VOCAL TRACT SHAPE ESTIMATION

M. Bilginer Gülmezoğlu[1] and Atalay Barkana[2]

[1,2] Department of Electrical and Electronics Engineering, Anadolu University, Eskişehir.

Abstract

The previous works done by other researchers to determine the cross_sectional areas of the vocal tract assumes that the glottal wave shape consists of only an impulse for each vibration of the vocal cords.

At the beginning of this work, it was expected that if the actual sawtooth input signal is taken as the glottal wave input, obtained cross_sectional areas of the vocal tract can be closer to actual ones. Input signals from the glottis and the output signals from the lips are measured and digitized simultaneously with the sampling frequency of 10kHz.

In this work, the transfer function of the vocal tract is obtained by using least_squares technique. The coefficients of the denominator polynomial of the transfer function are used to determine the cross_sectional areas of the vocal tract.

At first, the previous works which assume an impulse wave shape for the input is also repeated to find the cross_sectional areas of the vocal tract. It is observed that the obtained areas are similar to previous ones.

In the second step, the cross_sectional areas obtained by taking the actual input signal as the glottal wave input. The results of this work are compared with the work which takes the impulse signal as the glottal wave input.

The work with actual input signal seems to yield the position of the constriction point of the tongue better than the results of the previous work.

1.THE LEAST_SQUARES TECHNIQUE

The identification of systems with constant parameters which form the parameter vector is related to the measurements by linear matrix relation

$$\mathbf{y} = \mathbf{A}\mathbf{a} + \mathbf{v} \tag{1}$$

where $\mathbf{y}$ is an $N \times 1$ vector of measurements, $\mathbf{a}$ is a $p \times 1$ parameter vector to be estimated, $\mathbf{v}$ is an $N \times 1$ vector of noises or errors in the data taken, and $\mathbf{A}$ is an $N \times p$ matrix of data to be transformed by the model. Minimization of the sum of the squares of the errors yields[1],

$$\hat{\mathbf{a}} = (\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T\mathbf{y} \tag{2}$$

where $\hat{\mathbf{a}}$ is the estimate of $\mathbf{a}$. The matrix $\mathbf{A}^T\mathbf{A}$ is $p \times p$ and must be inverted. If $p > N$, then the rank of $\mathbf{A}^T\mathbf{A}$ is less than p and it will be singular hence not invertible.

2. VOCAL TRACT AREA FUNCTION

The discrete_time transfer function of the vocal tract is known [2] as

$$G(z) = \frac{Y(z)}{U(z)} = \frac{1}{1+\alpha_1 z^{-1}+\ldots+\alpha_n z^{-n}} \tag{3}$$

where $Y(z)$ is the discrete_time output signal, $U(z)$ is the discrete_time input signal, n is the order of the denominator and α_i's are coefficients of denominator polynomial or linear prediction coefficients (LPC's).

The equivalent difference equation of the transfer function is

$$y(k) = -\alpha_1 y(k-1) - \alpha_2 y(k-2) - \ldots - \alpha_n y(k-n) + u(k) \tag{4}$$

$y(k)$ and $u(k)$ are assumed zero for negative indices k, so that the difference equation of (4) in matrix form is

$$\begin{pmatrix} y(1) \\ y(2) \\ \vdots \\ \\ \\ \vdots \\ y(N) \end{pmatrix} = \begin{pmatrix} u(1) & -y(0) & 0 & \ldots & 0 \\ u(2) & -y(1) & -y(0) & \ldots & 0 \\ \vdots & \vdots & & & \vdots \\ & & & & 0 \\ & & & & -y(0) \\ \vdots & \vdots & & & \vdots \\ u(N) & -y(N-1) & -y(N-2) & \ldots & -y(N-n) \end{pmatrix} \begin{pmatrix} 1 \\ \alpha_1 \\ \vdots \\ \\ \\ \vdots \\ \alpha_n \end{pmatrix} + \begin{pmatrix} v(1) \\ v(2) \\ \vdots \\ \\ \\ \vdots \\ v(N) \end{pmatrix} \tag{5}$$

The expression (2) is applied to (5) to get the best α_i's

Reflection coefficients (k_m) can be obtained from the LPC's. Thus the computational expression [2] is

$$a_{m-1,i} = \frac{a_{m,i} - k_m a_{m,m-i}}{1-k_m^2} \tag{6}$$

with $k_m = a_{mm}$ for $m = M, M-1, \ldots, 1$ and $i = 0, 1, \ldots, m-1$ and $|k_m| < 1$. M is the number of sections from the lips.

The discrete area function of the estimated vocal tract shape are then computed from the reflection coefficients [2] as

$$A_{m-1} = \frac{1+k_m}{1-k_m} A_m \tag{7}$$

for $m = M, M-1, \ldots, 1$, keeping in mind that A_M is an artificial area. Having no absolute reference value, A_M is usually assumed to be unity.

The relation between sampling frequency (f_s), number of sections (M), length of the acoustic tube (L=Ml; l is section length) and speed of sound (c) is given [2] as

$$f_s = \frac{Mc}{2L} \tag{8}$$

The values of f_s and c are constant as 10kHz and 350m/sec respectively. Since L can be chosen as 17 cm, M is approximately taken as 10 for this work.

3. RESULTS AND COMPARISON

In this work, two periods of the glottal and speech signals are chosen as the interval of analysis. Preemphasis by a filter $1 - z^{-1}$ is applied to the speech wave output and Hamming window is applied to both glottal wave input and speech wave output. In the first step, unit impulse is taken as the glottal wave input. The cross_sectional areas of the vocal tract are determined for eight Turkish vowels. When the results of five Turkish vowels /a/, /e/, /i/, /u/ and /o/ are compared with those of Ishizaka Flanagan [3] model and those of Sadaoki Furui's [4] works, it is seen that the obtained results are globally similar to Ishizaka Flanagan and Furui's results. Differences at some points must be expected because Ishizaka Flanagan's results are obtained for phonetics from Russian language and Furui's results are obtained for phonetics from Japanese language. Since the areas vary in the case of strong voice, the results are normalized.

In the second step, measured and digitized values at the glottal wave from piezo crystal are used as an input in expression (5). It must be noted that the location of piezo crystal on the vocal cords is important in measuring the glottal wave shape. The cross_sectional areas of the vocal tract are also determined for eight Turkish vowels. From Figure 1, it can be seen that the obtained results are a lot closer to X_ray data [5]. Especially, the position of the constriction point of the tongue is more significant compared with the case of impulse input.

4. REFERENCES

1. Jacquot, R. G., Modern Digital Control Systems (Marcel Dekker Inc., New York and basel, 1981)
2. Markel, J.D. and Gray, A.H., Linear Prediction of Speech, (Springer Verlag, New York,1976).
3. Ishizaka, K. and Flanagan, J. L., Synthesis of Voiced Sounds from a Two-mass Model of the Vocal Cords, in Bell System Tech. J. 51, 1233-1286 (1972).
4. Furui, S. Digital Speech Processing, Synthesis and Recognition, (Marcel Dekker, Inc. New York and Basel, 1989).
5. Flanagan, J. L., Speech Analysis, Synthesis and Perception, (Springer Verlag, New York, 1972).

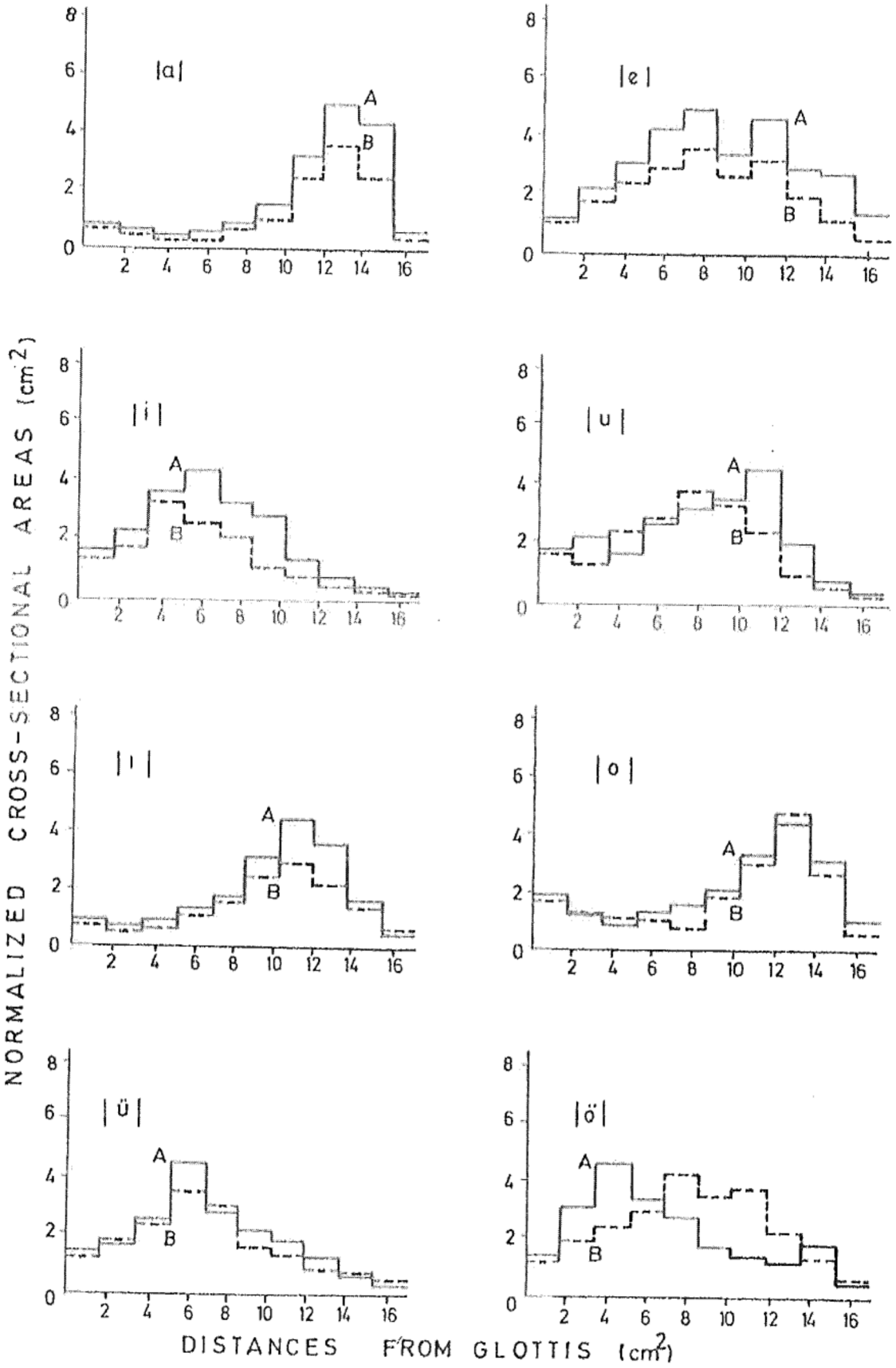

Figure 1. Estimated area functions: (A) impulse glottal wave; (B) actual glottal wave

COMPARISON OF HUMAN AND LOVEBIRD SPEECH

Osman Parlaktuna[1] and Atalay Barkana[2]

[1,2] Electrical and Electronics Engineering Department, Anadolu University, Eskişehir.

Abstract

It is well known that parrots and lovebirds imitate the human words if they are trained well. The speech production systems of parrots and lovebirds are different from humans'. For example, the length of the vocal tract of a human is almost equal to the whole length of a lovebird. Therefore, it is expected that the time signal of word produced by a lovebird would be different from the time signal of the same word produced by a human. Although, the time signals for the same word produced by a human and a lovebird are different, we still could understand the word produced by a lovebird. It is important to determine the parameters of lovebird speech which cause us to understand the words spoken by a lovebird. Knowing those parameters will help us in the field of speech recognition. In this study, the speech produced by a lovebird and by its trainer is analyzed in the time and frequency domain and those properties of the lovebirds' speech which are similar with the humans' are discovered.

INTRODUCTION

In this study the speech produced by a lovebird trained by a teenage girl is recorded. Name of the lovebird is "Dilosh" and this word is analyzed in the time and frequency domain, and compared with the same word spoken by the trainer.

The vocal tracts of the humans and lovebirds are very different from each other. The length of the vocal tract of a human is almost equal to the whole length of a lovebird. Singing birds have pepulus which vibrates as the human vocal cords. A vibrating trachea and a beak replaces the resonant frequency vibrating cavities of a human vocal tract. The schematic diagram of the vocal tract of a lovebird is shown in Figure 1 [1].

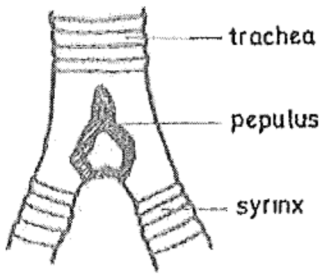

Figure 1. Schematic Diagram of the Vocal Tract of a Lovebird.

Time Analysis of the Word "Dilosh"

Since the speech production systems of a human and a lovebird are very different, it is expected that the word "Dilosh" produced by the lovebird and its trainer would be very different. This can be seen from Figures 2 and 3 which are the parts of the word "Dilosh" produced by human and lovebird, respectively. The periodic structure of the phonemes "d", "i", "l" and "o" of the human speech is not so obvious in the lovebird speech. The word produced by the lovebird has a shorter duration than the human word duration. But, average durations of each phoneme as the percentage of the whole word of lovebird and human speech are very close to each other as shown in Table 1 [2].

Table 1. The Average Durations of Phonemes

Phoneme	Human	Lovebird
D	0.050	0.055
I	0.250	0.210
L	0.076	0.060
O	0.320	0.250
Sh	0.270	0.345

Careful examination of "i" and "o" vowels of lovebird speech reveals that lovebird tries to imitate the pitch frequency of the human speech. The pitch frequency of the trainer is 350Hz on the average. Energy envelopes of the lovebird speech is almost periodic with the frequencies 280Hz and 350Hz for "i" and "o" vowels, respectively. The high frequency components are very effective in the lovebird speech which does not exist in the human speech.

Frequency Analysis of the Vowels of "Dilosh"

The 3-D spectrograms of the word "Dilosh" for the lovebird and its trainer are shown in Figures 4 and 5. In the bird speech, the high frequency components above 1500Hz seem to carry the whole information. The peak frequencies of the vowel "o" of the lovebird are between 1600Hz and 3500Hz. The formant frequencies of the vowel "o" of the trainer are 614, 1228 and 1848Hz. Although the lovebird speech seems to be the shifted version of human speech to the high frequencies, there is no one to one correspondence between high energy frequencies.

Modulation of the Vowels "i" and "o"

From the time and frequency domain graphics it seems that "i" and "o" vowels are modulated as DSBSC by the lovebird with carrier frequencies 2400Hz and 1600Hz, respectively. In order to generate those vowels, "di" and "losh" parts of the human speech are multiplied by sinusoidals of frequencies 2400Hz and 1600Hz, respectively [3]. To avoid overlaps human speech is filtered by a low pass filter with the cutoff frequency 700Hz. When the word is listened back a sound close to the bird speech is heard. 3-D spectrogram of this modulated wave is given in Figure 6.

CONCLUSION

Although the lovebird spends less time than the trainer to generate the word "Dilosh", duration ratio for each phoneme seems to be the same both for the lovebird and the trainer. The lovebird tries to imitate the pitch frequency by using energy envelopes. High energy components are highly effective in the lovebird speech. "i" and "o" vowels of the lovebird may be generated from the human vowels using DSBSC modulation technique. After examining other vowels generated by the lovebird, the lovebird speech may be synthesized and the synthesized words may be used to train the young lovebirds [3, 4].

REFERENCES

1. Öktay, Melekper, "Omurgalı Hayvanların Karşılaştırmalı Anatomisi," Istanbul, 1988.
2. Arslan, Beyhan, "Muhabbet Kuşu Konuşmasının Analizi," Bitirme Ödevi, Anadolu Üniversitesi, Elektrik-Elektronik Mühendisliği Bölümü, Eskişehir, 1991.
3. Kaya, Akile, "Muhabbet Kuşu Sesinin Sentezi," Bitirme Ödevi, Anadolu Üniversitesi, Elektrik-Elektronik Mühendisliği Bölümü, Eskişehir, 1992.
4. Hulse, Stewart H. et al., "An Integrative Approach to Auditory Perception by Songbirds," Comparative Perception, Volume II: Complex Signals, Edited by William C. Stebbins and Mark A. Berkley, John Wiley & Sons, 1990.

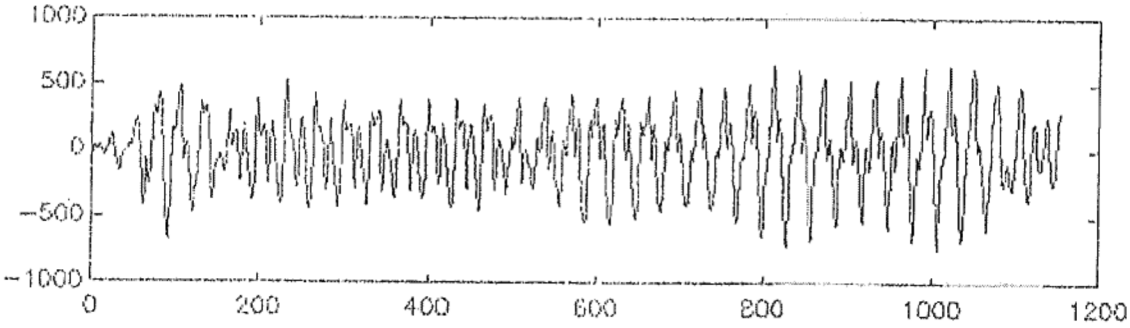

Figure 2. Sample of "Dilosh" Produced by Trainer.

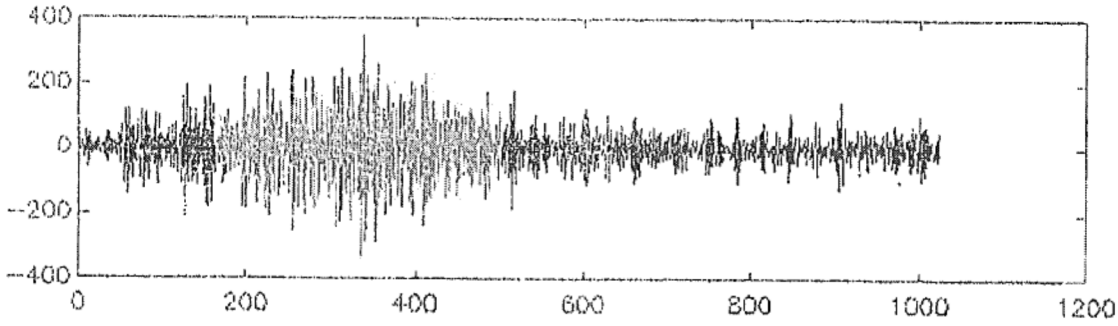

Figure 3. Sample of "Dilosh" Produced by Lovebird.

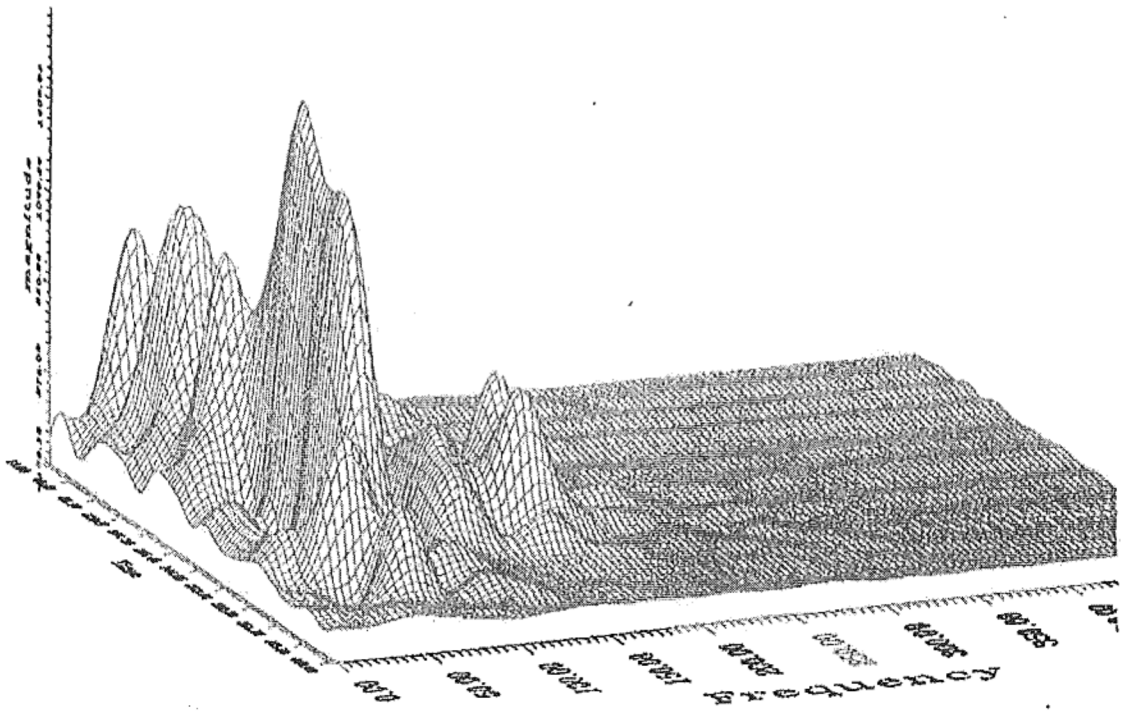

Figure 4. 3-D Spectrogram of "Dilosh" Produced by Trainer.

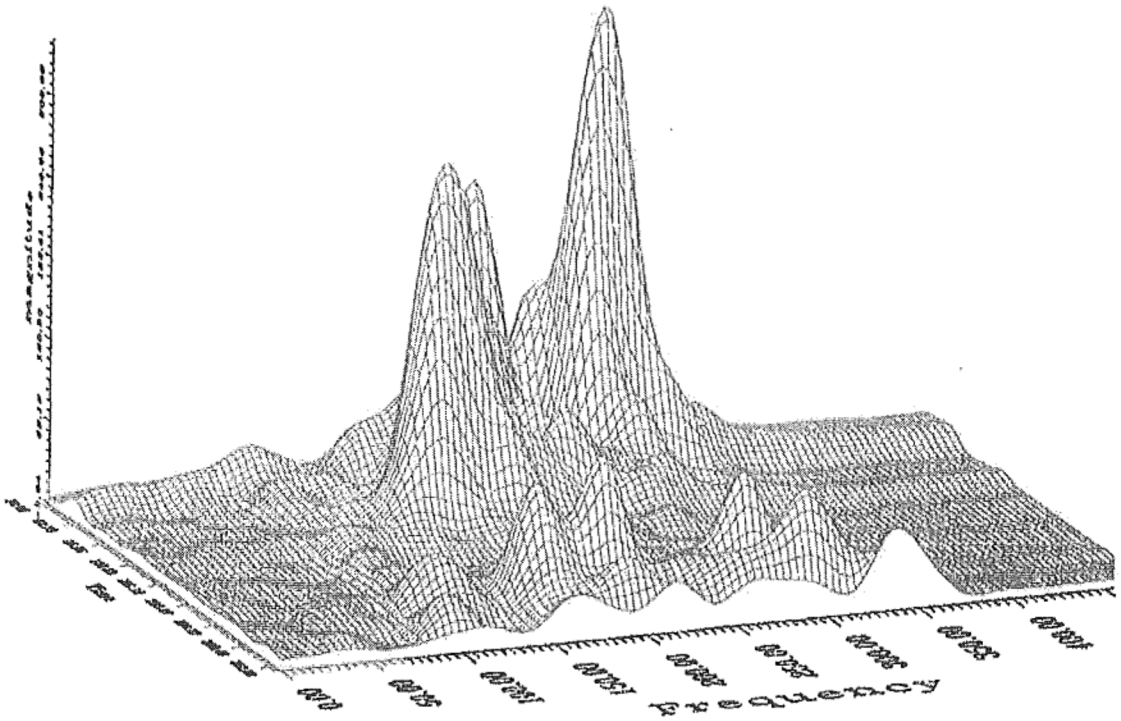

Figure 5. 3-D Spectrogram of "Dilosh" Produced by Lovebird.

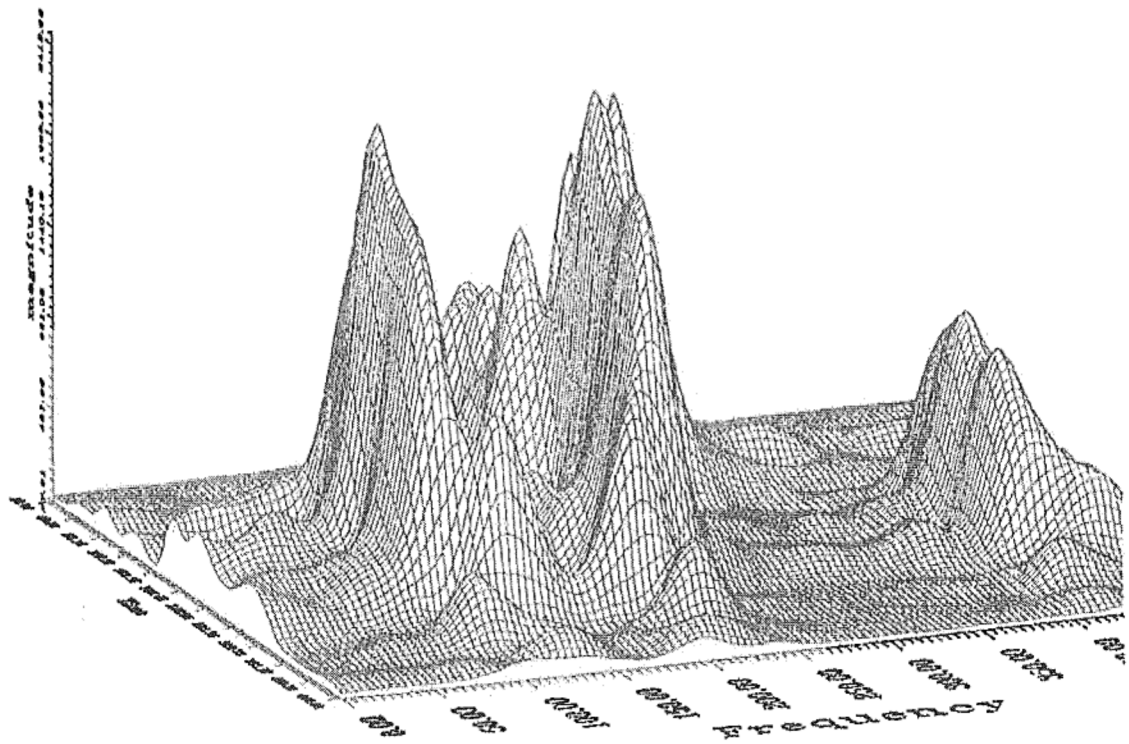

Figure 6. 3-D Spectrogram of the Modulated "Dilosh".