Bull. Fac. Bioresources, Mie Univ.
No.18 : 35 – 46
August 10, 1997

A Study on Human Voice and Other Sound in Bio-production Environment

– Possibility of Use of Voice Recognition with Agricultural Machinery –

Kunio SATO, Makoto HOKI and Katsusi AGATA
Faculty of Bioresources, Mie University

Abstract

A software system that displays spectrograms of sound signals was developed to carry out image processing. Some analysis on the difference between human voice and tractor noise and various kinds of sound signals found under bio-production environment were performed by using this system.

Image processing methods used mainly are "Fourier transformation" and "projection". As a result of the analysis, valuable characteristics of various kinds of sound signals found under bio-production environment are made clear. The possibility of identifying various sounds including tractor engine noise, animal and insect sound and human voice was considered with examination of various samples.

It was confirmed that human voices had a continuous structure that exists in quefrency domain unlike tractor engine noise. However, the same structures were observed in some sounds of animals, birds and humming of honeybees. Thus the need for further examinations of details was also pointed out to achieve the separation of human voice from other sounds for purposes of better utilization in future.

Key words : voice recognition • image processing • tractor noise • sheep bleating • crow cawing • honeybee humming

1. Introduction

In recent years, the labor-saving and automation of agriculture has been going on. Also in agricultural work using tractors, it is required to improve safety by automating functions that are controlled by an operator at the same time with the driving and steering operation. Usually only one person operates a tractor and his hands and legs are used for the driving and steering operation of the tractor. Accordingly, a specific man-machine interface needs to be developed for the operator to indicate the computer to control the other functions[1),2)].

For such a purpose, it is conceivable to control a machine through voice. However, the noise of an agricultural tractor, that exceeds 90 dB at the driver's seat, is so large that it becomes a major factor that

Accepted: March 15,1997. 1515 Kamihama, Tsu, Mie 514, Japan
@ sato. bio. mie-u.ac.jp

prevents the recognition of voice.

As for the conventional voice recognition methods, the recognition rate should be decreased theoretically when noise mixes, because they use spectrum envelope[3] of voice in the voice analysis stage. Accordingly, new technology is required with an analysis stage of voice to establish the speech recognition method that is available under noise[4].

To aim for the future development of such new technologies, a software system that analyzes characteristics of the voice under tractor noise visually is constructed[5] in this study. First of all, spectrogram as an image data is prepared from voice in this system. Secondly, systematic analysis is conducted using the image data by carrying out Fourier transform and other image processing techniques such as projection, etc.

Finally, characteristics of various sounds that are found in bio-production environment are investigated by using the software system developed. Especially, the possibility of identifying various sounds, including tractor engine noise, animal and insect sound and human voice is considered. Though researches concerning the sounds of animals were conducted from the view point of behavioral science6), researches comparing animal sounds against human voices are not general yet.

2. Methods

2.1 Scheme of the program system

Schematic diagram of the system developed is shown in Fig.1. The function of this system consists of (1) sampling of sound signal, (2) preparation of spectrogram data, (3) image processing and (4) image data presentation. First of all, a set of sound signals is sampled and a sound file is prepared. Next, a set of spectrogram data is prepared by the system. Then the needed image processing is carried out. In this stage as occasion demands, several image processings are continuously carried out. Finally, or on the way of series of image processing work, obtained sets of image data are displayed to confirm the results.

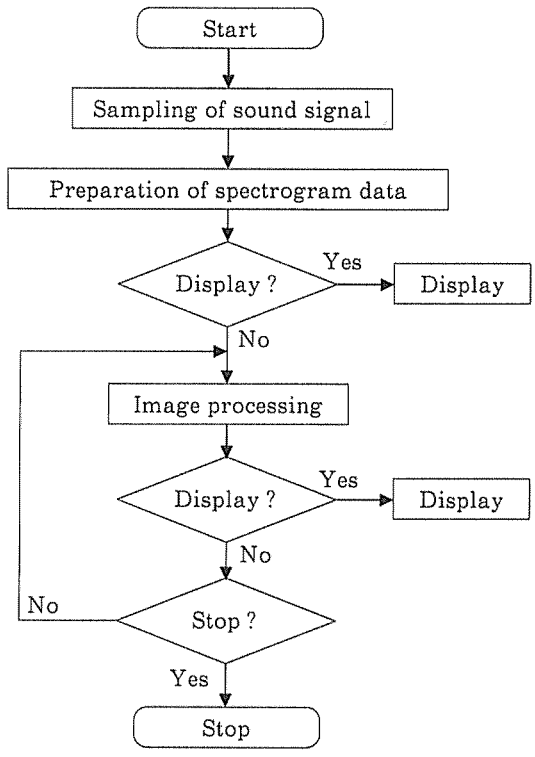

Fig. 1 Schematic diagram of the program system

2.2 Samplings of sound data

In this study, the sampling is performed using the voice input function attached with the PC. Sound signals were sampled at the rate of 44100 *Hz* on the monaural and were accumulated in the PC as sound data files.

2.3 Preparation of spectrogram data

The sound data obtained in the preceding paragraph were divided into frames with the method shown in Fig.2. One frame was composed of continuous 1024 points, and n frames were extracted from the beginning of a file with skipping the frame period q. The frame number n was 256 and the frame period q was eighth of the frame length, namely 128 points. Then short time spectrums were calculated to be the original data for a spectrogram.

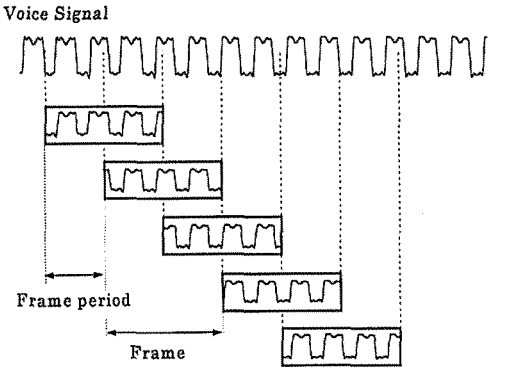

Fig. 2 Preparation of spectrogram data

Fourier transformation of consecutive signal $g(t)$ is defined as

$$G(f) = \int_{-\infty}^{\infty} g(t)e^{-2\pi ift}dt, \qquad (1)$$

where f is frequency. To adapt this transformation to N points discrete data $x(nT)$ $(n=0\sim N-1)$, the discrete Fourier transformation (DFT)

$$X(k\Omega) = \sum_{n=0}^{N-1} x(nT)e^{-j2\pi nk/N}, \qquad (2)$$

where $\Omega=2\pi/(NT)$ is angular frequency, k $(k=0\sim N\text{-}1)$ is the order of Ω and T is the sampling period, was employed with the FFT (Fast Fourier transform) algorithm. In calculation, the hamming window was used to reduce the terminal effect.

Next, the short time power spectrum of the period $Z(k\Omega)$ was calculated with

$$Z(k\Omega) = r(k\Omega)^2 + i(k\Omega)^2, \qquad (3)$$

where $r(k\Omega)$ and $i(k\Omega)$ are real and imaginary part respectively of the complex value obtained by (2).

In the DFT of the sound data for Frequency-Time spectrum, namely spectrogram data, $N=1024$ and $T=1/44100=2.2675\ 10^{-5}$ [s] were adopted. Then the logarithms of $Z(k\Omega)$ s from (3) were used as the logarithmic spectrum.

To display those data as a spectrogram image, each series of spectrum was ordered along the time axis and the gray scale (black and white) level was assigned to each spectrum data. Gray scale has the vector value (r,g,b) composed of red r, green g and blue b. To show the intensity of spectrum, its maximum M and minimum m are assigned to (255,255,255) and (0,0,0) respectively, and the intermediate intensity v $(m<v<M)$ is represented as

$$(n,n,n), \qquad (4)$$

where

$$q = \{255/(M\text{-}m)\}(v\text{-}m), \qquad (5)$$

$n=[q]$ ([ ] is the gaussian symbol).

The total number of the colors that can be displayed with this method are 256 stages.

2.4 Image Processing

The following image processing methods were used in this study, to extract further characteristics from obtained spectrograms.

2.4.1 Fourier transformation

Though the Fourier transformation used is the DFT expressed by (2), $x(nT)$ is not necessarily time series data since the objective data is image. For instance, $x(nT)$ means the time series data when the DFT is applied along the time axis, but it means the numerical series data that have the independent variable of frequency when the DFT is applied along the frequency axis. The data transformed with the DFT applied along the frequency axis is called cepstrum that has the independent variable of quefrency, whose dimension is time. The actual meaning of this transformation will be explained later.

Concerning the calculation of the Quefrency-Time spectrum, namely scan power cepstrum, that can be obtained by DFT applied along the frequency axis of above-mentioned spectrogram, $N=256$ and $T=44100/1024=43.0665$ $[1/s]$ were used. For the Frequency-Frequency spectrum, that can be obtained by DFT applying along the time axis of spectrogram, $N=256$ and $T=1024/44100/8=0.002902$ $[s]$ were used.

Though the two dimensional Fourier transformation was also applied for the image processing, no significant characteristics could be found in each case.

2.4.2 Projection

The projection data concerning the x axis or y axis can be obtained through adding all data along y axis or x axis of displayed two dimensional image respectively.

2.5 Displaying Image

Images were displayed after the processing explained in 2.3 or 2.4. Displayed image can be classified into three categories as

(1) Frequency-Time spectrum (spectrogram)
y-axis (256 *pixels*) : frequency, 0 to 11.025 *kHz*
x-axis (256 *pixels*) : time, 0 to 0.74304 *s*

(2) Quefrency-Time spectrum (scan power cepstrum)
y-axis (128 *pixels*) : quefrency, 0 to 11.60998 *ms*
x-axis (256 *pixels*) : time, 0 to 0.74304 *s*

(3) Frequency-Frequency spectrum
y-axis (256 *pixels*) : frequency, 0 to 11.025 *kHz*
x-axis (128 *pixels*) : frequency, 0 to 172.266 *Hz*

The drawing software for those images is written in FORTRAN and it displays the image online.

3. Human voice under tractor noise

When we use human voice as a man-machine interface for the control of a tractor, the biggest obstacle is the engine noise. Whether or not the presence of a cabin, the sound pressure level of engine noise frequently exceeds 90 *dB*. Therefore, the investigations are carried out using the image processing of human voice and engine noise to grasp the characteristics of respective sound.

3.1 Human voice and cepstrum

Fig.3 shows a spectrogram of a male voice 'shita'. In the area of a consonant of the spectrogram,

unforeseen signal that extends to high frequency domain along the frequency axis emerges, as shown in the figure. On the other hand, in the area of a vowel of the spectrogram, some striped pattern called the pitch emerges along the time axis. The interval of this stripe pattern is called the fundamental frequency that can be conceived as the fundamental period of the source of a human voice, that is to say, the pitch period. An image of the scan power cepstrum can be obtained by Fourier transformation of the spectrogram along the frequency axis. The definition of the power cepstrum is written as

$$C(\tau) = |F\{\log S(t)\}|^2, \quad (6)$$

where τ is the quefrency explained later. The symbol F represents the Fourier transformation expressed as (1) and S is the power spectrum of time function $g(t)$ as

$$S(f) = |F\{g(t)\}|^2. \quad (7)$$

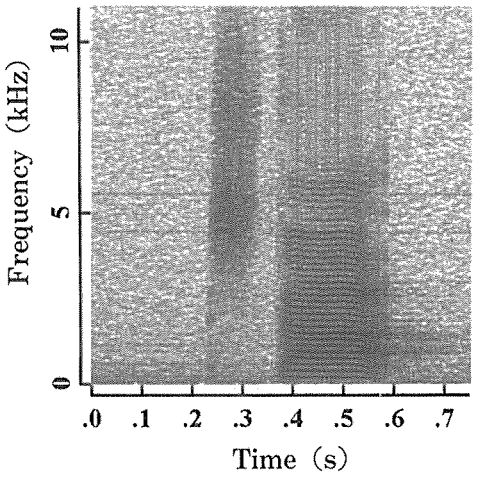

Fig. 3 Spectrogram of 'shita'

Fig.4 shows the scan power cepstrum of Fig.3. In this figure, the stripe pattern of Fig.3 emerged as one line. The y-axis of the scan power cepstrum is called quefrency, whose dimension is time *s*. The line emerged in Fig.4 corresponds to the quefrency of 5 *ms*, that can be conceived as the period of the fundamental pitch of the vowel. Besides the dominant line that shows the pitch period, the second pitch period is emerged in Fig.4 at the two times longer quefrency even though it is unclear. This is caused by the clear stripe pattern even in the high frequency domain in Fig.3. Then the periodical element of the pitches, that have the two times or more higher frequency than the fundamental one, can be detected. The clear feature of those harmonic elements of quefrency is that they are located at just the position of integer multiplication of the fundamental pitch.

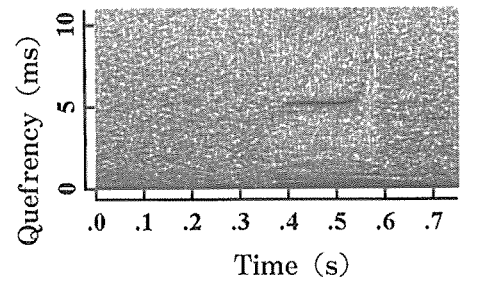

Fig. 4 Scan power cepstrum of 'shita'

In the Japanese pronunciation of 'shita', when it is pronounced slowly, it includes the vowel 'i', although when it is pronounced quickly it does not include any vowel 'i' as shown in Fig.3.

3.2 Tractor engine noise

Fig.5 shows the engine noise of a non-moving tractor with 6 cylinders 90 *Hp* engine at the revolution rate of 2000 *rpm*. In this case the microphone was set in front of the operator's mouth, since the absolute sound pressure level of the noise and voice at this point is needed. The overall sound pressure level of this case was 95.3 *dB* and in this condition the normal communication between two close persons was almost

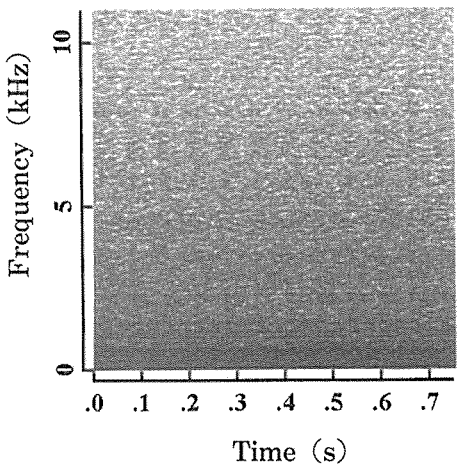

Fig. 5 Spectrogram of a tractor engine noise

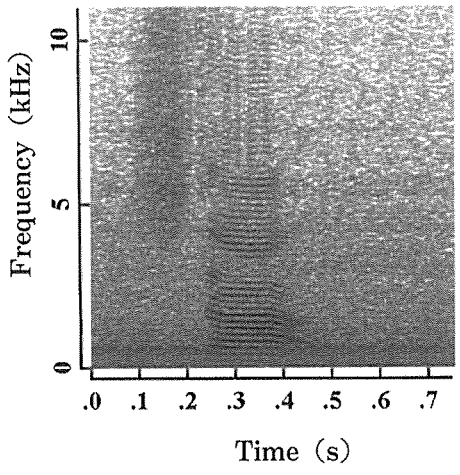

Fig. 6 Spectrogram of 'shita' under tractor engine noise

impossible.

3.3 Human voice under engine noise

Fig.6 shows the human voice 'shita' under tractor engine noise. Compared with Fig.3, in the low frequency part of consonant 'sh' and 't', the feature was almost hidden behind the engine noise. Contrasted with this, vowel 'a' keeps its feature well like Fig.3.

Those phenomena are caused by the fact that the engine noise ranges vastly from low frequency to high with comparably gradual decline of power. On the other hand the vowel pitches concentrate at such a narrow area that they are not seriously affected by the noise.

3.4 Discussion

With such comparisons, it is predicted that the vowel pitches could be a clue to voice recognition under tractor engine noise. As shown in Fig.6, although the consonant feature is hidden behind the engine noise below about 4 *kHz*, its element extended to higher frequency domain would hardly be affected by the noise.

According to those investigations, the method, that detects the voice existing part using a vowel pitch first then searches consonant parts back and forth, is conceivable as an effective method.

4. Sound in the bio-production environment

When a tractor is controlled with voices, various sounds besides the engine noise are supposed to be mixed in the sound signals. To evaluate the availability of voice control of tractors, the various as well as more realistic considerations are required.

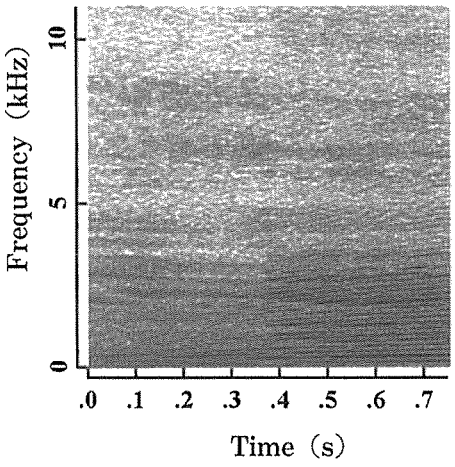

Fig. 7 Spectrogram of lowing of a cow

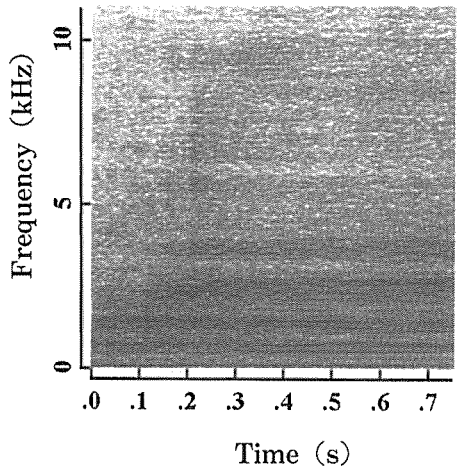

Fig. 9 Spectrogram of bleating of a sheep (case 1)

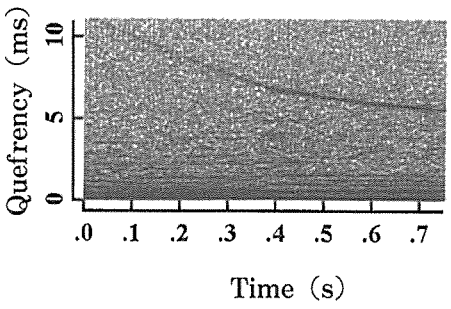

Fig. 8 Scan power cepstrum of lowing of a cow

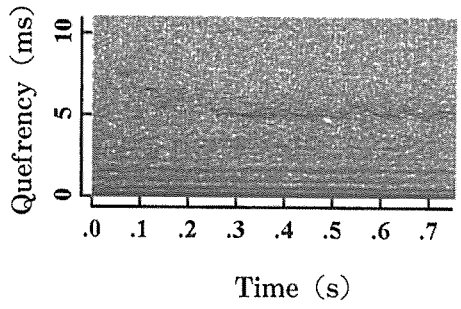

Fig. 10 Scan power cepstrum of bleating of a sheep (case 1)

In this chapter the further study is carried out on the possibilities of the presence of the pitches, that distinguish human voice from tractor engine noise, in the bio-production environment.

4.1 Sound of livestock

A cow (Japanese Black, *Bos tauras*) and a sheep (Corridale, *Ovis aries*) are investigated as examples of livestock. Fig.7 shows a spectrogram of the general lowing of a cow, and Fig.8 shows its scan power cepstrum. From those figures, we can see the same kind of pitch element as human vowel in the lowing of a cow. In Fig.7, it is observed that during early 0.35 seconds the characteristic of the voice pass is different from that of the later part. In Fig.8, the pitch element last comparatively long and its period changes gradually.

Fig.9 and 10 show a spectrogram and a scan power cepstrum of the bleating of a sheep (case 1)

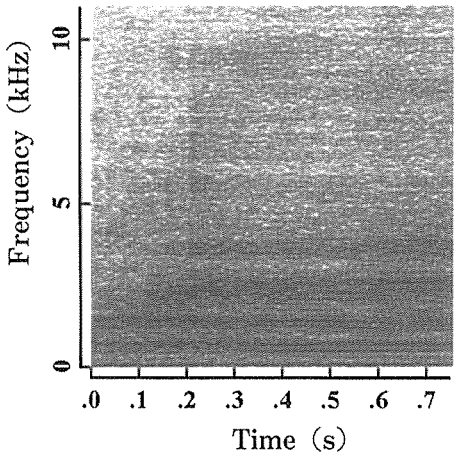

Fig. 11 Spectrogram of bleating of a sheep (case 2)

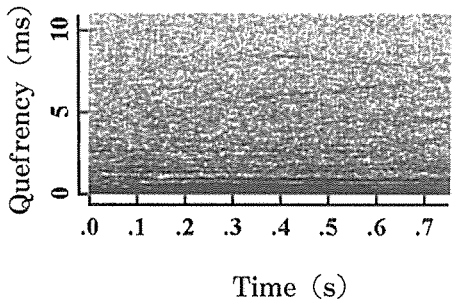

Fig. 12 Scan power cepstrum of bleating of a sheep (case 2)

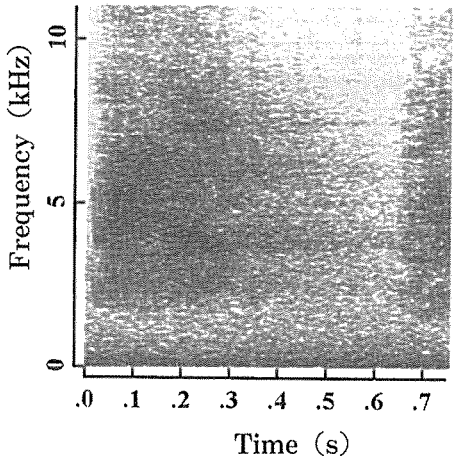

Fig. 13 Spectrogram of chirping of a single bulbul

respectively. Though this sound has the fundamental pitch of about 0.5 *ms*, it also has the predominant quefrency element at about 5 *ms* and further at two times longer quefrency. Especially, the quefrency at about 5 *ms* has a very complicated structure. The relationship between the quefrency element of 0.5 *ms* and 5 *ms*, however, needs more detailed considerations to be concluded whether or not they have the relationship of integer multiplication.

Next, Fig.11 and 12 show a spectrogram and a scan power cepstrum of the bleating of a sheep (case 2) respectively. In Fig.12, at the quefrency of about 8 *ms* and 6 *ms*, there emerge two predominant quefrency elements. Further careful observation leads to another intermittently continuous quefrency element at about 4 *ms*. What is significant in this figure is the obvious absence of the relationship of integer multiplication between the element of 8 *ms* and 6 *ms*. If this phenomenon is explained simply, there are two or more sound sources in this sheep. Considering Fig.11, we can see a very thick stripe in the frequency elements of the spectrogram and it varies gradually along the time axis. The actual mechanism of this phenomenon requires further biological and anatomical investigations to be solved.

4.2 Sound of birds

Fig.13 and 14 show spectrograms of chirping of a single Bulbul (*Hysipetes amaaurotis*) and a flock of Bulbul. The major feature of those images is that in neither case of a single bird or a flock, there is

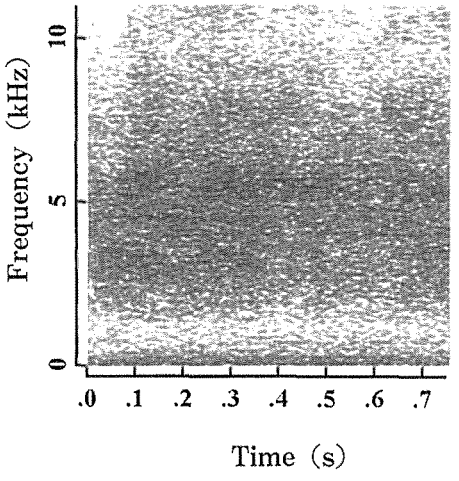

Fig. 14 Spectrogram of chirping of a flock of bulbul

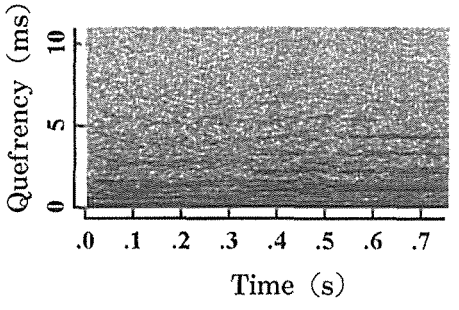

Fig. 15 Spectrogram of crowing of a bantam

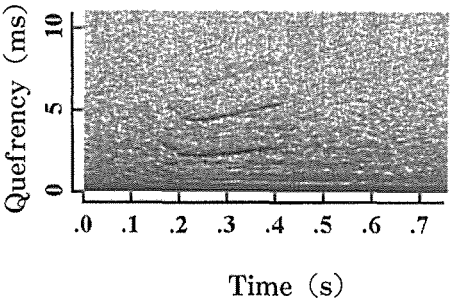

Fig. 16 Scan power cepstrum of cawing of a crow (case 1)

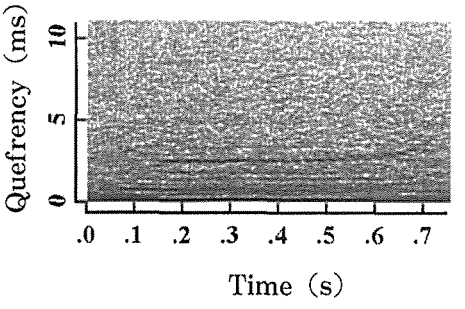

Fig. 17 Scan power cepstrum of cawing of a crow (case 2)

seldom much volume in the area below 2 *kHz*, where usually high volume is projected in cases of general noise and sounds of animals. When the figures are compared with Fig.3, it can be said that the chirping of this bird has power at the same area as sound 'sh' of human. In the scan power cepstrums of those cases, we could not find any remarkable structure, as for instance pitches.

Fig.15 shows the crowing of a bantam (Japanese bantam, *Gallus gallus*). The fundamental pitch period was about 1.1 ms and it has the feature of keeping an almost constant time period during the crowing.

Fig.16 and 17 show the scan power cepstrum of cawing of a crow (*Corvus macrorhynchos*). This time we could record two types of typical cawings. In the first one, as shown in Fig.16, the pitch element continues smoothly like human vowels. In the other one, as shown in Fig.17, the pitch element has the periodical continuation. The Fourier transformation along the time axis was carried out with Fig.17 to investigate the periodic features and a Frequency-Frequency image was displayed as Fig.18. According to this image, some spectrum can be observed at about 40 *Hz* of x-axis. To make it clear, the projection of Fig.18 was calculated along the x-axis shown as Fig.19. In this figure, the small but clear peak can be seen at about 40 *Hz*. Thus the feature of cawing that has low frequency periodical continuation was

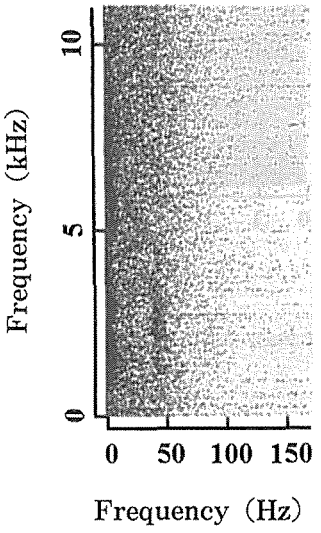

Fig. 18 Frequency-Frequency diagram of cawing of a crow (case 2)

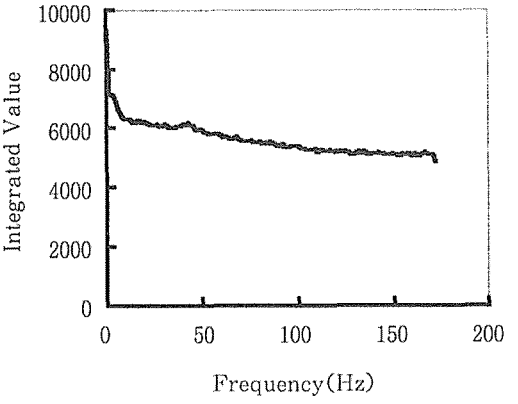

Fig. 19 Projection of Fig. 18

confirmed quantitatively using Fig.18. From this figure, the interesting suggestion was obtained, that the crow transfer such low frequency, that is hard to be transferred, with a carrier of between 2 and 4 *kHz*. The availability of the Frequency-Frequency image was also certified.

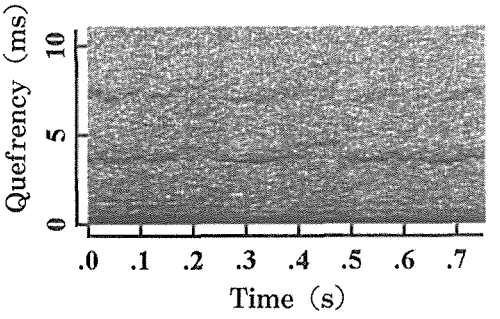

Fig. 20 Scan power cepstrum of humming of a honeybee (*Apis mellifera*)

4.3 Humming of honeybee

Voice control system used in the bio-production environment may have the mix of humming of small insects. Fig.20 and 21 shows the scan power cepstrums of humming of two species of honeybee, European honeybee (*Apis mellifera*) and Japanese honeybee (*Apis cerana japonica*) respectively. What is interesting is that, there are comparatively short quefrency elements between the long elements in both cases. The short and long quefrency elements seem to be harmonic sounds produced by the same couple of wings but with different movement.

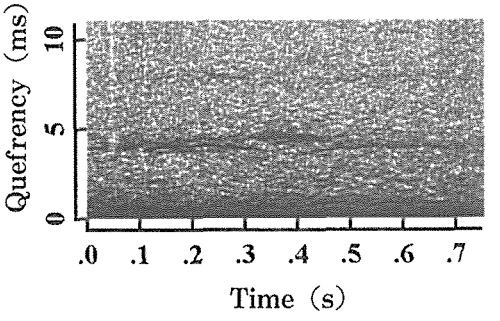

Fig. 21 Scan power cepstrum of humming of a honeybee (*Apis cerana japonica*)

Though the quefrency elements of Fig.20 have the form of convex downward, those of Fig.21 have the form of convex upward. Whether these differences depend on individuals or species needs to be investigated further with more cases.

4.4 Discussion

Concerning the typical sound signals in the bio-production environment, several samples were investigated and valuable data could be obtained. Though further biological and anatomical studies are needed for the future investigations, the following conclusions can be listed concerning our purposes.

(1) In bio-production environment, it is unable to recognize human voice correctly just analyzing pitch elements.

(2) To recognize human voice correctly, one or all of following counter plots should be taken.

(a) More precise information concerning the intensity of pitch elements should be used for recognition. This process is needed to recognize the category of human vowels.

(b) Position information of the sound source should be taken to narrow the candidates down. With this process the sound source can be identified more correctly.

(c) Various kinds of sound of animals and noises should be recognized positively to distinguish human voices clearly.

5. Conclusions

A software system that displays spectrograms of sound signals was developed to carry out image processing. Some analysis on the difference between human voice and tractor noise and various kinds of sound signals found under bio-production environment were performed by using this system. Image processing methods using "Fourier transformation", "projection" etc. revealed to be valuable in our purposes. Moreover, several valuable phenomena were observed in some creatures including sheep and honeybees using our system.

As the results of these considerations, the following conclusions were obtained.

(1) Human voices have a continuous structure that exists in quefrency domain unlike tractor engine noise. It seems to be possible to use this feature to apply for voice recognition even under big noises.

(2) However, the same structures were observed in some sounds of animals, birds and humming of honeybees. Therefore to achieve the separation of human voice from other sounds, further measures are proposed to be taken as follows ;

(a) More precise information concerning the intensity of pitch element should be used for recognition. This process is needed to recognize the category of human vowels.

(b) Position information of the sound source should be taken to narrow the candidates down. With this process the sound source can be identified more correctly.

(c) Various kinds of sound of animals and noises should be recognized positively to distinguish human voices clearly.

Acknowledgment

We are grateful to Mr. H. IMUKAI, who was a strudent in 1994, and Mr. H. HATAE who is a graduate student of Mie University for their help in construction of the software system. We also would like to thank Dr. Y. KOBAYASHI, Dr. M. MATSUURA and Mr. Y. MIYAZAKI for their help in sampling of animal and

honeybee sounds. Especially, Dr. KOBAYASHI gave us authentic suggestions in animal sounds.

References

1) SATO K., M. HOKI and A. MATSUDA. Word Recognition under Tractor Noise. Proc. AAAE. pp.375−381 (1992).
2) SATO K., M. HOKI and V. M. SALOKHE. Voice Recognition by Neural Network under Tractor Noise. Trans. ASAE. Vol.36 (4), pp.1223-1227 (1993).
3) I MAI S. Voice Recognition. Kyoritu Syuppan, Tokyo (1995). (in Japanese)
4) SATO K., M. HOKI and H. HATAE. Analysis of Tractor Noise with Voice Visualization. Research Report JSAM Kansai Branch, Vol. 80, pp.15−16 (1996). (in Japanese)
5) AGATA K., K. SATO and M. HOKI. Voice and Tractor Noise Interaction. Proc. The Tri-University International Joint Seminar and Symposium 1996. pp. 199-204 (1996).
6) IKEDA Y. and Y. ISHII. Identification of Individuals with Sounds of Livestock. Research Report JSAM Kansai Branch, Vol. 79, pp.53-56 (1996). (in Japanese)

人音声と生物生産環境音に関する考察

−農業機械への音声認識技術の適応可能性について−

佐藤邦夫, 法貴　誠, 縣　克司

三重大学生物資源学部

音声信号をスペクトログラムにより表示し, 画像処理を行うソフトウェアシステムを開発した。これを用い, トラクタエンジン騒音と人音声の相異, 生物生産環境下で考えられる各種音声信号の解析を行った。ここで用いた画像処理手法は「フーリエ変換」,「投影」などである。解析の結果, 生物生産環境下で予想される各種音声信号の特性の一端が明らかとなった。特に, トラクタエンジン騒音および動物や昆虫の発生する音と人音声の区別が可能かどうかについて, 種々の例を計測し検討した。

その結果, 人音声はトラクタエンジン騒音と異なり, ケフレンシ領域に継続する構造を持つことが確認された。しかし動物や鳥類の鳴き声のほか, 蜂の羽音にも類似する構造があることが分かり, それらと人音声の分離のためにはさらに詳細を検討する必要性が指摘された。