

# Analysis of Anomalous Vocalizations

UNLV Auditory Perception Laboratory

Michael D. Hall, Ph.D., Supervisor

Kimberly Wieberg and Melissa Griffith, Graduate Research Assistants

## *Original and Digitized Recordings*

Fourteen sets of vocalizations could be determined from the recording. The recording was very noisy, reflecting a mixture of influences from the recording media and additional sound sources. For example, at the onset of each vocalization, a periodic mechanical sound can be heard. The periodic nature of the sound is consistent with circular motion, perhaps reflecting the initiation of the recording device, such as a tape feed from a reel-to-reel tape recorder. If the original recording was made on a reel-to-reel recorder, then at best the laboratory was working from a cassette copy of the recording. It is more likely that the laboratory was provided a secondary cassette copy. Each additional copy should result in further degradation of signal-to-noise ratio, as the medium will introduce original noise atop the original signal amplitude. [No information was provided about microphone sensitivity. Thus, it is unclear what filtering contributions may have been made by the selected microphone]. To make matters worse, signal-to-noise ratio was already very low due to the fact that the recordings were necessarily collected at a great distance from the source (i.e., resulting in signals with very low amplitude). Furthermore, the recording team made frequent interjections during the vocalizations, which had to be excised (along with a portion of the signal) following digitization.

Digitization and editing of vocalizations was accomplished using Syntrillium's *Cool Edit Pro* software. Monaural output from the cassette was digitized at a 44.1 kHz (16-bit) sample rate. Each vocalization constituted a separate .wav file. Any silence preceding or following each vocalization was excised. Similarly, all spurious noises that exceeded the maximum amplitude of the given vocalization were excised at zero (voltage) crossing points.

For noise reduction procedures, the longest uninterrupted section of ambient noise, including environmental and machine noises, was identified and saved as a separate .wav file. The entire noise file then was used to generate an average spectral profile consisting of 9,999 samples of the noise file to be filtered out of the vocalization (i.e., signal+noise) files. Vocalization files were submitted to maximal noise reduction levels using a fast-Fourier transform of 8,192 sampling points (for fine frequency resolution), a high precision factor (17, to minimize amplitude distortion), minimal smoothing across frequency bands (1, so as not to affect background noise level), and no transition width (0 dB) between noise and signal amplitudes.

Following noise reduction, vocalization files were normalized to maximum peak amplitude. Also, any residual, low-amplitude noises that did not reflect the target signal were excised at zero-crossing points. On the accompanying CD, these noise-reduced versions of the

vocalization files have been converted to stereo files with amplitude ramped linearly over 10 ms from zero to maximum amplitude at vocalization onset, and conversely, at offset. Monaural versions of the waveforms without amplitude ramps were submitted to spectral analyses, and were reduced to vowel steady-states to insure consistent formant frequencies (i.e., reflecting contribution from a single vowel) in the computation of average spectra.

### *Formant Analyses*

Spectral analyses of the steady-state versions of the vocalizations were conducted within the *Computerized Speech Research Environment* (CSRE, version 4.5). Spectrograms were generated according to an autocorrelation technique using a Hanning window of 512 samples, 256 frequency bands, 60 percent overlap (to smooth output) and 98 percent signal pre-emphasis (to maintain a spectral tilt consistent with speech—around -6dB/octave). For each steady-state, an average spectral window was calculated across the entire signal. Formant center frequencies (indicating spectral peaks) and their corresponding amplitudes (in relative dB) were recorded.

The data obtained from the spectrograms is summarized for each vocalization in Table 1, along with the corresponding vowel category that was perceived by the laboratory staff. Most vocalizations reflected sustained vowel sounds from a single vocal tract configuration, primarily /a/; the indication of multiple categories for a given vocalization in Table 1 reflects that the steady-state was perceived to be between two categories and/or to move at offset to an alternative vowel category. As Table 1 shows, only two discernable formants could be determined following noise reduction for most of the productions. Fortunately, two formants are all that are required to reliably distinguish vowel categories.

*Table 1. Formant frequencies (F1-F3), corresponding amplitudes in relative dB (i.e., 0 is maximum), and dispersion (F<sub>D</sub>) measures for each vocalization.*

Beginning Time Code	Vowel(s)	F1(Hz)	F1 (dB)	F2 (Hz)	F2 (dB)	F3 (Hz)	F3 (dB)	F <sub>D</sub> (Hz)
5:35	/a/	500	-45.7	1031	-61.2			531
7:16	/a/	656	-46.0					
4:10	/a/-/o/	625	-38.2	968	-52.7			343
5:48	/a/-/o/	562	-40.3	937	-63.5			375
7:26	/a/-/o/	656	-50.4					
2:16	/a/-/o/	531	-52.9	750	-57.2	1968	-61.1	719
6:53	/a/-/o/	625	-50.9	937	-74.9			312
0:54	/a/-/o/	625	-48.5					
3:56	/o/	625	-38.5	937	-57.6			312
4:33	/u/	625	-40.1	1343	-67.9			718
4:49	/u/	718	-36.7	2015	-72.6			1297
5:16	/ʌ/	500	-44.4	1250	-62.0			750
							<b>M</b>	<b>595</b>
							<b>SE</b>	<b>107</b>

In order to determine whether or not the produced formants were consistent with human productions, results from each vocalization were compared against recently obtained data from humans. The human comparison sample came from Hillenbrand, Getty, Clark, &

Wheeler (1995), who provided an extensive set of acoustic analyses for productions of the desired vowel categories by 45 men, 48 women, and 46 children. The obtained F1 and F2 values from each vocalization are plotted against the Hillenbrand, et al. (1995) data in Figure 1. As indicated by the ellipses in Figure 1, the majority of productions fit within human vowel categories. However, most of the vocalizations occurred at category boundaries, thereby representing outliers. Furthermore, one /a/ production failed to fall within any ellipse fit to human data, indicating that these vocalizations were very unusual. It is worth noting that this production does fall within an ellipse fit to the seminal Peterson & Barney (1952) data for an alternative vowel category, but still clearly represents an outlier according to that alternative data set.

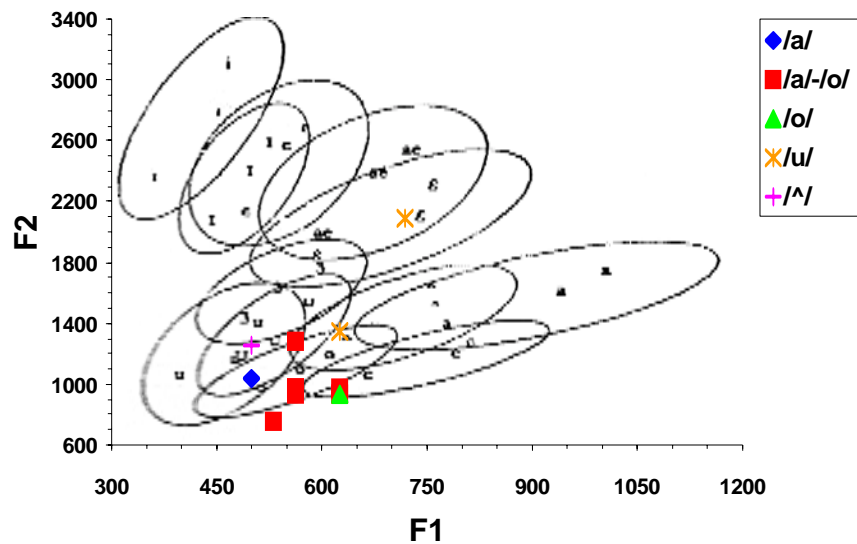


Figure 1. F1 and F2 center frequencies (in Hz) for each recorded vocalization. Values are plotted against ellipses fit to production data from each vowel category by adult males and females, as well as children (from Hillenbrand, et al., 1995). Within each ellipse/vowel category, mean values for each human sample population are indicated by phonetic symbols (lower values within each category correspond to adult males).

Also included in the rightmost column of Table 1 are formant dispersion values, which reflect the average difference (in Hz) between adjacent formants. Recent data from Fitch (1997) indicate that formant dispersion measures are highly correlated with measures of vocal tract length in macaques, and vocal tract length is a strong indicator of body size/weight. In fact, formant dispersion provided correlation coefficients ( $r$ ) of  $-0.922$  and  $-0.868$  for oral and nasal vocal tract lengths, plus coefficients of  $-0.816$ ,  $-0.886$ , and  $-0.869$  for body length, log weight, and skull length, respectively. Fitch also revealed that the obtained dispersion values for macaques could be predicted effectively by a simple one-tube model  $D_{pred} = c/2L$ , where  $D_{pred}$  is the predicted dispersion (in Hz),  $c$  is the speed of sound (335 m/s), and  $L$  is the length (in m) of the vocal tract.

In a similar fashion, an attempt was made to use the dispersion values from Table 1 to estimate vocal tract length for the target sound-producing object. However, there were a couple of major obstacles to obtaining an effective estimate of vocal tract length. First, only two formants could be derived for most of the productions. Second, the vocal tract configurations were typically consistent with /a/, which reflects a compact spectrum between F1 and F2 center frequencies. Taken in combination, these factors should reduce dispersion, which should result in inappropriately elevated estimates of vocal tract length.

In order to minimize the contribution of these factors, estimation of vocal tract length was restricted to signals with three or more formants. Unfortunately, there was only one such vocalization. The estimate of vocal tract length derived from this signal is provided along with 2- and 3-formant dispersion values for /a/ and /o/ vocalizations and corresponding human comparisons in Table 2. The obtained dispersion values are lower than the average values obtained for human males (according to Hillenbrand, et al., 1995), resulting in a longer estimated vocal tract length than for human males. However, both estimates in Table 2 appear to be inappropriately high. The estimated length for the human male's vocal tract is clearly elevated from the average length of 16 cm that is typically assumed in tube models. Also, the two-formant dispersion values in Table 2 are comparably low for humans and the target set of vocalizations, reflecting similarly compact spectra. Furthermore, the estimate based on the sole three-formant vocalization from the cassette reflects a clear problem of insufficient sampling. Estimates of vocal tract length are likely to become more comparable to average values from human males given a larger sample of vocalizations consisting of more formants. To what degree such estimates approach human values remains to be determined.

*Table 2. Comparison of mean formant dispersion measures [based on either the first two (F1-F2) or first three (F1-F3) formant center frequencies] for /a/ and /o/ vowels produced by human males (from Hillenbrand, et al., 1995) and the corresponding recorded vocalizations. Estimates of vocal tract length [i.e.,  $L = (c/F_D)/2$ ] derived from the three-formant dispersion values also are provided (right column).*

	F <sub>D</sub> (Hz)		Estimated Vocal Tract Length (cm)
	F1-F2	F1-F3	
human	489	929	18.0
recorded vocalization	432	719*	23.3

\*based on 1 signal (2:16)

### *Other Vocalizations*

Excluded from the analyses above are a short series of nine productions. A subset of five complete, more intense productions have been included on the CD for archiving purposes. Summary values for these alternative productions are provided in Table 3. The stimuli were excluded for several reasons. First, they are inconsistent with all the other, sustained vowel productions in that they reflect movement between two vowel categories (/uwa/). Second, these short vowel movements sound like primate vocalizations, but there are no primates native to the wooded areas of the Northwest. Third, they are more intense than most of the other productions on the cassette tape, including adjacent regions of that tape.

This suggests that the recording of these vocalizations may have been obtained at a closer distance to the microphone.

Table 3. Formant frequencies, corresponding amplitudes in relative dB (i.e., 0 is maximum), and dispersion (FD) measures for each /uwa/ vocalization.

Beginning Time Code	Vowel(s)	F1(Hz)	F1 (dB)	F2 (Hz)	F2 (dB)	F3 (Hz)	F3 (dB)	F <sub>D</sub> (Hz)
3:39	/uwa/	718	-40.1	1562	-59.7			844
	/uwa/	687	-32.9	1593	-48.6			906
	/uwa/	687	-47.5	937	-53.2	1593	-71.2	453
	/uwa/	625	-46.3	1062	-60.7	1562	-59.4	468.5
	/uwa/	656	-44.4	1000	-56.7	1593	-56.1	468.5
4:25	/o/	562	-51.3	687	-40.4	2031	-74.1	735

Also excluded from analyses, but included on the CD and summarized in Table 3, is a raspy /o/ production that was identified on the cassette by members of the recording team as \*\*\*\*\*. Dispersion measures from this signal suggest an estimated vocal tract of almost 23 cm. If signal degradation (i.e., lack of formants) is responsible for this gross overestimation of vocal tract length, then it would be reasonable to argue that other sounds on the cassette also could have been produced by a human male. It also is noteworthy that the spectral distribution in this /o/ production is similar to other vocalizations. A visual depiction of spectral similarity to another /o/ vocalization is provided in Figure 2. Whereas the second formant is less distinguishable in the vowel that is known to be produced by a human male (right), both productions reflect a similar compacting and center frequency of F1 and F2, as well as a similar distribution of higher frequency components.

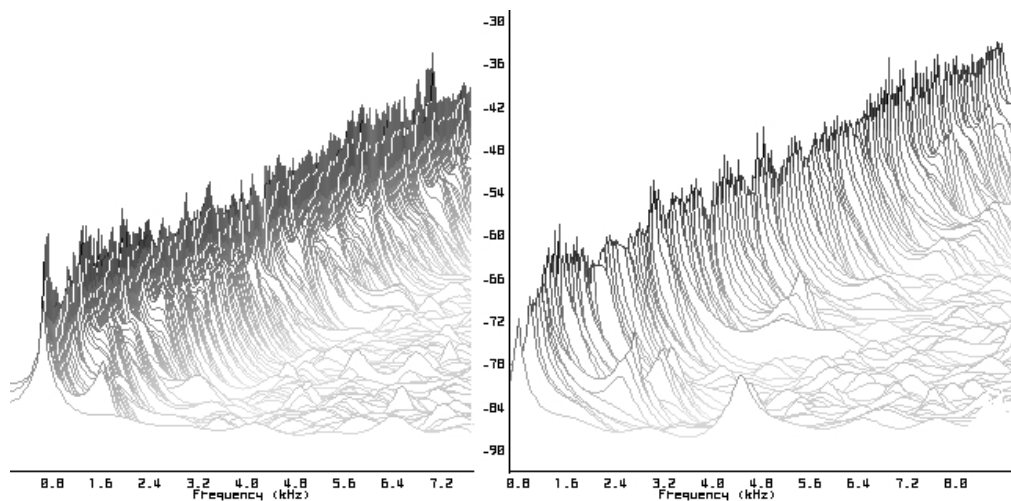


Figure 2. Waterfall display of spectrum from an /o/ vocalization (time code 3:56, left) and the raspy /o/ production (4:25) that was identified as being produced by a human male (right). Amplitude (in relative dB) is displayed on the vertical axis, and frequency is displayed on the horizontal axis.

In light of these observations, the collection of vocalizations that were excluded from analyses may represent a good point of further inquiry with the recording team. For instance, it would be helpful to know whether any of these vocalizations were believed to be produced by the same sound-producing object that produced the other isolated vowels. If the recording team indicates that all vocalizations on the cassette were produced by the same object, then it would be possible that all of the recordings were produced by a human male. Alternatively, if the /uwa/ vocalizations were not produced from the same source, then there would be an additional basis for excluding them from analyses beyond the fact that they consist of multiple vowel sounds.

### *Fundamental Frequency Analysis*

An additional goal of the current research project was to provide for each vocalization an evaluation of fundamental frequency (F0), which is a primary contributor to pitch. Unfortunately, F0 analyses with CSRE were unsuccessful using either Cepstrum- or Comb-filtering, though the latter (using a 1,000 Hz cut-off frequency for zero-crossing calculations) produced slightly more reasonable results. The noisy recording made it impossible for the pitch-tracking algorithm to consistently assign zero-crossings at the beginning/end of each period of the waveform, which forms the basis of F0 calculations. An alternative would be to assign such zero-crossings by hand for each waveform. However, given the number of signals and the length of each signal, this alternative would require much more time than the funded period allows. Zero-crossings also would be very difficult to assign in low-amplitude portions of the signals. The aforementioned evidence from primate studies also suggests that formant dispersion correlates more strongly with body size than does pitch. Thus, it is possible that even if F0 measurements from the current sample could be obtained, conclusions about body size would be no more valid.

### *Preliminary Conclusions*

At this early stage of analysis, it appears that the vocalizations could have been produced by a human male. However, there are a few unusual aspects of the vocalizations that clearly warrant further investigation. First, the majority of sustained productions are at or near vowel category boundaries, which is very unusual. If these productions were produced by a human male, then an explanation would need to be sought for why the set of vocalizations consistently represent outliers. For example, it is possible for a human to intentionally produce such outliers. It also is possible that such outliers could reflect a limitation in early speech development, such as exceeding a critical period by being raised in isolation. Outliers also could reflect a physiological limitation, such as an anomalous or damaged vocal tract. A second finding that warrants investigation is that many of the vocalizations reflect formants with low center frequencies, suggesting that they may have been produced by a large body. While it is acknowledged that this suggestion at least in part reflects measurement limitations

due to reliance on compact spectra with a minimal number of discernable formants, the currently observed pattern of formant dispersion is not conclusively human.

Remaining questions about the vocalizations are likely to be effectively addressed by future research. Although it would be preferable to obtain recordings with a high sample rate directly on a laptop computer or comparable device (using a sensitive microphone accompanied by a windscreen, and without additional discussion/noise sources), it is probably not feasible to obtain such recordings. Alternatively, additional work could be done with the existing recordings to obtain a clearer depiction of signal properties. For example, waveforms could be digitized from the original recording in order to minimize data loss that has certainly occurred due to the recording media. Furthermore, an array of noise-reduction levels and techniques could be compared in an attempt to obtain a larger complement of formants for dispersion measures. An analysis of F0 also could be accomplished using a Comb-filter algorithm via hand placement of zero-crossings. Finally, the obtained vocalizations could be simultaneously compared with existing catalogues of primate vocalizations and data from humans in order to determine whether the vocalizations more closely approximate one or the other population. Insofar as there should be at least rough existing approximations of the size of primates that produced sounds for a database, this latter approach also might provide further insight about the size of the sound-producing object in question.

*Send correspondence to:*

Michael D. Hall, Ph.D.  
Psychology Department  
University of Nevada, Las Vegas  
4505 Maryland Parkway, Box 455030  
Las Vegas, NV 89154-5030