodon and NewYork, PP : 5-54, 1977.
4. Pickles J. O., "An Introduction to the Physiology of Hearing", Academic Press, London and NewYork, PP : 24-70, 1982.
5. Schroeder M. R., "An Integrable Model for the Basilar Membrane", J. Acoust. Soc. Am., Vol. 53(1), PP : 429-434, 1973.
6. Neely S. T. and Kim D. O., "A model for Active Elements in Cochlear Biomechanics", J. Acoust. Soc. Am., Vol. 79, PP : 1472-1480, 1986.
7. Zwicker E., "A Hardware Cochlear Nonlinear

Preprocessing Model with Active Feedback", J. Acoust. Soc. Am., Vol. 80(1), PP : 146-153, 1986.

8. Diependaal R. J., Viergever M. A., "Nonlinear and Active two-dimensional Cochlear models : Time domain solution", J. Acoust. Soc. Am., Vol. 85(2), PP : 803-812, 1989.
9. Jarng S. S., "Electrical transmission line modelling of the cochlear basilar membrane", Journal of the Korea Society of Medical and Biological Engineering, Vol. 14, No. 2, PP : 125-136, 1993.
10. Golub G. H. and Loan C. F. V, "Matrix computations", North Oxford Academic Ltd., PP : 52-80, 1983.
11. Shiavi R., "Introduction to applied statistical signal analysis", Aksen Associated Inc., PP : 18 : 67, 1991.
12. Kim S. L., "A study on implementation of voice recognition system for the korean continuous word", Ph. D. thesis, University of Dan-Kook, PP : 6-16, 1991.

▲Soon Suck Jarng

Member of the Acoustical Society of Korea

Date of Birth: 11th December 1961

Feb. 1984 : B.Eng. in Electronic Engineering at University of Han-Yang

Sep. 1985 : M.Eng. in Electronic Engineering at University of Hull (U.K.)

Sep. 1988 : M.Sc. in Physiology at University of Birmingham (U.K.)

Dec. 1991 : Ph.D. in Electronic & Electrical Engineering at University of Birmingham (U.K.)

From March 1991 To Present : Assistant Professor in Control & Instrumentation Dept. at University of Chosun

March 1996 : Elected as a member of the Institute of Electrical Engineering (IEE)

Main Research Field : Cochlear bio-mechanics, Digital signal processing, Underwater acoustics etc.