Mapping the phonetic structure of humpback whale song units: extraction, classification, and Shannon-Zipf confirmation of sixty sub-units

Howard Pines

Citation: Proc. Mtgs. Acoust. **35**, 010003 (2018); doi: 10.1121/2.0000957
View online: https://doi.org/10.1121/2.0000957
View Table of Contents: https://asa.scitation.org/toc/pma/35/1
Published by the Acoustical Society of America

Volume 35 http://acousticalsociety.org/

176th Meeting of Acoustical Society of America
2018 Acoustics Week in Canada
Victoria, Canada
5-9 Nov 2018

Animal Bioacoustics: Paper 5aAB8

Mapping the phonetic structure of humpback whale song units: extraction, classification, and Shannon-Zipf confirmation of sixty sub-units

Howard Pines
Wireless Networking Business Unit, Cisco Systems - Retired, El Cerrito, CA, 94530; howardpines@ieee.org

The striking similarities of time-frequency spectrograms of voiced human speech and humpback whale vocalizations suggested a common targeted frequency-modulated phonetic basis. To map the sub-unit structure of humpback whale song units, a time-frequency contour segmentation, extraction, and classification procedure was developed and tested on streaming voiced human speech. When the procedure was applied to humpback vocalizations and the tone-pairs of the two most energetic "vocal fold" harmonic frequencies were plotted in x-y coordinates, the plot exhibited properties of an optimally structured Shannon "modem symbol constellation" diagram of 14 distinct sub-regions and 60 acoustically distinct sub-unit symbols. The humpback symbol constellation is structurally comparable to the tone-pair symbol constellations of English and Asian language vowels. The information entropy and plot of the humpback sub-unit symbol set's cumulative probability vs. ranked frequency distribution function are nearly identical to the entropy and Zipf power-law profile of the English language phoneme set. The precise specification of the sub-unit structure of more than one hundred song units analyzed to date has resulted in a significant expansion in the number of unique units identified in previous studies, suggesting that the "lexicon" of humpback song units is potentially much larger than cited in the literature.

Published by the Acoustical Society of America

 https://doi.org/10.1121/2.0000957

Introduction

Nearly a half-century has elapsed since Roger Payne and Scott McVay discovered the unique structure of humpback whale "songs": that they are comprised of a sequence of acoustically distinct units of variable duration up to a few seconds, separated by silences, and arrayed into patterns exhibiting short and long-term periodicity.[1] They concluded that these lengthy and complex sonic arrangements are structured like the short phrases and longer themes of human songs.[2]

In follow-up studies spanning four decades, researchers have attempted to determine the set size of these acoustically distinct units using a variety of manual and automated analytic methods and statistical classification techniques.[3] Depending on the source data and methods used in multiple studies, the estimates of the number of distinct units varies from twelve to more than one-hundred.[4 5 6] Based upon these unit set-size assessments, other studies have calculated the information content of humpback songs to be significantly less than the information entropy measurement of human language messages.[7 8] In conjunction with the extreme tonal quality of song units, these results reinforce the human perception that humpback songs are of a musical rather than a speech-like nature. A few studies have attempted to characterize the sub-unit structure of song units utilizing spectral segmentation, extraction, and classification methods.[6 9 10a] Others have attempted to identify the syntactical and "cultural" aspects of song structure and transmission.[10b] But if humpback songs are indeed an expression of communication, how can one accurately classify these units without attempting to understand their "phonetic" or sub-unit architecture in relation to the fundamental principles of digital communication and information theory? Since the characterization, classification, and the number of distinct units varies substantially from study to study, we propose that a systematic analysis of whale songs should rely on a communication-theory-based procedure for the extraction and classification of sub-units.

Using Fast Fourier Transform (FFT) spectral analysis, the FFT-derived time-frequency, log-power spectrograms, i.e. waterfall plots, of streaming voiced human speech and humpback whale vocalizations were generated and compared. The striking similarities of the pitch- and formant-structured time-frequency contours indicated a sub-structural basis of humpback vocalizations similar to human speech. Visual assessment of waterfall plots of songs recorded in Maui during the 1984 breeding season also suggested that the song unit set size is potentially much larger than current assessments.[11] This paper describes the formulation, testing, and analysis results of a methodology for the segmentation, extraction and classification of the time-frequency profiles characterizing these sub-structural elements.

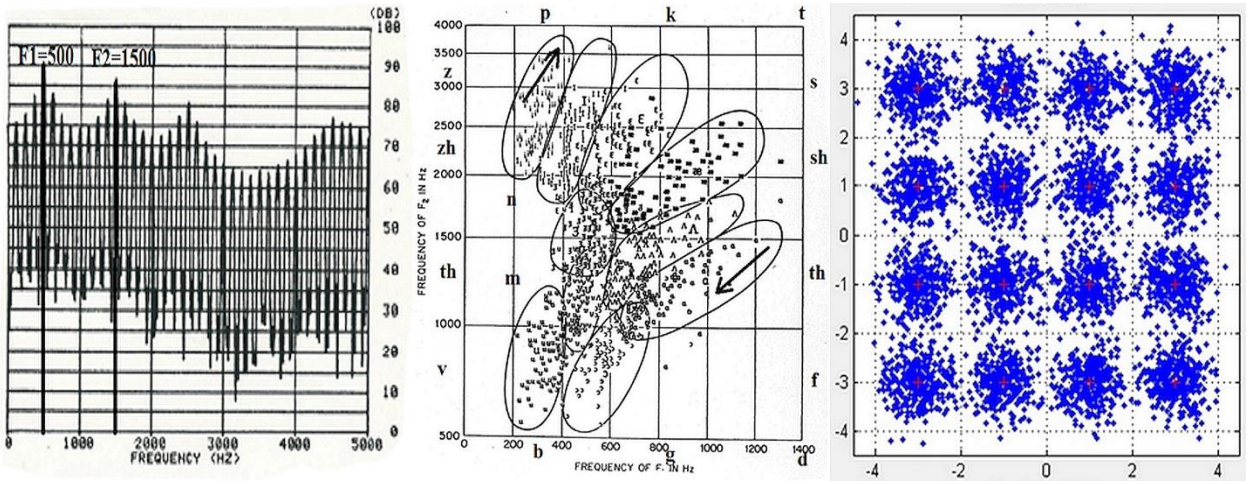
Human Speech Production: English Language Vowels

Several recent spectrographic studies comparing humpback whale song vocalizations to human voiced speech have concluded they share similar acoustical and articulatory characteristics: a frequency-modulated series of harmonics generated by vibrating membranes either filtered or nulled by a resonant vocal chamber.[12 13]Since all whale song data analyzed in this study has distinct harmonic structure, this report will focus on the physiology and acoustics of vowel generation.[11]

Vowel sounds are produced when the vocal cords vibrate at a fundamental frequency, aka 'pitch,' and a series of harmonic vibrational modes which are integral multiples of the fundamental.[14] The pitch harmonics are filtered by the resonance frequencies of the vocal tract composed of the throat, mouth, tongue, lips and nasal passages. Changing the shape and the dimensions of the vocal tract generates the differential resonant frequency combinations specific to each vowel. A time-snapshot plot of the log-power spectrum for a typical vowel, Fig. 1A, shows the saw-tooth structure of the vocal cord harmonics being filtered by the envelope of the vocal tract's frequency-response power spectrum.

The power spectrum of human vocal cord harmonics decreases by about 6 Db per octave with increasing frequency.[14] The filtered harmonics corresponding to the two lowest vowel resonant center frequencies, F1 and F2 in Fig. 1A, aka *formants* or *tone-pairs*, are therefore typically more energetic than harmonics filtered by higher resonant frequencies. The different F1, F2 tone-pair combinations thus serve to acoustically and perceptually distinguish one vowel from another. This is evident in Fig. 1B, a plot of the first two formant resonant center frequencies for the ten primary English language vowels for a wide range of speakers, where the x-axis represents

the range of frequencies of the lower frequency formant and the y-axis plots the frequency range of the higher frequency formant. [15]



(A) **(B)** **(C)**

Figure 1A. Frequency vs. Log-Power spectrogram of an English language vowel. Characterized by the Resonant Tone-Pair (F1,F2) = (500,1500) Hz. X-axis = frequency (Hz). Y-axis = log-power (DB).

Figure 1B. X-Y scatter plot of the first two vocal tract resonant center frequencies (F1, F2), aka *formants*, for the ten primary English language vowels for a wide range of speakers. X-axis = F1(Hz). Y-axis = F2(Hz). Arrows = Mandarin tonal vowels. Low, medium, high frequency consonants depicted on plot perimeter. Reproduced from Figure 8 of *The Journal of the Acoustical Society of America*, **24 (2),** 175 (1952) - Peterson and Barney, "Control Methods Used in a Study of the Vowels," with the permission of the Acoustical Society of America.[15]

Figure 1C. MATLAB©-simulated sixteen-symbol QAM, Quadrature Amplitude Modulated, scatter plot of received symbols corrupted by Gaussian channel noise. Reproduced from Mathworks online documentation site with the permission of Mathworks Education Marketing.[18]

Digital Communication Theory - Core Symbol Set - Vowels

According to the Shannon-Hartley theorem, the maximum rate at which information can be transmitted essentially error-free in a communication channel of fixed bandwidth, like a phone wire, the air between two human speakers, or through the water separating two humpback whales, is limited by the ratio of average signal power to the channel's average Gaussian noise level, i.e., SNR.[16] To approach this theoretical throughput limit, digital modem designers select signal modulation techniques, e.g. frequency modulation, and optimized communication symbol sets, like the symmetrically distributed sixteen dual-tone frequency symbols used in DTMF touch-tone phone signaling.[17] The effect of channel noise on SNR and transmitted modem symbols is illustrated in the Fig. 1C MATLAB©-simulated sixteen-symbol QAM, quadrature amplitude modulated, scatter plot of received symbols corrupted by Gaussian channel noise.[18]

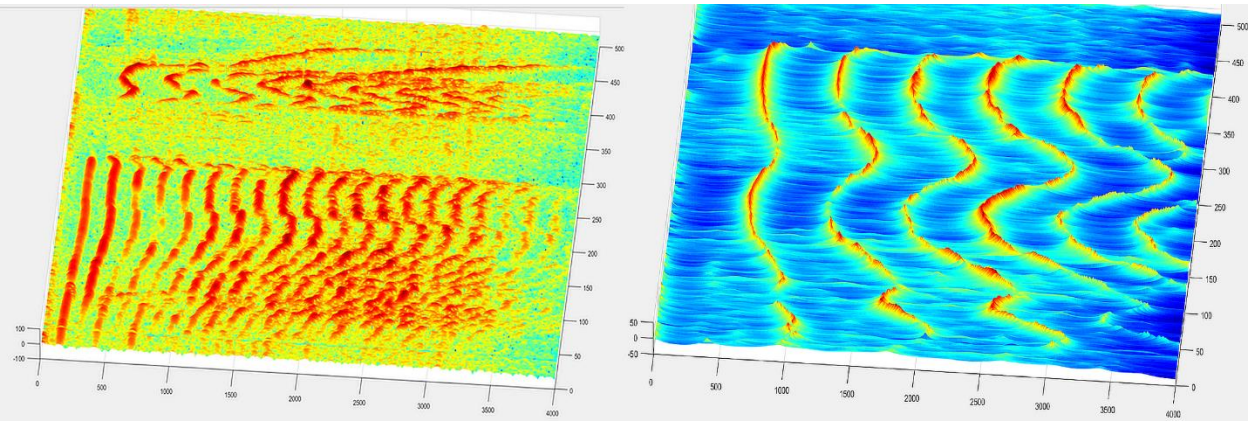
In the Fig. 1B plot of F1, F2 tone-pair data points representing the ten primary English language vowels, the points are clustered into nearly equally spaced, essentially non-overlapping elliptical regions which span the available bandwidth of the communication channel, in compliance with the criteria for an optimally designed modem constellation symbol set illustrated in Fig. 1C. The dimensions of the elliptical regions specific to each vowel are inversely proportional to the SNR, i.e., the ellipses grow larger in response to increased speaker production variations and ambient channel background noise. The maximum number of primary English language vowels is therefore limited by Shannon to the ten regions depicted in Fig. 1B.

English vowels are differentiated by the fixed pair of resonant frequencies in Fig. 1A and generated by the shaping of the vocal tract, irrespective of a speaker's fundamental vocal cord vibrational frequency. Tonal languages such as Mandarin and Cantonese, however, generate even more perceptually distinct combinations of a particular vowel by coupling the excitation of the vocal tract with rising and/or falling pitch intonation induced by dynamic adjustments of cord tension. Up to four different pitch-intonated vowel combinations can be produced for each specific shape of the vocal tract.[20] Cantonese language speakers generate even more pitch-inflected vowel

combinations by causing the pitch to rise or fall between multiple levels of low, medium, and high fundamental frequencies. Two examples of pitch-inflected Mandarin vowels are illustrated as either an up or down "pitch transition vector" embedded in the tone-pair sub-regions in Fig. 1B.

Humpback Sound Production

A recent study proposed a new theoretical acoustic model of humpback sound production "totally different from the human vocal model. The humpback whale sound generator involves vocal folds located in the larynx, which can be set into oscillations in both directions by an airflow powered either by the lungs or by the laryngeal sac."[13] Perhaps this bi-directional model, with the vocal folds mediating airflow between the lungs and the laryngeal sac, explicates the nature of the alternating spectral structure of streaming song units analyzed in this study.[11] Fig. 2A illustrates these distinctive differences in the FFT spectral waterfall plots of two consecutive units. The spectrogram of the three-second unit in the foreground exhibits the frequency-modulated characteristics of human voiced speech: nearly constant pitch in the 100 to 200 Hz range and a series of harmonics with a formant-shaped time-frequency modulation profile. The one-second unit in the background is radically different in that the fundamental pitch frequency is much higher and varies over an extreme range. Since approximately half of the units comprising this study's Maui whale song data exhibit these extreme-pitch-modulated time-frequency contour profiles and, for reasons which will become apparent, this investigation will focus on the latter type of vocalization.



(A) **(B)**

Figure 2A. Five-second time-frequency, log-power, FFT (Fast Fourier Transform) spectrogram of humpback vocalizations. Comparison of a three-second, constant-pitch, "formant modulated" song unit in the foreground vs. a one-second pitch-varying unit in the background.

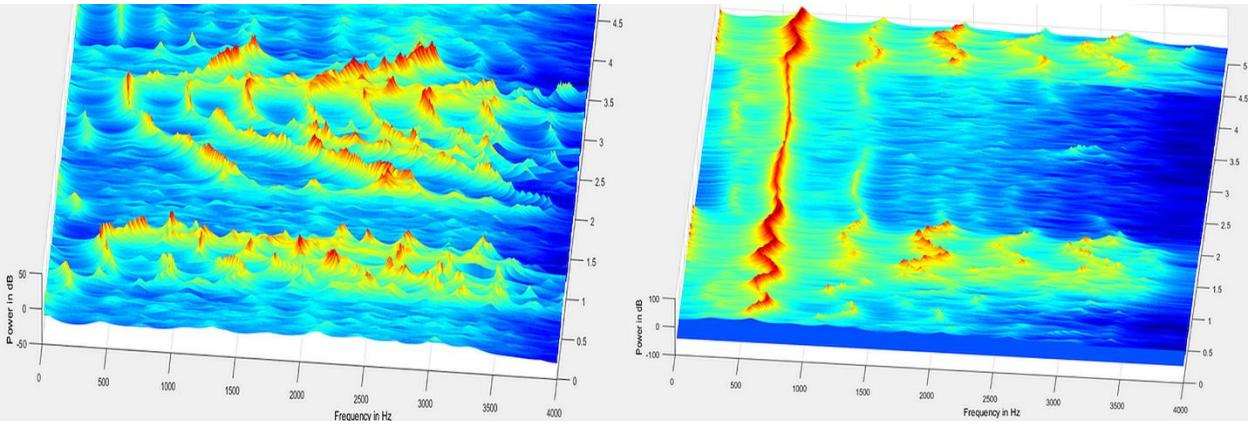
Figure 2B. Five-second time-frequency, log-power, LPC (Linear Predictive Coding) spectrogram of a single complex humpback whale song unit. An example of rising and falling pitch variation over a wide frequency range and the ability to attenuate pitch harmonics, including the fundamental.

Humpback Vocalizations, Human Vowels, and Symbol Target Frequencies

Although little is known about the anatomy and physiology of humpback whale vocalizations,[12] much can be deduced from time-frequency spectrographic analysis. Comparing and contrasting the time-frequency, log-power spectrograms generated by 30-pole Linear Predictive Coding (LPC) analysis of human and humpback whale vocalizations, humpback whales can produce as rich a variety of acoustically distinct frequency-modulated sounds as human speakers. In the collection of waterfall plots of Fig. 2 and Fig. 3, it's apparent that humpbacks have the unique ability to vary pitch over a wide frequency range up to a maximum of nearly four octaves during a time interval ranging from 0.2 to over 2.0 seconds. This dramatic pitch-ranging capability enables a wide range of frequency combinations corresponding to the ratio of the fundamental to the first, second, and higher harmonic frequencies. The humpback's mechanism of extreme pitch variability potentially serves the same function employed by human speakers who change the ratio of the vocal tract resonance frequencies to generate acoustically and perceptually differentiated voiced phonemes. It's also apparent from Fig. 2 and Fig. 3 that humpbacks utilize the four types of constant and variable pitch intonations that characterize Asian tonal languages.

The waterfall plots of humpback vocalizations also indicate that the time-duration during which the harmonic

ratios are sustained is much longer than the time durations of the phonation of the resonant frequency ratio corresponding to a specific English language vowel. This applies to both the constant pitch and pitch-varying voiced modes of humpback vocalizations. This is illustrated in Fig. 3B, where the constant pitch segment is sustained for nearly three seconds and the two alternating, rising-and-falling pitch-varying segments, akin to Mandarin, are maintained at a constant average pitch for nearly a second. In human speech, however, the target frequencies of vowels and consonants are sustained for only a very small fraction of a second. Further discussion about this time-dependency relationship and its impact on interpreting the data derived by the sub-unit symbol classification analysis is described in the "Results" section of this study.



(A) **(B)**

Figure 3A. Five-second time-frequency, log-power, LPC (Linear Predictive Coding) spectrogram of a humpback vocalization. Dramatic pitch-ranging capability enables rising, falling, and constant pitch-harmonic segments and harmonic-frequency combinations corresponding to the ratio of the fundamental to the first, second, and higher harmonic frequencies.

Figure 3B. Five-second time-frequency, log-power, LPC (Linear Predictive Coding) spectrogram of a humpback vocalization. An example of rising and falling pitch variation and the ability to null the first and second pitch harmonics.

The tone-pair target frequencies associated with each English language vowel can be precisely determined at the inflection and start/end points in the time-frequency contour profiles of waterfall plots. This is apparent in Fig. 4A, an F1 vs. F2 contour profile plot of the tone-pair frequencies for the half-second utterance of the phrase *a year*, where the vowel target frequencies appear at the "begin" and "end" points and at intermediate locations of pronounced curvature in the time-frequency contour profile. At each such inflection (edge) point, the course trajectory of the contour profile "pivots" in the direction of the next tone-pair frequency target. Contour profiles like Fig. 4A also result when plotting the sequence of transitional and target baseband symbols transmitted by a digital modem using the 16-QAM coding scheme shown in Fig. 1C.

The similarities of the time-frequency contour profiles, i.e. start, end, and edge points, in the human and humpback waterfall plots, Fig. 2, Fig. 3, and Fig. 4, suggests that humpbacks might also utilize target-frequency-modulated communication symbols in a manner similar to human speech and digital modems. The relationship between symbol target-frequencies and the endpoint and inflection-point regions characterizing the time-frequency-log-power contour profiles in Fig. 2, Fig. 3, and Fig. 4, suggests a methodology for the extraction of tone-pair target frequencies.

Methods

Recordings

The humpback song data for this study was obtained from Maui's Pacific Whale Foundation (PWF).[11] The hydrophones and recording equipment used by the PWF reportedly had linear frequency responses from 50 Hz to at

least 10,000 Hz.[21] The audio feed of the twenty-minute PWF whale song recording was digitized at the Nyquist rate of 8000 samples per second for 0- to 4-kHz audio bandwidth and stored in .wav file format.

Data Extraction Methodology - Goals

A strategy devised to map the phonetic structure of human and humpback vocalizations, a time-frequency-log-power contour segmentation, extraction, and classification procedure, based upon the identification of the targeted "tone-pair" resonant center frequencies, F1 and F2, of English language vowels in a continuous stream of primarily voiced speech, was developed and tested. The same procedure was applied to the extraction of the two most-energetic tone-pair, pitch-harmonic target frequencies of voiced humpback whale vocalizations. The tone-pairs are plotted in X-Y coordinates and subjected to data-clustering analysis to identify postulated tone-pair symbol sub-regions similar to the Fig. 1B and Fig. 1C plots of English vowel and modem-symbol tone-pairs.

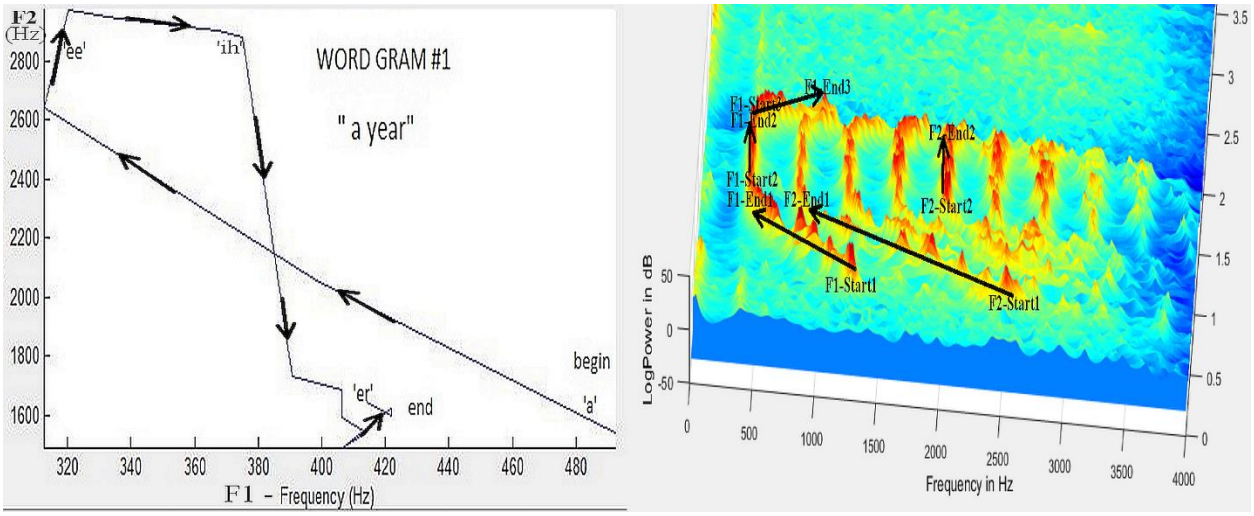
Data Input and Spectral Analysis

Streaming 25-second sections of the digitized .wav file data were buffered to the author's MATLAB©-coded app for subsequent spectral analysis and peak frequency detection. Each twenty-five second data buffer was processed in five-second increments by the MATLAB© program. To capture the humpback's wide-ranging pitch harmonic structural details, a 1024-point FFT algorithm generated frequency-log-power spectral data every 10 milliseconds over 40-millisecond 320 sample-wide hamming-windowed frames with 75% overlap between successive windows. Every fourth 10-millisecond frequency-log-power spectral profile was buffered for subsequent analysis, i.e. 25 spectral profiles per second. A JPEG file of the LPC-derived time-frequency waterfall plot of each five-second segment was stored for subsequent review and verification purposes. All LPC waterfall plots were generated by standard MATLAB© plotting functions. Fig. 2, Fig. 3, and Fig. 4 were generated by the author's MATLAB© app.

For each of the 125-per-five-second buffered FFT spectral profiles, the bin center frequencies F1, F2, F3 and the corresponding log-powers G1, G2, G3 of the three most energetic harmonics were computed and stored for subsequent processing by the sub-unit endpoint and edge extraction algorithm.

Example of Endpoint and Edge Data Extraction and Sub-Unit Segmentation/Pre-Classification

Sub-unit pre-classification is based on the four types of constant, rising, falling, or rising and falling pitch variation along the time-frequency contour profile segments which span target frequencies. An illustration of the procedure for the extraction of start, edge, and endpoint tone-pair and single-tone target frequencies is depicted in the Fig. 4B waterfall plot of a 1.5 second song unit composed of three distinct sub-units:



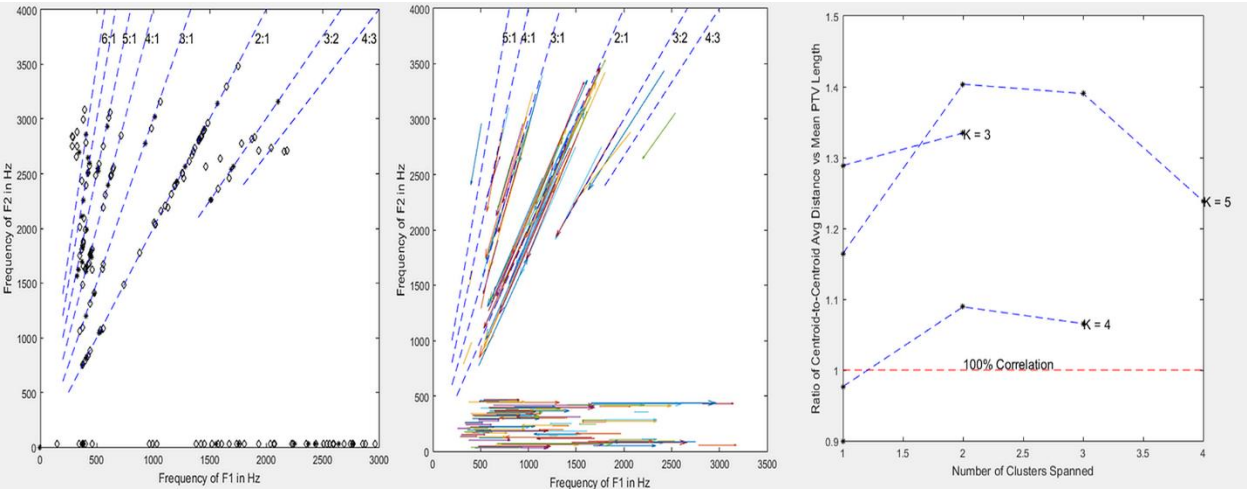
(A) **(B)**

Figure 4A. F1 vs. F2 time-frequency contour plot for 0.5 second, all-vowel utterance of "a year." Vowel target frequencies appear at the beginning, end, and locations of pronounced curvature in the frequency contour profile.
Figure 4B. An example of the procedure for sub-unit pre-classification of a 1.5 second song unit from the extracted start, edge, and end-point tone-pair and single-tone frequencies.

Sub-unit 1. The tone-pairs ($F1_{Start1}$, $F2_{Start1}$) and ($F1_{End1}$, $F2_{End1}$) correspond respectively to the tail and head of the decreasing-frequency PTV sub-unit spanning the tone-pair fundamental and first-harmonic target frequencies (1250,2500) Hz → (425,850) Hz.
Sub-unit 2. The tone-pairs ($F1_{Start2}$, $F2_{Start2}$) and ($F1_{End2}$, $F2_{End2}$) correspond to the 0.5 second duration constant pitch sub-unit whose target frequencies (400,2000) Hz correspond to the fundamental and fourth harmonic. Note: The discontinuity in the transition from the end of sub-unit 1, ($F1_{End1}$, $F2_{End1}$), to the onset of sub-unit 2, ($F1_{Start2}$, $F2_{Start2}$), is captured by the edge-detection algorithm.
Sub-unit 3. The single-tone extracted target frequencies, $F1_{Start3}$ and $F1_{End3}$, correspond to the tail and head respectively of the single-tone rising-frequency PTV sub-unit spanning frequencies 500 → 1000 Hz. From an algorithmic perspective, $F1_{End2}$ and $F1_{Start3}$ are duplicate single-tone points sharing the same extracted edge point frequency. Edge points almost always demarcate the transitional boundary between contiguous sub-units within a unit.

Results

Applying the target-frequency extraction algorithm to the 130 voiced units in the Maui recording, Fig. 5A is a scatter plot of all extracted tone-pair and single-tone data points and pre-classified sub-units differentiated by constant or small-amplitude ("pitch-wiggle") pitch generation. Fig. 5B plots all *pitch transition vector,* aka PTV, sub-units that span extracted target-frequencies separated by more than 100 Hz.[20]



(A) **(B)** **(C)**

Figure 5A. Scatter plot of extracted tone-pair and single-tone sub-unit frequencies. Δ = Constant Pitch, * = Pitch Wiggle. Single-tone frequencies along x-axis, and quantized harmonic-ratio slope lines, e.g. Ratio 2:1 = first-harmonic:fundamental, indicated as dashed --- lines.
Figure 5B. Extracted dual-tone and single-tone pitch transition vectors, aka PTVs. Single-tone vectors horizontally along x-axis and quantized harmonic-ratio slope lines indicated as dashed --- lines. Rate 2:1, 3:1, and single-tone vectors are randomly offset with respect to the slope lines for improved visualization.
Figure 5C. Ratio of K-Means-computed centroid-to-centroid distances vs. mean PTV lengths for K = 3,4,5.

Symbol Classification Methodology – Linear Regression and Harmonic Frequencies

In order to identify and classify sub-unit symbol regions from the extracted data points, linear regression and data clustering methods are essential tools in the determination of statistically significant patterns in X-Y scatter plot data.[22] In the plots of extracted target-frequencies, Fig. 5A and Fig. 5B, tone-pair frequency data is delineated into banded regions that span the same 0- to 4-kHz frequency range as the two primary English vowel resonant center frequencies, F1 and F2, in Fig. 1. Most of the data points straddle the three primary frequency bands which transect the plot from lower-left to upper-right and their slopes correspond to the ratios of pitch harmonic frequencies to the fundamental. The 2:1 slope of the right-most band corresponds to the ratio of the frequency of the first pitch harmonic to the fundamental. Likewise, the 3:1 slope of the middle band reflects that it is populated by tone-pairs

whose ratio is the second pitch harmonic to the fundamental. The left-most band is composed of tone-pairs of ratios 4:1, 5:1, and 6:1. The three slope lines, 2:1, 3:1, and 5:1, respectively pass through the center lines of each band.

In Fig. 5B it's apparent that the PTVs constitute the backbone of the banded regions centered on the 2:1 and 3:1 slope lines and harmonic ratios are preserved in the PTV transition between the "start/tail" and the "end/head" tone-pairs. Since harmonic ratios remain constant during the generation of all tone-pairs, i.e. tone-pairs are quantized in ratios of 2:1, 3:1, … etc., plotted data points should coincide with the quantized slope lines. However, due to measurement errors introduced during the spectral analysis and data extraction processes, the data is statistically distributed with respect to the slope lines as described in the supplementary materials documentation.

Data Clustering Considerations

Within each of the three primary frequency bands in Fig. 5A and Fig. 5B, the tone-pair data points appear to be clustered into sub-regions whose centroids straddle the 2:1, 3:1, and 5:1 slope lines. The task of mapping sub-regions is comparable to a phonetician challenged to delineate the vowel cluster regions "hidden" in the seemingly random tone-pair scatter-plot data in Fig. 1B. Statistical data clustering algorithms are crucial to the identification of centroids and regional boundaries for datasets of arbitrary complexity. However, information about the time-frequency contour profiles at the extracted edge and endpoint target frequencies can assist the identification and mapping of postulated Shannon-constrained sub-regions like those depicted in Fig. 1B.

If the distribution of tone-pair data into sub-regions is guided by a coding process such that the PTVs in Fig. 5B deterministically span sub-unit target regions, then the centroids of the sub-regions "hidden" in the Fig. 5A scatter plot data are theoretically located near the start and end points of these PTVs. More specifically, the lengths of PTV's should correlate with the mean distances separating sub-regional centroids. As described in the next section, this assumption is instrumental in the formulation of a cross-correlational data-clustering strategy which exploits the relationship between the sub-regional centroids and the PTV's which span them.

K-Means Data Clustering Analysis – Target Frequencies and PTV Lengths

K-means data clustering and silhouette algorithms were applied to the statistical analysis of *all* extracted tone-pair and single-tone data points plotted in Fig. 5A and Fig. 5B. K-means, or Lloyd's algorithm, is popular for cluster analysis in data mining applications where *n* raw data observations are partitioned into *K* clusters in which each observation belongs to the cluster with the nearest mean.[22] K-means analysis can generate an estimate of the centroids and boundaries of these data 'clusters' corresponding to sub-unit symbol regions. Silhouette clustering evaluation analysis uses clustering data and silhouette criterion values to evaluate the optimal number of data clusters.[29]

An unsupervised machine learning algorithm such as K-means cannot be expected to accurately determine the centroids and boundaries of the complete set of sub-unit regions spanning the entire tone-pair dataset plotted in Fig. 5A from a single batch analysis. However, since all tone-pairs and PTVs within each of the three primary frequency bands share the same quantized harmonic ratio, i.e. 2:1, 3:1, etc., these three regions can be considered independently of one another. Therefore, a more manageable 'semi-supervised' approach is to break the problem down into three independent K-means clustering sub-tasks analyzing only the data contained within each of the three primary frequency bands. All K-means analysis for this study used MATLAB©'s *KMEANS* standard function.

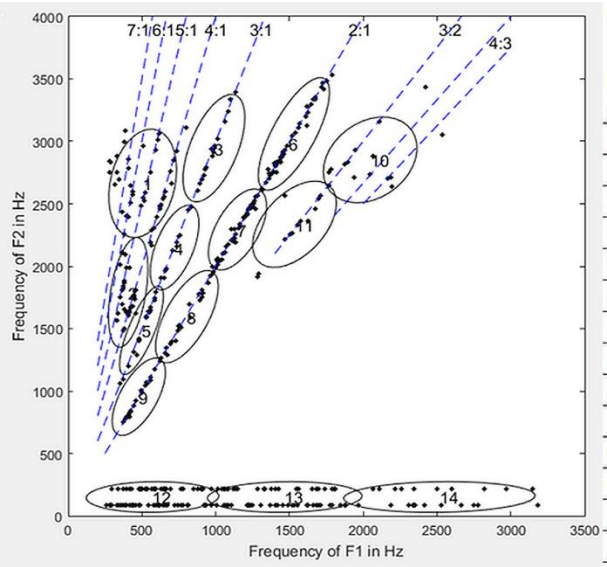
Since K-means measures the average variance in the data spread with respect to the centroid in each cluster, a standard procedure for determining the optimum number of clusters in each dataset is the *Elbow* method.[23] The optimum number of clusters occurs at the 'elbow' inflection region in the variance curve since adding more clusters results in only marginal gains in variance reduction.

The metrics generated by K-Means-elbow and silhouette-clustering analysis of the extracted data points with an F2/F1 ratio of 2:1 are inconclusive as to whether the optimum number of clusters corresponds to a value of K=3, 4, or 5. If there is a relationship between sub-regions and the lengths of the PTV's which span them, we formulated a cross-correlational data-clustering technique to determine the value of K which best correlates the K-Means-computed, centroid-to-centroid, sub-regional distances with the K-Means-computed mean PTV lengths spanning these sub-regions. Fig. 5C is a plot of the mean PTV lengths as a percentage of the mean inter-regional, centroid-to-centroid distances for values of K=3,4,5. It's clear that K=4 is the best match compared to the 100% correlation, straight-line plot and a strong indication that PTV's span centroid-to-centroid sub-regional distances.

Applying the same cross-correlational procedure to the tone-pairs in the frequency band straddling the 3:1 slope line and to the single-tone data "region" generated K=3 clusters as the best match for both regions. Since there was

insufficient PTV data for the regions centered on the 5:1 and 2:3 slope lines, K-Means silhouette clustering evaluation analysis generated optimum values of K=2 clusters for both regions. Data and plots for these cases is contained in the supplementary materials documentation.

All sub-unit centroid and sub-regional boundary data derived by K-means clustering analysis was incorporated into the Fig. 6 comprehensive plot of the fourteen sub-unit regions spanning the Fig. 5 scatter plots of all extracted start, edge, and stop tone-pairs and single-tone frequencies. A summary of the K-means-derived analysis parameters for all fourteen sub-regions is listed in Table 1.



Sub-Unit Region #	Sub-Unit Centroid F1,F2 (Hz)	Sub-Unit Centroid F2/F1 Ratio	Major Axis Length (Hz)	K-Means Per-Point Avg. Std. Dev.(Hz)	Adjacent Sub-Unit #'s	Adjacent Sub-Unit Slope Line	Adjacent Sub-Unit Boundary F1,F2(Hz)
1	534,2580	4.835	880	267	2,1	4:1, 5:1	428,2178
2	388,1737	4.480	890	189		4:1, 5:1	
3	962,2886	3.001	900	237	3,4	3:1	840,2516
4	709,2125	2.997	710	202	4,5	3:1	589,1769
5	466,1412	3.031	750	229		3:1	
6	1533,3055	1.993	900	270	6,7	2:1	1363,2727
7	1212,2426	2.002	700	223	7,8	2:1	1063,2133
8	906,1809	1.996	800	226	8,9	2:1	696,1393
9	489,980	2.006	670	179		2:1	
10	2135,2866	1.342	750	297		3:2, 4:3	
11	1567,2441	1.557	780	276	10,11	3:2	1822,2673
12	599*		950	211	12,13		982
13	1391*		1050	262	13,14		1890
14	2446*		1100	285			

Figure 6. Fourteen sub-unit regional centroids and boundaries derived by K-Means analysis of all extracted Start, Edge, and Stop tone-pairs and single-tone frequencies – denoted by *. Centroid Locations marked by region number symbols '1' to '14'. Single-tone frequencies along x-axis. Data error-bar is [+-]8 Hz orthogonal to slope lines and [+-]28 Hz parallel to slope lines. Details in author's Open Science Framework project directory.
Table 1. Centroid and regional parameters for all fourteen K-means-computed sub-regions. (*=Single-Tone)

Articulatory and Acoustic Basis of Dual-Tone Symbol Set Structure

Direct comparison of Fig. 1B and Fig. 6 reveals the similarity of human phonetic and humpback sub-unit symbol structures: approximately a dozen optimally spaced dual-tone symbol regions to maximize available channel bandwidth.[16] Humans span the dual-tone bandwidth range in Fig. 1B by vocal tract resonant frequency shaping. Humpbacks achieve similar results in Fig. 6 using two distinct articulatory/acoustical mechanisms: extreme pitch variability in combination with the attenuation and/or nulling of specific harmonic frequencies, including the fundamental. By comparing the X-axes of Fig. 1B and Fig. 6, these mechanisms enable humpbacks to generate a set of quantized harmonic-ratio tone-pairs to exploit bandwidth availability beyond the range of English speakers.

Within each of the three-primary tone-pair frequency bands in Fig. 5 and Fig. 6, humpback fundamental pitch and harmonic frequencies vary over a wide range with the fundamental ranging from 325 Hz to nearly 2000 Hz. This extreme pitch ranging capability enables humpbacks to generate tone-pair frequency combinations that span the entirety of and partition each band into perceptually distinct sub-unit symbol regions. For example, the center frequencies or centroids of the four sub-regions straddling the 2:1 slope line are listed in Table 1. These sub-regions are generated by the ratio of the fundamental and first harmonic frequency for four distinct frequency regimes centered at these values of the fundamental: low=489 Hz, medium-low=906 Hz, medium-high=1212 Hz, and high = 1533 Hz. The two to three sub-regions comprising each of the other bands are generated in a similar fashion.

The three primary frequency bands spanning Fig. 5 and Fig. 6 correspond respectively to the F2:F1 tone-pair ratio of vocal fold harmonics: ratio 2:1 right-most band, ratio 3:1 middle band, and ratios 4:1, 5:1, 6:1 comprising the left-most band. The time-frequency contour profiles of humpback waterfall plots in Fig. 3 illustrate how tone-pairs populating the ratio 3:1 band are generated by the nulling of the first pitch harmonic. Similarly, the time-frequency contours of Fig. 2B and Fig. 4B illustrate how tone-pairs occupying the left-most band straddling the rate

4:1, 5:1, and 6:1 slope lines result from the nulling or attenuation of the second, third, or fourth harmonics. Indeed, since the data points comprising the orphan sub-unit regions in the right-most section of the Fig. 6 plot have an F2:F1 ratio of 3:2 or even 4:3, it appears that humpbacks are capable of nulling even the fundamental pitch frequency to create additional sub-unit symbol regions.

Dual-Tone and Single-Tone Core Symbol Set Classification Referenced to Fourteen Symbol Sub-Regions

Using the articulatory and acoustical mechanisms described in the previous section, humpback whales generate tone-pair harmonic ratios to construct the eleven dual-tone symbol sub-regions spanning the interior of Fig. 6 and the three single-tone sub-regions at the bottom of the plot. These fourteen sub-regions form the basis of the humpback's core, constant-pitch, sub-unit symbol set.

Expansion of the humpback's core symbol set is achieved by utilizing the four distinct types of constant- and variable-pitch to generate numerous intra- and inter-regional pitch transition sub-units akin to Mandarin and Cantonese vowels. The shorter intra-regional PTVs in Fig. 5B are similar to the Mandarin language's rising and/or falling pitch-intonated vowels represented by the two vectors plotted in Fig. 1B. The longer inter-regional PTVs are akin to the Cantonese and Vietnamese generation of vowels by rising or falling pitch between different levels of low, medium, and high pitch.[24] Compared to Cantonese, however, some of these humpback pitch-intonated transitions are spectacular glissando-like slides across multiple regions, utilizing the maximum bandwidth available by spanning the entirety of the 2:1 and 3:1 slope lines. Just as only some vowels in Mandarin and Cantonese are pitch-intonated, i.e. tonal, only about half of the humpback's potential intra- and inter-regional fixed-frequency, single-tone and dual-tone regions are populated and/or linked by PTVs.

Using pitch intonation, humpbacks have more than tripled the core set size of fourteen single-tone and dual-tone constant-pitch voiced sub-unit symbols to a total of 60 sub-units. Table 2A lists the complete set of sixty derived sub-unit symbols in relation to the four types of pitch transitions referenced to the fourteen single- and dual-tone sub-regions in Fig. 6 and Table 1.

A) 1C	B) 2C	C) 3C	D) 4C	E) 5C	F) 6C	G) 7C
H) 8C	I) 9C	J) 10C	K) 11C	L) 12C	M) 13C	N) 14C
O) 1W	P) 2W	Q) 3W	R) 4W	S) 5W	T) 6W	U) 7W
V) 9↓	W) 9W	X) 10W	Y) 11W	Z) 12W	α) 13W	Γ) 14W
θ) 1↑	Δ) 5↑	a) 6↑	b) 2↑	c) 12↑	d) 13↑	e) 14↑
f) 1↓	g) 3↓	h) 12↓	i) 7↓	j) 10↓	k)10→11	l) 13↓
m) 14↓	n) 1→2	o) 3→4	p) 3→5	q) 6→7	r) 6←7	s) 7→8
t) 4→5	u)12←13	v) 8←9	w) 6→8	x) 6←8	y) 7→9	z) 6→9
Ψ)6←9	Π)12→13	Φ)13→14	Ω)13←14	λ) BB		

(A)

soAD	ofAAN	BΠofAJ	PcG	iRoD	jJc	wkFr
BP	aNq	YΠ	MNN	PΦe	Pr	SΠ
OΠ	lΠ	Wx	PrΦ	NABAB	OrM	Px
UL	PΦNΩ	PE	NΦNeq	BSc	ΦmN	UslEJc
LLZ	MαNu	BWAc	MFHJ	ZAc	GJC	ZAc
dMl	ZAc*	GJl	ALO	lFqAc	LPΔ	yAc
λLAc	uIII	λWc	FlpIc	cIc	yBc	ΓAc
zAc	GdlM	MAPc	MFy	MFy*	lNl	aMNw
lMΦΓw	Xkfv	iGs	Bbc	kqOΠ	Fu	stAvd
Tu	PHc	ySc	αN	LBc	qtOc	TGL
LEΠ	uConOv	Oc	ipJθ	KBM	gkAAΓcGo	QYoOYA
BΠ	esA	KcG	NLtEΠ	kcGd	qhtJΠ	WEcM
FG	YΠ	lΠFd	BlxΦ	NLGΨN	WEΠ	LAΦ
BKGaΦ	LNrFe	PNrΦN	TwOΠ	wΓrMFG	GIF	aMFMG
FsEBc	whBLc	FghBVc	SLS**	cVlH*^	yBc***	

(B)

Table 2A. Complete Sub-Unit List – "Symbol-Code) Acoustic-Attributes" Per Table Entry

Legend: 1-14 = Region #; C=Constant Pitch; W = IntraRegional Up&Down Pitch Wiggle;
↑ = IntraRegional Rising Pitch; ↓ = IntraRegionional Falling Pitch;
→ = InterRegional Rising Pitch; ← = InterRegional Falling Pitch; BB = Non-Voiced BroadBand

Table 2B. Humpback song unit lexicon. Sub-unit structure of 104 song units from Table 2A sub-unit symbol codes. * = Duplicate Unit *** = Figure 4B Decoded Unit ** = Figure 3B Decoded Unit *^ = Tonga Unit

Dual-tone and single-tone sub-unit symbols classified by mechanisms of symbol set expansion

An inventory of the number and types of acoustically and perceptually distinct constant-pitch and pitch-varying sub-unit symbols, referenced to the sub-unit regions plotted in Fig. 6, is listed below:

1. Core Set: Fixed Frequency – Constant Pitch: Dual-tone: 11 sub-units, Single-tone: 3 sub-units.
2. Intra-regional rising or falling pitch: Dual-tone: 9 sub-units, Single-tone: 5 sub-units.

3. Intra-regional rising-and-falling pitch: Dual-tone: 11 sub-units, Single-tone: 3 sub-units.
4. Inter-regional pitch variation: Dual-tone: 15 sub-units, Single-tone: 3 sub-units.

Discussion

The rationale for classifying PTVs as primary tonal symbols, similar to Mandarin and Cantonese tonemes, rather than secondary transitional paths between two symbol regions, as is the case for non-tonal English language vowels, is indicated by the lengthy time-durations, ranging from 0.2 to 2.0 seconds, of the constant-pitch tone-pairs plotted as 'Δ' in the scatter plot of Fig. 5A. The generation of PTV's also varies over a similar range of time-durations. Humpbacks may therefore need to generate constant-pitch tone-pairs for an extended time, see Fig. 3 and Fig. 4, to differentiate the tonal symbols from the non-tonal symbols. The validity of this assumption and the accuracy of the derivation of the humpback's complete sub-unit symbol set can be verified by statistical analysis of the symbol set's frequency-of-occurrence probability distribution profile described in the next section.

The derived sub-unit symbol set listed in Table 2A enables the 104 song units analyzed in this study to be precisely specified by their constituent sub-unit sequence structure and sub-unit sequence length. Table 2B constitutes a first-of-its-kind "phonemic lexicon" of humpback song units. In a breakdown of the percentage of song units versus sub-unit length, the minimum sub-unit sequence length is two sub-units, the maximum length is 8 sub-units, and the average length is about 3.5 sub-units.

Only two of the more than one hundred units analyzed to date share identical sub-unit sequences. The 102 unique units represent a major increase in the maximum number of unique song units reported in most previous studies. Since 98% of the total units analyzed in this particular recording are unique, it's likely that analysis of additional humpback song data could result in a dramatic expansion in the number of unique units. The rather low song unit redundancy factor is noteworthy, especially in comparison to the English language where two words, "of" and "the," constitute 10% of all word occurrences and 150 of the most common words constitute 50% of the corpus of all English words.[25]

English Symbol (*19*)	Frequency-% English	Cumulative Frequency-%	Humpback Symbol	Frequency-% Humpback	Cumulative Frequency-%
ə	10.0%	10.0%	c	8.79%	8.79%
I	9.8 %	19.7%	A	6.32%	15.11%
n	8.0%	27.7%	N	5.22%	20.33%
t	7.6%	35.3%	B	4.40%	24.73%
r	7.1%	42.4%	M	4.40%	29.12%
s	4.9%	47.2%	F	4.12%	33.24%
l	3.7%	50.9%	G	4.12%	37.36%
đ	3.3%	54.2%	L	4.12%	41.48%
d	3.2%	57.5%	II	4.12%	45.60%
æ	3.1%	60.5%	P	3.30%	48.90%

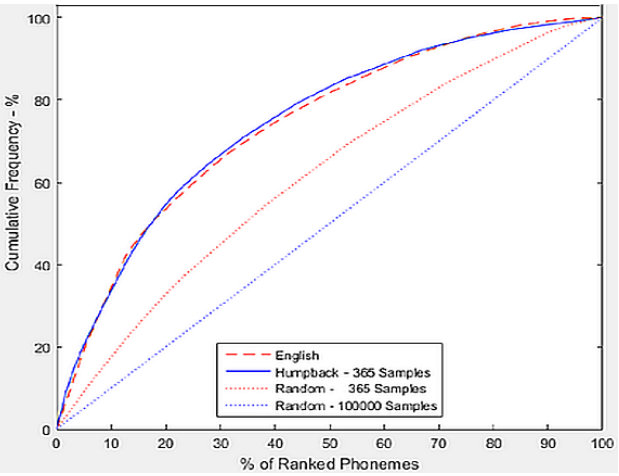


Table 3. Ten highest ranked phonemes by frequency of occurrence: English vs. Humpback. Humpback symbols referenced to Table 2A sub-unit symbol codes.

Figure 7. Cumulative frequency vs. % of 39 ranked English phonemes and 60 ranked humpback sub-unit symbols. The straight-line distribution of a set of equally-likely symbols is shown at the bottom of the figure. The results utilized 365 humpback sub-unit symbol sample occurrences.

Information Theory: Phonetic and Lexical Statistical Analysis

Statistical analysis of the frequencies of occurrence of the 60 sub-unit symbols comprising the 104 song units in Table 2B provides compelling evidence about the accuracy of the extraction and classification procedures in mapping the structure of humpback song units. Table 3 lists a side-by-side comparison of the frequencies of occurrence of the ten most-frequently observed English phonemes and humpback sub-units. Fig. 7 clearly shows that a plot of the humpback sub-unit set's cumulative probability vs. ranked frequency distribution profile is nearly identical to the same plot for English language phonemes. Both profiles conform to the Zipf power law's signature

geometric probability distribution profile for natural language symbol sets exhibiting pronounced information redundancy.[26] This contrasts with the straight-line distribution profile at the bottom of Fig. 7 for a sixty-element symbol set where all elements are equally likely.

These Zipf-like probability distribution curves contain all the statistical data needed to compute the first-order Shannon information entropy values of the complete set of 39 English phonemes and the 39 most-likely elements of the humpback "phonemic" symbol set: 4.71 bits and 4.61 bits respectively.[27] This compares to a Shannon entropy of 5.29 bits for a 39-element set of equally likely symbols. When comparing same-size symbol sets, lower entropy values correspond to increased redundancy and lower information content. The Shannon entropy calculation for the humpback's entire sixty element symbol set is 5.31 bits, compared to a value of 5.9 bits for a sixty-element set of equally likely symbols.

The large sub-unit set size and the affirmation of Zipf's law are compelling evidence for humpback phonetic complexity vis-à-vis human speech. This suggests that the "lexicon" of humpback whale song units is potentially much larger than currently cited in the research literature. The humpback's "alphabet" of 60 voiced sub-unit symbols is comparable to the phonemic set size of human tonal languages like Vietnamese and Burmese and is about 50% larger than the complete set of 39 English language phonemes.[19 24] It's interesting to note that the 59% to 41% ratio of tone-pair to single-tone sub-unit symbol occurrences is comparable to the 62% to 38% ratio of consonant to vowel phoneme usage in American English.[19]

Conclusions

By the joint criteria of digital communication and information theory, humpback whales are precisely modulating frequency to generate a set of perceptually distinct acoustical patterns with the phonetic properties of human speech. In compliance with the Shannon Communication Theorem, humpbacks and humans have each crafted optimized, sixty-symbol, frequency-modulated phonetic architectures (Fig. 1B and Fig. 6). As humans have demonstrated, a sixty-symbol phonemic alphabet exceeds the information entropy threshold to support the construction of a 100,000-plus "word" lexicon. Although the rate of production of sub-unit and unit symbols is slow in comparison to human speakers, humpbacks have encoded a necessary and sufficient set of primitive acoustical symbols required for the expression of speech.

That the tone-pair target frequencies which encode humpback voiced sub-unit symbols are generated by vocal fold harmonic ratios over an extremely wide range of fundamental frequencies, rather than by vocal tract resonance frequencies, imbues humpback song units with a distinct tonal quality. It's not surprising, therefore, that humans are predisposed to perceive humpback vocalizations as an expression of musical communication. This study's findings, however, provide compelling evidence that humpback whale vocalizations exhibit natural language phonetic structure. The phrase and theme structure of humpback songs can be analyzed with enhanced precision now that voiced units can be characterized and differentiated with a sub-unit-level of specificity listed in Table 2.

Side-by-side comparison of human and humpback whale phonetic structural mappings suggests an intriguing question: Which species was the first to manifest the complex phonetic architecture enabling the generation of speech? Since mitochondrial DNA analysis substantiates the presence of modern Megaptera novaangliae in the North Pacific Ocean for at least the past 175,000 to 200,000 years, it's humbling to speculate if humpbacks harnessed sound for the expression of complex acoustic symbols before modern humans.[28]

Acknowledgments: This paper is dedicated to the memory of Greg Kaufman, co-founder and long-time Director of Maui's Pacific Whale Foundation.

Notes: Additional details about the data extraction, post-extraction sub-unit segmentation, pre-classification and data-clustering procedures and results, as well as error-analysis results and dataset sample size sufficiency considerations, are described in the Supplementary Materials document posted at the author's Open Science Framework project directory which is available by request.

References:

[1] R. S. Payne, S. McVay, Songs of Humpback Whales, *Science* **173**, 585–597 (1971).
[2] K. Payne, P. Tyack, R. Payne, "Progressive changes in the songs of humpback whales (Megaptera novaeangliae): A detailed analysis of two seasons in Hawaii" in *Communication and Behavior of Whale,* edited by R. Payne

(Westview, Boulder, CO, 1983), 9–57.
[3] F. Pace, P. White, O. Adam, "Hidden Markov Modeling for humpback whale (Megaptera novaeangliae) call classification," presented at the Soci´et´e Fran¸caise d'Acoustique. Acoustics 2012, Nantes, France, April 2012.
[4] F. Pace, "Comparison of Feature Sets for Humpback Whale Song Classification," MSc dissertation, University of Southampton, UK (2008).
[5] S. Guan, A. Takemura, T. Koido, "An introduction to the structure of humpback whale, Megaptera novaeangliae, song off Ryukyu Islands, 1991/1992", *Aquatic Mammals* **25.1**, 35–42 (1999).
[6] J. Allen *el al*., "Using self-organizing maps to classify humpback whale song units and quantify their similarity," *J. Acoust. Soc. Am.* **142**, 1943-1952 (2017).
[7] P. Suzuki, J. R. Buck, P. L. Tyack, "Information entropy of humpback whale songs," *J. Acoust. Soc. Am.* **119**, 1849-1866 (2005).
[8] J. L. Miksis-Olds *el al*., "Information theory analysis of Australian humpback whale song," *J. Acoust. Soc. Am.* **128**, 2385-2393 (2008).
[9] F. Pace, F. Benard, H. Glotin, O. Adam, P. White, "Subunit definition and analysis for humpback whale call classification," *Applied Acoustics* **71**, 1107-1112 (2010).
[10a] H. Ou, W. Au, L. M. Zurk, M. O. Lammers, "Automated extraction and classification of time frequency contours in humpback vocalizations," *J. Acoust. Soc. Am.* **133**, 301–310 (2013).
[10b] E. C. Garland *el al*., "Song hybridization events during revolutionary song change provide insights into cultural transmission in humpback whales," *Proc. Natl. Acad. Sci. U.S.A.* **114**, 7822–7829 (2017).
[11] Pacific Whale Foundation, "Hawaiian Whale Song – February, 1984 – Maui, Hawaii," PWF, Kihei HI (1984).
[12] E. Mercado, J. Schneider, A. A. Pack, L. M. Herman, "Sound production by singing humpback whales," *J. Acoust. Soc. Am.* **127**, 2678-2691 (2010).
[13] O. Adam *el al*., "New acoustic model for Humpback whale sound production", *Applied Acoustics* **74**, 1182-1190 (2013).
[14] L. R. Rabiner, R. W. Schafer, *Digital Processing of Speech Signals* (Prentice-Hall International, London, UK 1978).
[15] G. E. Peterson, H. L. Barney, "Control Methods Used in a Study of Vowels," *J. Acoust. Soc. Am.* **24**, 175-184 (1952).
[16] C. E. Shannon, "A mathematical theory of communication," Bell Syst. Tech. J. **27**, 379–423, 623–656 (1948).
[17] J. E. McNamara, *Technical Aspects of Data Communication*, (Digital Press, Bedford, Mass., USA 1982).
[18] WWW.Mathworks.Com/help/comm/ug/digital-modulation.html.
[19] R. E. Hayden, "The Relative Frequency of Phonemes in General-American English," *WORD* **6:3**, 217-223 (1950).
[20] P. Keating, G. Kuo, "Comparison of speaking fundamental frequency in English and Mandarin," *J. Acoust. Soc. Am.* **132(2)**, 1050-1060 (2012).
[21] D. A. Helweg *et al*., "Comparison of songs of humpback whales(Megaptera novaeangliae) recorded in Japan, Hawaii, and Mexico during the winter of 1989," *Sci. Rep. Cetacean Res.* **1**, 1-20 (1990).
[22] C. Bishop, *Pattern Recognition and Machine Learning*, (Springer Science + Business Media, New York, USA 2006).
[23] G. James, D. Witten, T. Hastie, R. Tibshirani, *An Introduction to Statistical Learning*, (Springer Science + Business Media, New York, USA 2013).
[24] D. Hwa-Froelich, B. W. Hodson, H. T. Edwards, "Characteristics of Vietnamese phonology," *American Journal of Speech-Language Pathology*, **11**, 264-273 (2002).
[25] W.N. Francis, H. Kučera, *Frequency Analysis of English Usage: Lexicon and Grammar*, (Houghton Mifflin, Boston, USA 1983).
[26] G. K. Zipf, *Selected Studies of the Principle of Relative Frequency in Language*, (Harvard University Press, Cambridge, USA 1932).
[27] C. E. Shannon, "Prediction and entropy of printed English," *Bell Syst. Tech. J.* **30**, 47-51 (1951).
[28] J. A. Jackson *el al*., "Global diversity and oceanic divergence of humpback whales (Megaptera novaeangliae)," *Proc. of the Royal Society B – Biological Sciences*, (2014).
[29] WWW.Mathworks.Com/help/stats/silhouette.html.