R. Lynn Kirlin and Lasse Hertel

Estimates of Pitch and Vocal Tract Length from Recorded Vocalizations of Purported Bigfoot*

Having analyzed a taperecording of purported Bigfoot speech using accepted techniques of signal processing, the authors conclude that the means and ranges of the recorded pitch and estimated vocal tract length of the speakers indicate that the sounds were made by a creature with "vocal features corresponding to a larger physical size than man." They also conclude that the tape shows none of the expected signs of being prerecorded or rerecorded at altered speed and hence diminish the probability of a hoax.

This paper is based on the analysis of a tape recording which was received by the authors in the spring of 1977. The circumstances under which the recording was made were reported as follows. On the night of 21 October 1972, Alan Berry, a journalist presently living in Sacramento, California, participated in the recording of what he and others believed to be one or more Bigfoot.[1] The event took place in the High Sierras of northern California "at about 8500 feet in late October after the first snowfall, some 2000 feet higher than the nearest road and about eight miles distant to the nearest established trail."[2] There were previous and subsequent recordings by members of the group at the same location, but the recording of 21 October is of exceptionally high quality and allows direct processing of the vocalizations without first specially filtering the noise. In addition, there is a wide range of vocalization, much of which shows a human-like level of articulation. There are also considerable lengths of what might be termed moans, whines, growls, grunts, and even some whistles, which no primates other than man are known to produce. The phrase might be written, "Gob-uh-gob-uh-gob, ugh, muy tail." Other professionals have listened to the tapes and have expressed their opinions, which have essentially been qualitative.[3]

The authors of this paper are neither linguists, anthropologists, nor speech pathologists, but have skills applicable to the processing of signals, including speech. The information which might be derived from speech is considerable, but only some of it is useful in attempting to answer the questions raised by the existence of these recordings. Given the constraints of the available equipment, which is really quite state-of-the-art, the first problem to the researchers was to determine what features of the vocalizations might lead to a decision as to the authenticity of the tapes. It was quickly determined that pitch frequency, the rate of opening and closing of the glottis, would be easy to extract from vowel segments and should be indicative of vocalizer size, reasoning that an extraordinarily low distribution of pitch in comparison with that of human would correspond to heavier or larger vocal chords.

Subsequently, it was also realized that formant frequencies, the resonances in speech, are an indication of the size of the vocal tract. Indeed, a review of the literature showed that speech signals can provide estimates of not only vocal tract length but also vocal tract cross-sectional area as a function of distance from the glottis to the lips.[4] However, using present techniques, the area functions are apt to be quite inaccurate for small errors in length estimation. Therefore, only length estimates and not area estimates were subsequently found, but these are sufficient for statistical comparison with known lengths of potential vocalizers other than the hypothesized Bigfoot.

Estimates of both pitch and vocal tract length are therefore extracted from segments on the tapes. This information is displayed via scattergram of pitch versus length, which allows easy visual comparison with human data, probability intervals for which are shown on the same plot. This approach is suggested for comparing data with that of other potential vocalizers, and it also allows determination of results if tape speed were changed. Lastly, extrapolation of average pitch and length estimates to body size is given, corresponding to human proportions; the results indicate a significantly large size.

VOCAL TRACT LENGTH ESTIMATES

The known estimators of human vocal tract length all have inherent variances. An estimator, which we will refer to as L_1,[5] requires knowledge of both resonant and antiresonant frequencies, but was found by the authors to work fairly well with only the resonances (formants). A modification of that estimator, which we will call L_2, uses only known formants and iterates through possible tract lengths to find a "best" length.

A more recent paper by Wakita included considerable data on human inter-speaker formant variances and length estimates for each of nine English vowels.[6] This data allowed formulation by Kirlin of a third length estimator, L_3, using maximum *a posteriori* estimation, given the formants of the vowel.[7] L_3 is quite accurate for human speech. Without *a priori* information on the human tract lengths this estimator becomes a maximum-likelihood estimator,

L_4, which allows a greater, less accurate range of lengths, more appropriate to tracts which are larger than human but which are also human-like.

The human-like criteria for L_3 and L_4 warrants further comment. The literature dealing with speech production and the evolution of the necessary vocal tract reveals that tracts of non-human anthropoids are very different in that, when body size is normalized, human tracts are considerably longer.[8] This results from the fact that human vocal chords are low in the neck, whereas others are immediately at the rear of the oral cavity, as shown in Figure 1. This difference allows human-like tracts to produce certain unique plosive consonants (|g|, |k|, for example) and formant sets as in the vowels |i|, |a|, |u|.[9] Since |g| is used in the "gob" phrase on the tape, it cannot be produced by a known non-human-like anthropoid tract. That is, the speaker is either human or has a human-like tract. If it is human, the tract length will fall in the known range for humans. If it is exceptionally long, it is likely not human. However, if length falls within human range, that does not, of course, prove it to be human.

The estimators for tract length are given in Appendix A. All four were used and the results averaged. L_3 tends to force the results to be more typical-l y human.

PITCH PERIOD ESTIMATION

The reciprocal of pitch frequency is pitch period. A nominal frequency for an adult male is 115 Hz, and the corresponding period is 8.7 milliseconds. Longer periods would indicate longer or thicker vocal chords. Due to the wide range of pitch for any human, much less all humans, only extremely low pitches (or long pitch periods) could be considered conclusive, barring tape speed changes.

Estimation algorithms for pitch are also of wide variety, but one which has been considered the best recently is that given by the cepstrum.[10] The cepstrum is defined as the inverse Fourier transform of the log-magnitude of the frequency spectrum. When a vowel is sustained for 30–50 milliseconds the resulting sound wave will normally contain several pitch periods. Processing the speech segment to yield a cepstrum produces a plot as is shown in Figure 2. The peak will occur at a time equal to the pitch period. Only those segments which have a well-defined pitch are used in the results.

FORMANT EXTRACTION

In order to estimate vocal tract length, resonances in vowel sounds must

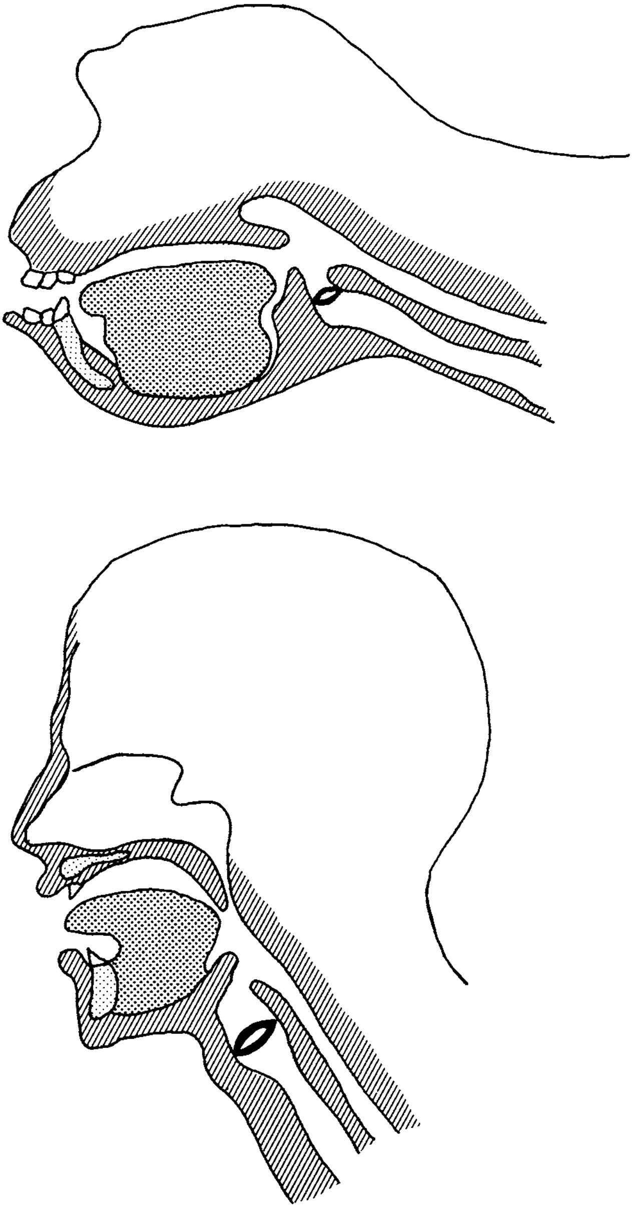

FIGURE 1: SKETCHES OF SCALED CHIMPANZEE AND HUMAN VOCAL TRACTS SHOWING DIFFERENCE IN PLACEMENT OF VOCAL CHORDS AND DIFFERENCE IN TRACT LENGTH

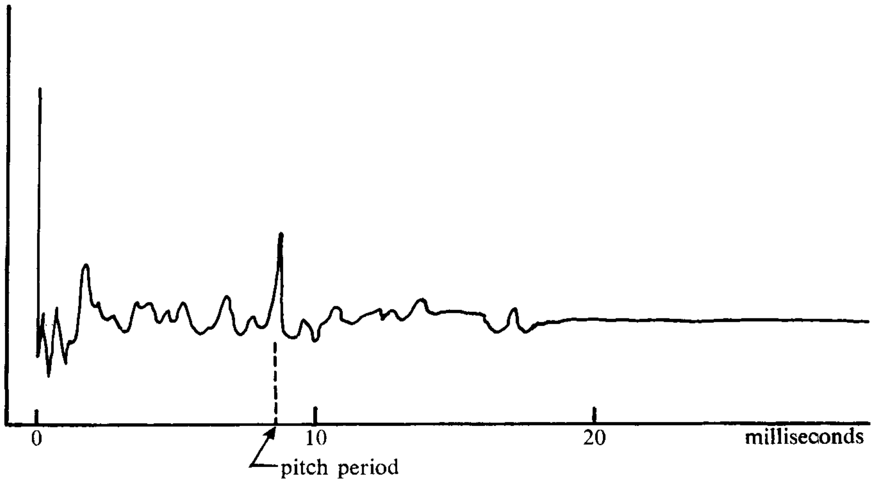

FIGURE 2: CEPSTRUM PLOT SHOWING TYPICAL HUMAN PITCH MARKER

be determined. The preliminary results of this research utilized the windowed (or weighted) cepstrum technique,[11] which essentially removes frequency variations in the power spectrum which are due to the pitch-rate impulses of glottal pressure. However, the smoothed frequency spectrum which results by Fourier-transforming the windowed cepstrum still often contains ambiguous peaks which may be erroneously interpreted as formants. The tendency is to count too many low-frequency peaks as formants, thus effecting vocal tract length estimates which become too long.

A more accurate means of determining formants is provided by "linear prediction" techniques which have been developed over the last six to eight years.[12] Linear prediction algorithms make a least-squared-error fit to the speech segment, using a predetermined number of resonances. This technique is much preferred by the authors, and the results obtained are quite reliable.[13]

Formants themselves allow comparison with human data, and this is the subject of continuing work.

RESULTS

The resulting estimates of pitch and vocal tract length are plotted in the scattergram of Figure 4. Superimposed on the data points are region borders approximately corresponding to ninety-five per cent probability intervals around means for equivalent data from human males. Human pitch statistics are given in various literature, some of which is reproduced in Figure 3.[14]

It can be seen in Figure 4 that, even though both pitch and length esti-

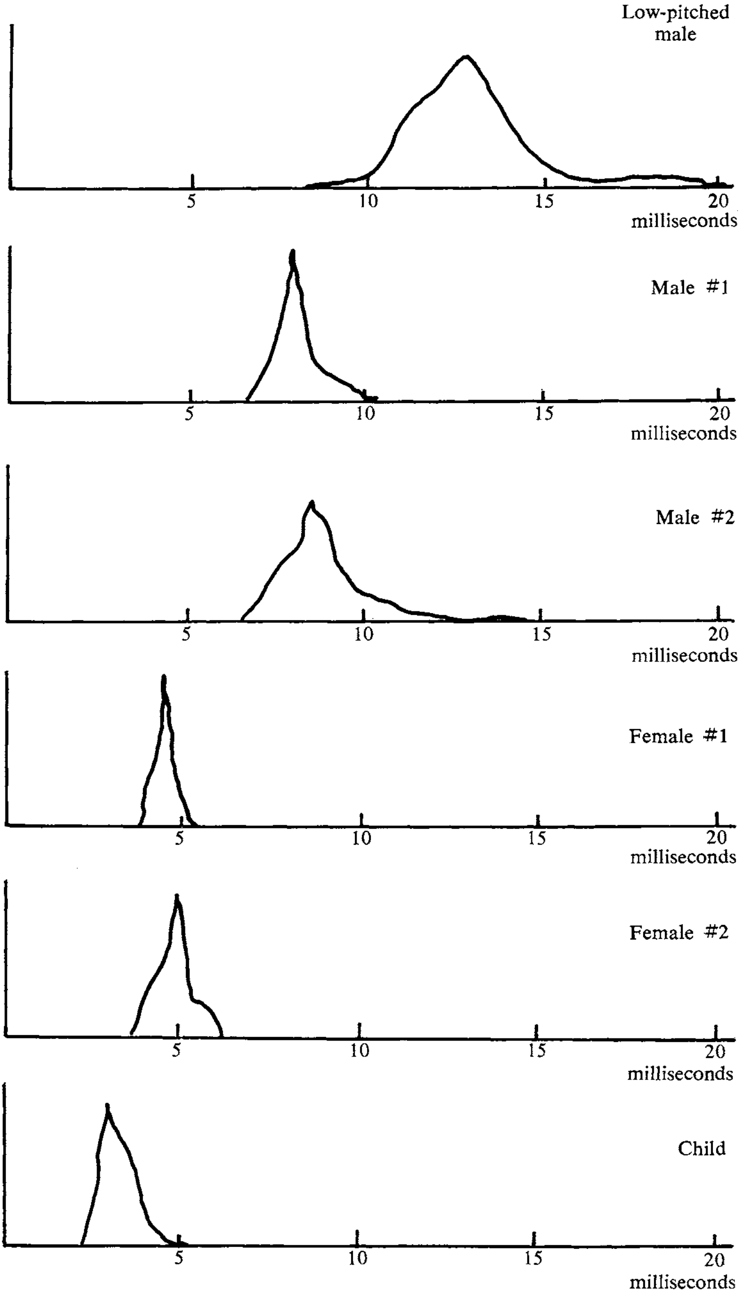

FIGURE 3: SMOOTHED HISTOGRAMS OF PITCH PERIOD

Source: After L.R. Rabiner et al., "A Comparative Study of Several Pitch Detection Algorithms," *IEEE Transactions on Acoustics, Speech, and Signal Processing* ASSP-24, no. 5 (October 1976): 399–423)

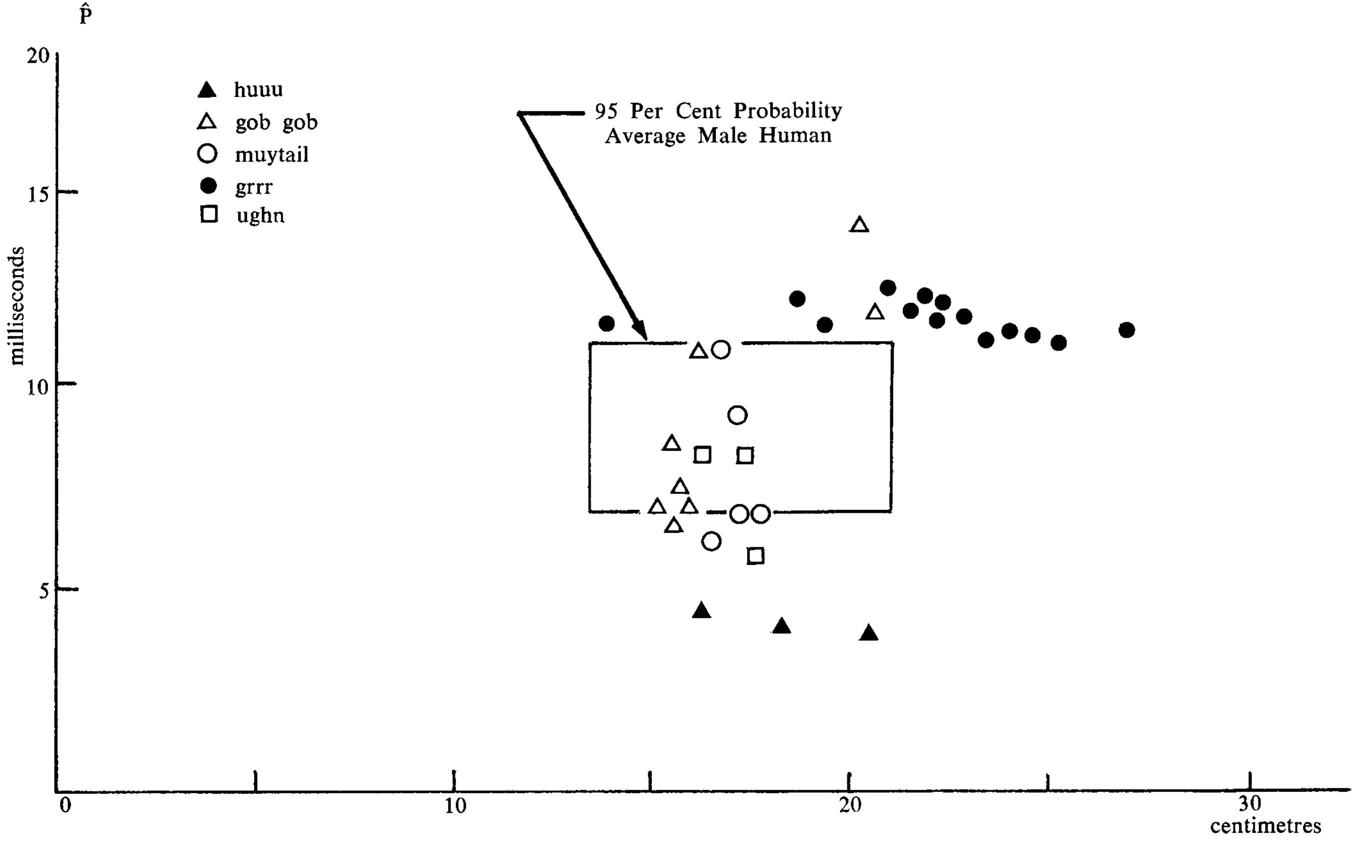

FIGURE 4. PITCH AND VOCAL TRACT LENGTH ESTIMATES WITH APPROXIMATE 95 PER CENT PROBABILITY REGION FOR NORMAL HUMAN MALE SUPERIMPOSED

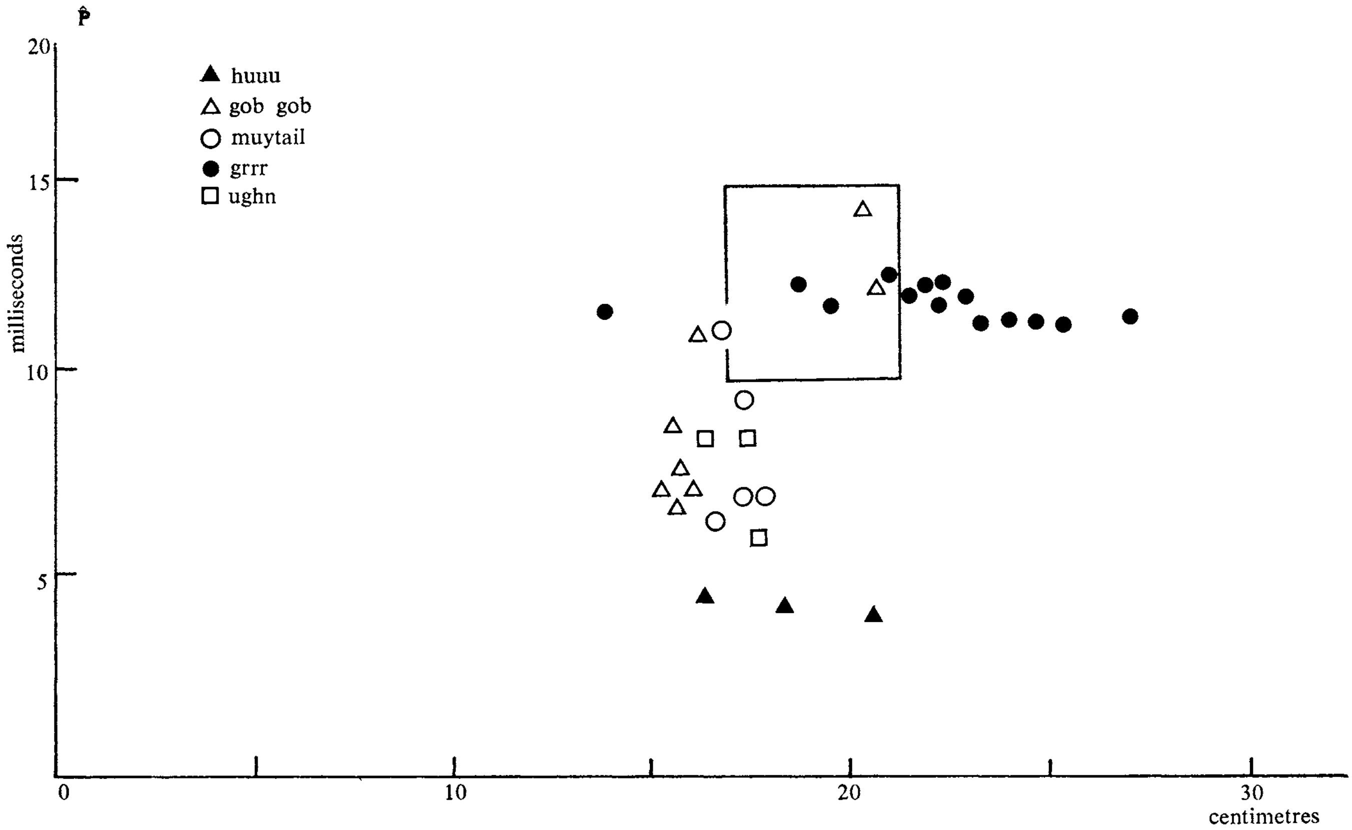

FIGURE 5: SAME DATA AS IN FIGURE 4, BUT 95 PER CENT PROBABILITY CORRESPONDS TO LOW-PITCHED HUMAN MALE AND THE VOWEL 3 WHICH REQUIRES A LONG TRACT LENGTH

mates vary considerably, the means and ranges indicate a creature with vocal features corresponding to a larger physical size than man.

Assuming 5′11″ to be the height for an average man, 115 Hz his average pitch, and 17.8 cm his average tract length, the creature or creatures on the recording, using all data shown, may be estimated to have a proportional height of 7′3″ by pitch or 6′4″ by tract length. Data from the "grr" or growl sounds alone shows quite different means, and yield heights of 8′2″ by pitch and 7′4″ by tract length.

Figure 5 repeats the same data, but superimposes the ninety-five per cent pitch and length region of a "deep voiced male" producing the vowel ɜ , which requires the longest human tract length. Note that the "grr" data falls outside this region.

The possibility of tape speed alteration should be considered. The effect of speed change on Figures 4 and 5 is easily determined. A speed-up on playback causes all recorded frequencies to appear higher; a slow-down on playback moves them lower. Playback slow-down is the situation of concern. Formant frequencies and pitch frequencies will both appear lower in proportion to the speed change. As both pitch period and vocal tract lengths are inversely proportionate to frequency, these estimates will be lengthened, both by the same proportion. For example, a tape slow-down by a factor of three would lengthen both pitch period and vocal tract length estimates by three; therefore, a data point will move along a line through the origin $\hat{p} = c\hat{L}$, where c is the constant which forces the line through the data point. This means that pitch-length ranges of any known creature could be shifted along lines of $\hat{p} = c\hat{L}$, as shown in Figure 6. Any resulting good match of these regions with the region of the Bigfoot data makes that creature a possible source of the vocalization, but on the basis of pitch and length comparison alone. Such a match concludes nothing with regard to linguistics or articulation rate. It is the opinion of the authors that the vocalizations on the tape were recorded at the speed they appear to be because the articulation rate and the range of vocal tract lengths are quite broad at constant pitch during the growl or "grr" sounds. However, the suggested matching of regions for other possible vocalizers should eventually be done.

Consideration of a human source should include the possibility of the human simply lowering his pitch. It should first be realized that 60–80 Hz pitches are difficult for most male humans to produce, and when one can it is with an accompanying decrease in volume which was not evident on the recordings.[15] An alternative possibility is prerecording with subsequent slow-down in playback, which would also proportionally increase vocal tract length estimates as shown in Figure 6. The mean pitch period estimate of about 12–13 milliseconds does show this corresponding lengthening of tract length with respect to the means of the other data, but the tract length

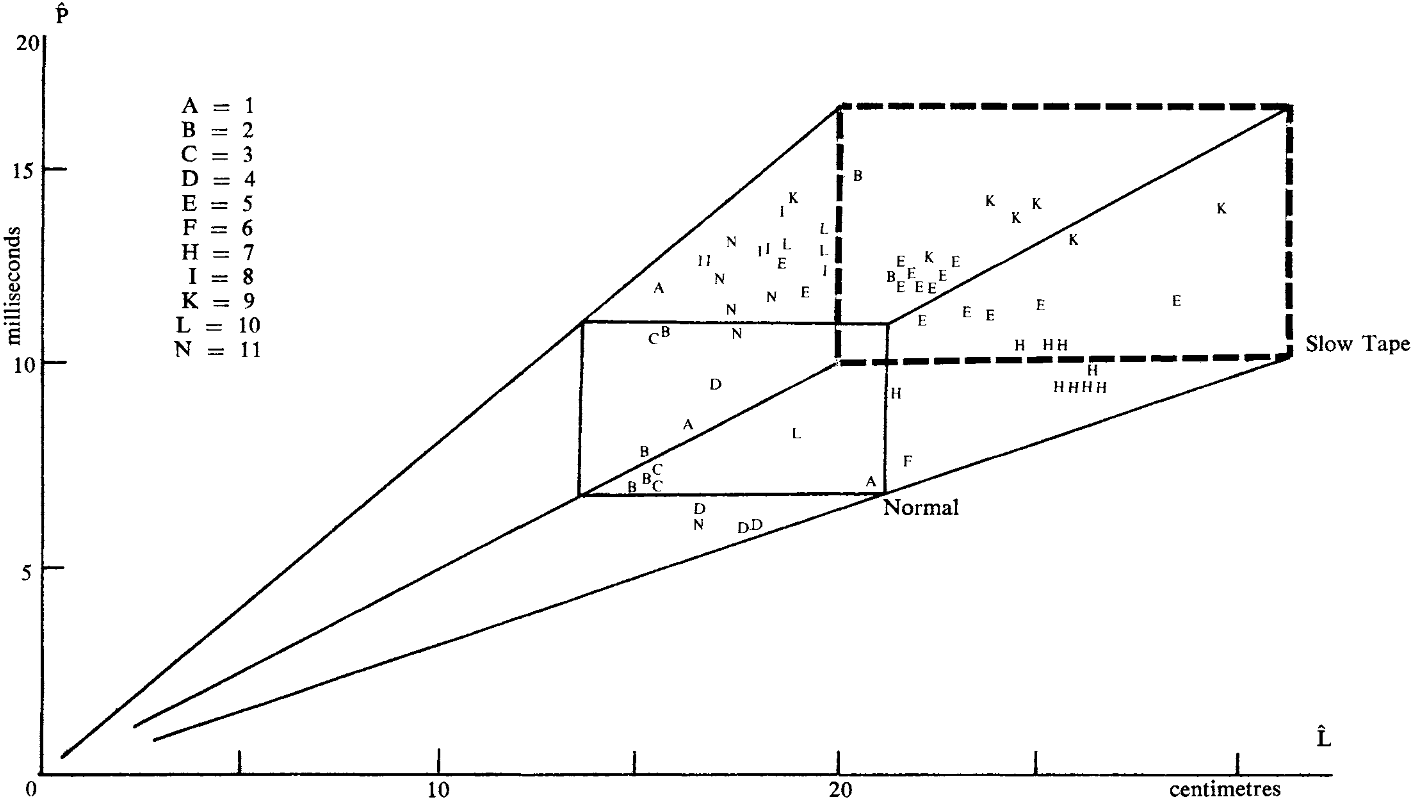

FIGURE 6: PITCH AND VOCAL TRACT LENGTH ESTIMATES WITH DIRECTION OF MOVEMENT OF DATA FOR TAPE SPEED ALTERATION SHOWN WITH APPROXIMATE 95 PER CENT CONFIDENCE INTERVALS FOR AVERAGE HUMAN PITCH AND TRACT LENGTH

range is considerably greater and not easily explained. A second alternative is prerecording with greater amplification or "close microphone" during segments of low pitch. Although this may be possible, examination of the original tape showed no 60 Hz frequencies, which would have been present in a prerecording if it had been recorded using alternating-current rather than battery power.[16] Thus any possible prerecording would fall under the constraints of battery power.

The possibility of prerecording normal language segments and rerecording by playing backward at varied speeds has been mentioned in some of the qualitative observations on spectrograms and listenings. The authors of this paper have played the tape backward and find no clearly identifiable speech. It should be realized that if any recording of any language were made and played backwards, eventually some phrase will occur which could "sound like" a known phrase in any language. Tape speed alteration is very unlikely in the "huu-u" and "gob" sequences because of the narrow range of vocal tract lengths extracted. Similarly, the growlings are quite consistent in pitch, even though tract length varies considerably. This fact is not consistent with tape speed alteration.

The possibility of more than one speaker, or even species, should also be explored. A look at the data in Figures 4 and 5 does show some gross separate clustering of tract length estimators between "grr" and the other data, but the two clusters overlap; the 2-σintervals are shown in Fig. 7. The sounds are potentially from the same species. The listener could very well imagine two creatures "conversing." (Three distinct sets of foot tracks were found the morning following the recordings session.) Vocal tract length estimates taken from these two separate segments do not show a significant difference, but even though pitch averages do, the suggestion of two creatures in these segments is not confirmed because wide pitch variations are too easy to produce. However, the "grr" cluster is a more acceptable reason for suspecting two creatures.

ANALYSIS OF THE WHISTLES

The recording contains some whistle exchanges between humans and the creatures. Analysis of the whistling is not included in the data groups used for analysis of pitch and vocal tract length, but is treated separately in this section.

There are two types of whistles found in the recordings. First, there are human types of whistling, both where there are no harmonics or formants present, and also where there are exact harmonics present, probably caused

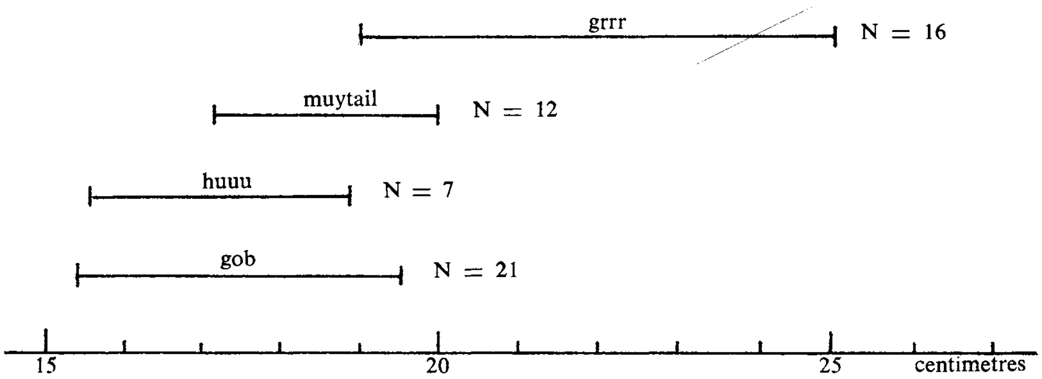

FIGURE 7: TRACT LENGTH 95 PER CENT PROBABILITY REGIONS FROM VARIOUS TAPE SEGMENTS; NUMBER OF SAMPLES = N

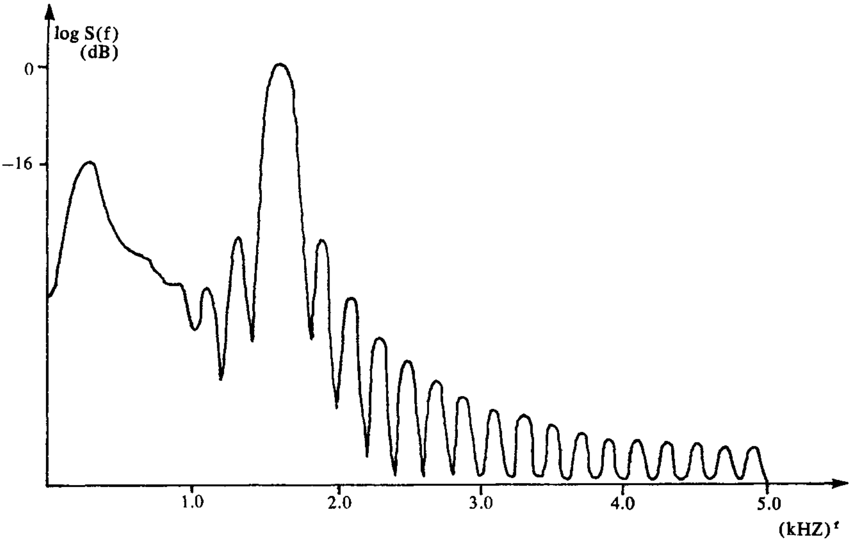

FIGURE 8: SMOOTHED POWER SPECTRUM OF A TYPICAL HUMAN WHISTLE

by a saturated microphone. A smoothed power spectrum of a typical human whistle is shown in Figure 8. Note that there are no formants or harmonics present. The low frequency components are due to the noise from the airstream. Second, there are whistles which are found to have non-harmonic formant frequencies, but no pitches.

Table 1 shows for six data segments the three first formant frequencies together with their respective vocal tract length estimates.

TABLE 1: THE THREE FIRST FORMANTS AND THE ESTIMATED VOCAL TRACT LENGTHS FOR THE ABNORMAL WHISTLES

F_1	F_2	F_3	L_1	L_2	L_3	L_4	$\bar{L}$	$^{\sigma}L$
732	1610	2509	14.8	15.2	16.5	16.5	15.7	0.86
1040	1869	3012	10.9	12.0	14.7	13.4	12.7	1.63
855	2196	2585	11.4	11.8	13.7	12.0	12.2	1.02
1102	2175	3794	9.9	10.5	13.1	11.1	11.3	1.40
359	2083	3232	14.6	14.9	15.2	14.3	14.8	0.4
959	1764	2929	12.4	13.4	15.0	14.0	13.7	1.09
Average			14.7	13.6	12.3	13.0	13.4	1.07
Standard Deviation			2.00	1.86	1.19	1.89	1.65	0.43
Average Variation between Estimators $\bar{V}$ = 8.3%								

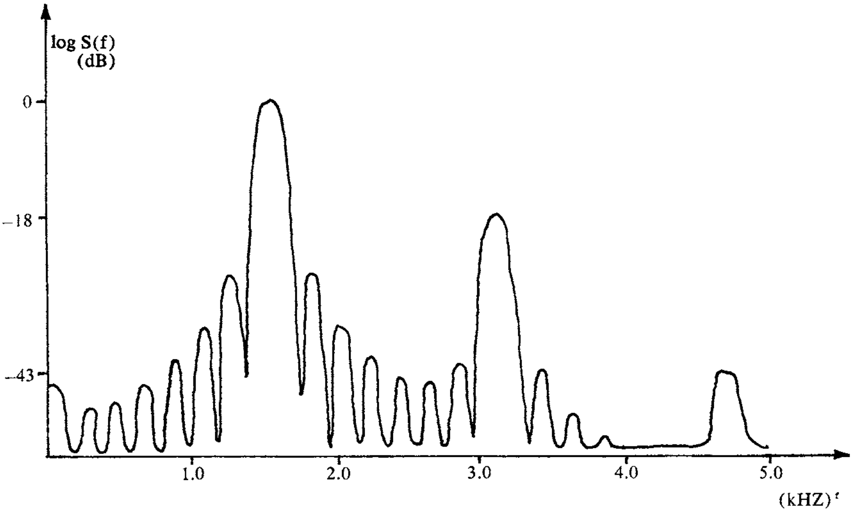

FIGURE 9: SMOOTHED POWER SPECTRUM OF A TYPICAL HUMAN WHISTLE, PRODUCED WITH A SATURATED MICROPHONE

By amplifying the whistle, the microphone can be saturated, and it will then produce harmonics as shown in Figure 9.

The formants were found using the linear prediction technique, and the values were checked using a smoothed power spectrum of each segment.

The formants and corresponding short vocal tract lengths found indicate the likelihood that the creatures could be able to whistle utilizing only a part of their vocal tract. If the creatures have a human-like vocal tract, they might be able to whistle using the constriction between the two vocal cavities. Such whistles can also be produced using some kind of a musical instrument, known to produce both harmonic and non-harmonic overtones.

CONCLUSIONS

The results indicate more than one speaker, one or more of which is of larger physical size than an average human adult male.

The formant frequencies found were clearly lower than for human data, and their distribution does not indicate that they were a product of human vocalizations and tape speed alteration. Although a time-varying speed could possibly produce such formant distributions, an objective hearing and the articulation rate do not support that hypothesis.

Statistical analysis was applied to groups of vocal tract estimates from different vocalizations, and a significant difference was found between the groups. When compared with human data the results indicated that there could possibly be three speakers, one of which is non-human. The average vocal tract length was found to be 20.2 cm. This is significantly longer than for a normal human male. Extrapolation of average estimators, using human proportions, gives height estimates of between 6′4″ and 8′2″.

Analysis of the rapid articulations in the beginning of the recording (gob-gob) resulted in human-like vocal tract lengths. Also, the sound /g/ in "gob" suggests a human-like vocal tract (two vocal cavities).

The pitch periods found cover the broad range of pitch periods for both normal human male and low pitched human male. However, they are mainly distributed around the data for the low-pitched human male.

Pitch and length estimates vary considerably but they are all found to be within the 95 per cent confidence interval for human speech with varying tape speed; however, assuming that there is only one vocalizer, then time-varying tape speed is necessary to produce data over such a wide range.

Both typical human whistles and some abnormal types of whistles were found. By using the formants from the abnormal whistles, very short vocal tract lengths were estimated. These whistles could either have been produced with some kind of a musical instrument or by the creature using only a part of its vocal tract.

It is hoped that the remaining uncertainties will not be considered reason for dismissing the recordings. The possibilities for prerecording are many, but there is no clear reason to believe it is likely. If Bigfoot is actually proven to exist, the vocalizations on these tapes may well be of great anthropological value, being a unique observation of Bigfoot in his natural environment.

Appendix: Vocal Tract Length Estimators

Four of the best vocal tract length estimators were used in producing the results in this paper.[17] All four involve a weighting of the speech signal's resonant and antiresonant (critical) frequencies f_k, $k = 1, 2, 3, \ldots$, which include the formants F_i, $k = 2i - 1$; that is, odd k correspond to the formants. Tract length is determined by [18]

$$\hat{L} = \frac{35300}{4\hat{f}_0} \text{ cm.} \tag{1}$$

where $\hat{f}_0$ is an estimate of a "fundamental" frequency which is the source of "harmonics" (the resonances), which can be thought of as being displaced from their normal position by the non-uniformity of the vocal tract tube.

Paige and Zue,[19] use

$$\hat{f}_0 = \sum(f_k/k)^2/\sum(f_k/k). \tag{2}$$

Knowledge of f_k for $k = 1, 3, 5$ gives the estimate L_1 through (2) and (1). Paige and Zue produced another estimator by choosing that $\hat{L}$ which, after extrapolating known f_k to higher $\hat{f}_k$ using the assumed $\hat{L}$, minimized the area-function perturbation. The estimator so produced is L_2.

Using formant mean and variance data listed by vowels and sex of speaker, given by Wakita,[20] Kirlin produced a third estimator,

$$L_3 = \frac{35300(\sum k^2/\sigma_k{}^2 + 1/\sigma_0{}^2)}{4\sum kf_k/\sigma_k{}^2 + \bar{f}_0/\sigma_0{}^2}, \tag{3}$$

where, over the speaker population and all vowels, σ_k^2, $= 1, 3, 5\ldots$, is the variance of the kth critical frequency, $\sigma_0^2 =$ the variance of the $\hat{f}_0$, and $\bar{f}_0$ is the mean tract length. By Wakita's data (mixed male and female), $\sigma_1 = 166$, $\sigma_3 = 417$, $\sigma_5 = 348$, $\sigma_0 = 62.3$ and $\bar{f}_0 = 537$.

If $\sigma_0{}^2 \rightarrow \infty$, L_3 becomes L_4, a maximum-likelihood estimator, which uses information about the variation of the formants, but not information about the population mean f_0. L_3 and L_4 are applicable to human-like vocal tracts.

The order of accuracy for humans is (best first) L_3, L_2, L_1, L_4, with the mean-squared errors (over nine vowels) ranging from 3.05 per cent to 10.4 per cent accepting Wakita's length estimates as correct.[21]

Notes

*This article includes, in addition to material presented at the Conference, data from Lasse Hertel, "An Application of Speech Processing Techniques to Recordings of Purported Bigfoot Vocalizations to Estimate Physical Parameters" (M.S. thesis, University of Wyoming, 1978).

[1] A description of the circumstances surrounding the recording is given in Alan Berry and A. Slate, *Bigfoot* (New York: Bantam Books, 1976), chapters 1, 2, and 3, and Appendix B.

[2] Ibid.

[3] Some of these are reproduced in ibid., Appendix B, including a spectrograph of about

four seconds of the recording, in which the "speech" is highly articulated and thus the subject of controversy.

4. See, especially, A. Paige and V. Zue, "Calculation of Vocal Tract Length," *IEEE Transactions on Audio and Electroacoustics* 18, no. 3 (1970): 268–70, and "Computation of Vocal Tract Area Functions," *IEEE Transactions on Audio and Electroacoustics* 18, no. 1 (1970): 7–18; M.R. Schroeder, "Determination of the Geometry of the Human Vocal Tract by Acoustic Measurements," *Journal of the Acoustic Society of America* 41, no. 4, part 2 (1967): 1002–10; P. Mermelstein, "Determination of the Vocal-Tract Shape from Measured Formant Frequencies," *Journal of the Acoustic Society of America* 41, no. 5 (1967): 1283–94; H. Wakita, "Direct Estimation of the Vocal Tract Shape by Inverse Filtering of Acoustic Speech Waveforms," *IEEE Transactions on Audio and Electroacoustics* 21, no. 5 (1973): 417–27, and "Normalization of Vowels by Vocal-Tract Length and Its Application to Vowel Identification," *IEEE Transactions on Audio and Electroacoustics* 25, no. 2 (1977): 183–92; and H. Wakita and A.H. Gray, Jr., "Numerical Determination of the Lip Impedance and Vocal Tract Area Functions," *IEEE Transactions on Audio and Electroacoustics* 23, no. 6 (1975): 574–80.
5. See "Calculation of Vocal Tract Length."
6. "Direct Estimation of Vocal Tract Length."
7. R.L. Kirlin, "A Maximum A-Posteriori Estimation of Vocal Tract Length," *IEEE Transactions on Acoustics, Speech and Signal Processing* (Dec. 1978): 571-74.
8. See P. Lieberman, "On the Evolution of Language: A Unified View," in *Primate Functional Morphology and Evolution*, ed. Russell H. Tuttle (The Hague and Paris: Mouton, 1975).
9. Ibid.
10. See A.M. Noll, "Cepstrum Speech Determination," *Journal of the Acoustic Society of America* 41 (1967): 293–309, and L.R. Rabiner et al., "A Comparative Study of Several Speech Detection Algorithms," *IEEE Transactions on Acoustics, Speech and Signal Processing* 24, no. 5 (1976): 399–423.
11. See J.L. Flanagan, *Speech Analysis, Synthesis, and Perception* (New York: Springer-Verlag, 1972).
12. See J.D. Markel and A.H. Gray, Jr., *Linear Prediction of Speech* (New York: Springer-Verlag, 1976).
13. The computer algorithm is described in ibid. A polynomial root-finding subroutine is also required.
14. Reproduced from ibid. Vocal tract estimates for male humans are given in Wakita, "Normalization of Vowels."
15. Some pitches in the low sixties were recorded but are not shown in this data.
16. See Berry and Slate, *Bigfoot*, Appendix A.
17. A comparison of many vocal tract length estimators is given in Kirlin, "Maximum A-Posteriori Estimation."
18. See ibid. and Paige and Zue, "Calculation of Vocal Tract Length."
19. "Calculation of Vocal Tract Length."
20. In "Normalization of Vowels."