HHS Public Access

Author manuscript

Anim Cogn. Author manuscript; available in PMC 2018 May 29.

Published in final edited form as:

Anim Cogn. 2016 March ; 19(2): 285–315. doi:10.1007/s10071-015-0933-6.

The Vocal Repertoire of the Domesticated Zebra Finch: a Data Driven Approach to Decipher the Information-bearing Acoustic Features of Communication Signals

Julie E. Elie and

Helen Wills Neuroscience Institute, UC Berkeley, 3210 Tolman Hall, University of California Berkeley, Berkeley CA 94720

Frédéric E. Theunissen

Department of Psychology and Helen Wills Neuroscience Institute, UC Berkeley, 3210 Tolman Hall, University of California Berkeley, Berkeley CA 94720

Abstract

Although a universal code for the acoustic features of animal vocal communication calls may not exist, the thorough analysis of the distinctive acoustical features of vocalization categories is important not only to decipher the acoustical code for a specific species but also to understand the evolution of communication signals and the mechanisms used to produce and understand them.

Here, we recorded more than 8,000 examples of almost all the vocalizations of the domesticated zebra finch, *Taeniopygia guttata*: vocalizations produced to establish contact, to form and maintain pair bonds, to sound an alarm, to communicate distress or to advertise hunger or aggressive intents. We characterized each vocalization type using complete representations that avoided any *a priori* assumptions on the acoustic code, as well as classical bioacoustics measures that could provide more intuitive interpretations. We then used these acoustical features to rigorously determine the potential information-bearing acoustical features for each vocalization type using both a novel regularized classifier and an unsupervised clustering algorithm. Vocalization categories are discriminated by the shape of their frequency spectrum and by their pitch saliency (noisy to tonal vocalizations) but not particularly by their fundamental frequency. Notably, the spectral shape of zebra finch vocalizations contains peaks or formants that vary systematically across categories and that would be generated by active control of both the vocal organ (source) and the upper vocal tract (filter).

Keywords

vocalization; Songbird; acoustic signature; meaning; classification; regularization

Corresponding Author: julie.elie@berkeley.edu, Tel: +1.510.643.1531, Fax: +1.510.642.5293.

Ethical Approval

All applicable international, national, and/or institutional guidelines for the care and use of animals were followed. All animal procedures were approved by the Animal Care and Use Committee of the University of California Berkeley and were in accordance with the NIH guidelines regarding the care and use of animals for experimental procedures. The authors declare that they have no conflict of interest.

Background

Many social animals have evolved complex vocal repertoires not only to facilitate cooperative behaviors, such as pair bonding or predator avoidance, but also in competitive interactions such as in the establishment of social ranks, mate guarding or territorial defense (Seyfarth and Cheney, 2003). For example, many songbirds use elaborate songs (Catchpole and Slater, 1995) as well as other short calls (Marler, 2004) as part of the courtship behavior and initial pair formation. This initial pair formation can lead to cooperative nest building, territory defense, reproduction, chick-rearing and the formation of a long-term partnership. The maintenance of such a stable pair bond requires close social contact that can be facilitated by vocal communication. For example, loud contact calls that carry individual information play a key role for songbird partners that attempt to reunite after having lost visual contact (Vignal et al., 2008). Similarly, vocal duets of songs (Farabaugh, 1982; Hall, 2004; Thorpe, 1972) or of calls (Elie et al., 2010) performed by partners can act as cooperative displays that signal commitment and reinforce the pair bond (Elie et al., 2010; Hall, 2004; Hall, 2009; Smith, 1994; Wickler and Seibt, 1980). In the context of cooperation for predator avoidance, birds produce alarm calls that can be predator specific (Evans et al., 1993). Finally, in competitive interactions, such as in territorial defense, songbirds can use both songs (Searcy and Beecher, 2009) and specific calls (Ballentine et al., 2008) to advertise aggressive intentions. The social context in which each of these vocalizations is emitted (e.g. affiliative interaction, alarming situation, etc.) can be used to classify vocalizations into behaviorally meaningful categories that define the vocal repertoire.

To provide a large range of information to the receivers, categories of vocal communication signals in social species must be acoustically separable. Animals can exploit two sources of variability for the production of acoustic signals: the acoustic structure of each individual sound element and how these elements are combined into sequences (see Zuberbühler and Lemasson, 2014 for a review in primates). How animals exploit sound variability to code the different categories of meanings is still an open question. Here, we explore how the acoustic structure of each individual sound element encodes meanings in vocalizations. Although these identifying acoustical features (or signatures) of specific vocal categories have been examined extensively in the vocal repertoires of both mammals (e.g. Brand, 1976; Deaux and Clarke, 2013; Kruuk, 1972; Salmi et al., 2013) and birds (e.g. Collias, 1987; Dragonetti et al., 2013; Ficken et al., 1978), a comprehensive and rigorous analysis of these distinguishing features in a given species has been difficult to achieve because of the limited number of acoustic features tested and/or because of the limited size or quality of the datasets. As a result, acoustical signatures for complete repertoires are often described qualitatively from a relatively small number of distinguishing features as experienced by human listeners or as observed visually on their spectrographic representation (but see Fuller, 2014; Stowell and Plumbley, 2014). A more quantitative analysis requires a very large dataset of calls for two reasons. First, animal vocalizations, as characterized by their sound pressure waveform, are inherently highly dimensional: they are represented by a large number of amplitude values in time that jointly can code an infinite number of unique combinations. Describing the sounds without *a priori* assumptions on the nature of the distinguishing features requires a representation that is also highly dimensional and

This is the exact thing that I wanted to do

Text

More explanation on **`highly dimensional feature space`:**

Imagine you're trying to describe the unique sound of an animal. Each sound can be described by several characteristics or parameters, like pitch, volume, duration, etc. When you have many such parameters, you're dealing with a "high-dimensional feature place," which is essentially a complex and detailed way of looking at all these characteristics together.

preferably complete or invertible in the sense of being equivalent to the sound pressure waveform. In order to estimate the distribution of the sounds in this highly dimensional feature space, one would optimally need more samples than the number of parameters that describe each sound; and even if dimensionality reduction approaches are used, many samples are still required. Second, the dataset of vocalizations should include examples from the complete vocal repertoire of as many animals as possible. This sampling is needed in order to properly assess if vocalizations produced during different social interactions indeed form separable acoustical categories and, if so, to obtain reliable estimates of the within category variability. Such a data driven approach results in an unbiased identification of the distinguishing acoustical features among categories and in a rigorous estimation of the maximally achievable discriminability of vocalization types based solely on acoustics.

1

هدف این مقاله :

This text describes a comparison between two different methods used to analyze vocal sounds. The first method, referred to as "assumption free," does not start with any predefined ideas about which features of the sounds are important. In contrast, the second, more traditional method, starts with a specific, small group of sound characteristics (acoustical parameters) that researchers already think might be important for conveying information in the sounds being studied.

We embarked on such a data driven exploration of a complete vocal repertoire for the zebra finch, *Taeniopygia guttata.* We compared the results of this assumption free approach to a more classical approach that investigates the potential role of a small subset of chosen acoustical parameters for determining the information-bearing features in vocal repertoires. First, as ethologists who have studied the vocal behavior of this species in the field and in the laboratory (Elie et al., 2010; Elie et al., 2011; Mouterde et al., 2014), we were interested in generating a detailed description of the complete vocal repertoire of this songbird both to contribute to our knowledge of its natural history and to contribute to the field of animal communication. More precisely, by using quantitative methods and encouraging comparative approaches, we wanted to gain insights into the evolution of vocal communication signals and assess the degree with which such acoustic codes share similarities across species. The idea of universal codes for vocal communication is hotly debated (Hauser, 2002; Seyfarth and Cheney, 2003), but common principles have been found at different levels. At the acoustical level, relationships between the coarse sound attributes and the meaning of vocal signals have been shown to hold in many species (Morton, 1977); for example, affiliative vocalizations are often soft and low frequency sounds, while loud and high-frequency sounds are often alarm vocalizations (Collias, 1987). At the production level, the source-filter mechanism describing the making and shaping of vocal communication calls is also present in many species (Taylor and Reby, 2010). By performing a quantitative analysis of the information-bearing features in the zebra finch vocal signals, we determined to what extent our data support these putative universal principles of animal communication.

2

خلاصه این بخش:

The main goal of their research was to explore how vocal communication has evolved in these birds and to see if the ways birds use sounds to communicate have any similarities across different species. This involves a debate about whether there are universal "codes" or patterns in how all animals communicate vocally, which some scientists argue exist and can be observed across various animal species.

Second, as neuroethologists, we wanted to determine precisely the acoustical features that distinguish vocalization categories in zebra finches in order to then investigate how these could be represented in their auditory system and from there what neural mechanism lead to vocal categorizations (Bennur et al., 2013; Elie and Theunissen, 2015).

Finally, the zebra finch provided a unique opportunity to obtain a very large data set of vocalizations of high audio quality that would be accurately labeled in terms of their behavioral context and the identity of the emitter. Indeed, the vocal repertoire of the zebra finch has been described in the field in the complete context of its natural history (Zann, 1996). Our own fieldwork had also given us insights on the range of vocal behaviors produced in the wild (Elie et al., 2010). Thus, as we describe in detail below, using this

خیلی مهم

داره میگه کاره ما به کاره های رفتاری هم اینسایت میده

knowledge and experience, we recorded a very large number of vocalizations emitted in clearly distinct social contexts from different groups of domesticated zebra finches in the laboratory. To obtain the appropriate range of vocal behaviors, these zebra finches were housed in enriched cages that were designed to encourage both affiliative and agonistic social interactions. Additional behavioral conditions (such as those required to produce alarm calls) were established experimentally. We were then able to describe these sounds using large feature spaces and to apply machine learning techniques including a novel regularized discriminant analysis for the identification of the acoustic features encoding the behavioral meaning embedded in these vocal signals.

Methods

Recording the Complete Repertoire of Captive Zebra Finches

اینجا اومده یه سری داده آمادری درباره فنچ ها داده- تعدادشون سنشون و اینا

Number of birds, age, sex, living conditions—We recorded calls and songs from 45 zebra finches: 18 chicks and 27 adults. Chicks were all recorded while they were 19 to 25 days old (note that only one chick was recorded on day 25 after hatch). Birds were considered as adults if they had molted into their mature sexual plumage, which is achieved around 60 days. For the 17 adults for which we had birth records, their age ranged between 2 months (>60 days) and 7 years (17.6±5.1 months, only one female was less than 90 days old). Of the 18 chicks, 7 were female, 9 were male and 2 of unknown sex (LblGre0000 and LblGre0001). Of the 27 adults, 13 were female and 14 were male. All the birds were born in one of the captive zebra finch colonies housed at University of California (UC) Berkeley or UC San Francisco. In these colonies birds are bred in large cages containing 1 to 15 families and can see and vocally interact with the rest of the colony. Therefore, birds were raised in a rich social and acoustic environment.

For recording purposes, adults were divided into groups of 4 to 6 individuals with even sex-ratio. Each group was housed in a cage (L = 56 cm, H = 36 cm, D = 41 cm) placed in a soundproof booth (L = 74 cm, H = 60 cm, D= 61 cm; Med Associates Inc, VT, USA) whose inside walls were coated with 5 cm of soundproof foam (Soundcoat, Irvine, CA, USA) and which was isolated in a room from the rest of the colony. The cage was provided with 3 nest boxes. Food, drinking water, grit, lettuce, bath access and nest material were provided *ad libitum* and the light cycle was 12/12. Adults were housed and daily recorded while freely interacting in these housing conditions for up to 4 months.

Chicks were housed with their siblings and parents in the same family cage (L = 56 cm, H = 36 cm, D = 41 cm) in one of the breeding colony rooms. Food, drinking water, grit and nest material were provided *ad libitum* and the light cycle was 12/12. Lettuce and bath access were provided once a week. Before each recording session, the cage was transferred into a sound proof booth (Acoustic Systems, MSR West, Louisville, CO, USA) and chicks were physically and acoustically isolated from their parents for 30 minutes to 1 hour to elicit their begging calls upon re-introduction into their parents' cage.

Recording methods: equipment, distance, methodology (selection by ear, etc)—All recordings were performed between 02/2011 and 06/2013 using a digital recorder (Zoom H4N Handy Recorder, Samson; recording parameters: stereo, 44100 Hz, gain 90 or

67) placed 20 cm above the top of the cage for adults' recordings or 19 cm from the side of the cage for chicks' recordings. The calls recorded with a gain of 67 (to prevent clipping during recording sessions) were adjusted once digitized to match those recorded with a gain of 90 and in this manner produce a set of audio files that included relative sound level information. Because the position of the birds from the recording device was limited to the size of the cage, the vocalizations in our recordings sampled a range of intensities corresponding to a range of distances from the microphone of 20 to 80 cm. The behaviors of the birds were monitored during the recording sessions (147 sessions of 60–90 min) by an expert observer (JEE) placed behind a blind, into the darkness of the room, for the adults' recordings, or by observation through a peephole in the sound proof booth for chicks' recordings. Note that chicks were placed back one by one in their parents' cage to record their *Begging* and *Long Tonal* calls to ensure identification of the calling chick. Indeed, chicks tend to beg and call together at exactly the same time, so recording individual calls was only possible by separating siblings during recording sessions. During recording sessions of the adult groups, the observer was tracking birds' behavior, sampling vocalizations for which both the behavioral context and the identity of the emitter were clearly identified and taking notes of this information and the exact time of emission of the vocalizations for as many as possible. Then, annotated vocalizations were manually extracted offline from the sessions' recordings and selected to be part of the vocalization library only if no overlap with noise (cage noise, wing flaps, *etc.*) or vocalizations from other individuals could be heard in the extract. Based on the distinctiveness of behavioral contexts and of acoustic structures, and on the grouping and nomenclature described by Zann in his fieldwork with wild zebra finches (Zann, 1996), vocalizations were classified into 11 categories: *Begging* calls, *Long Tonal* calls, *Distance* calls, *Tet* calls, *Nest* calls, *Whine* calls, *Wsst* calls, *Distress* calls, *Thuk* calls, *Tuck* calls and *Songs*. Vocalizations were either isolated as bouts for those emitted in bouts (*Song*, *Begging* calls, and occasionally *Nest* calls) or as individual vocalizations for all the others. A single expert observer (first author JEE) was used to this human based classification, as it required extensive experience with the birds' behavior. Dr. Elie obtained this experience observing zebra finch vocal and social behaviors both in the field and in the laboratory over a period of 6 years. The classification yielded results that for the more descriptive acoustical measures also agrees with previous accounts (Zann, 1996) and is further validated here by using unsupervised clustering algorithms.

Vocalization Preprocessing and the Generation of the Vocalization Data Base

The audio recordings described above resulted in a vocalization library of 3405 vocalization bouts. To prepare the sounds for various acoustical analyses, the vocalization bouts were filtered, segmented into examples of single call or song syllables and time centered. First, all the sounds were band-pass filtered between 250 Hz and 12 kHz to remove any potential unscreened low and high frequency noise that would be outside of the hearing range of zebra finches (Amin et al., 2007) and could affect acoustical measurements such as those pertaining to the shape of the temporal amplitude envelope. Second, vocalization bouts were segmented into individual calls or song syllables. For this purpose, we estimated the sequence of maxima and minima in the temporal amplitude envelope. The amplitude envelope was estimated by full rectification of the sound pressure waveforms followed by

low-pass filtering below 20 Hz. The maxima above 10% of maximum overall amplitude and minima below this threshold were found. When successive maxima were found without interleaved minima, the maximum with largest amplitude was picked. Similarly, successive minima were eliminated by choosing the one with the smallest amplitude. This succession of minima and maxima were used to cut the vocalization bouts into individual calls or syllables. Vocalization segments shorter than 30 ms were ignored. Third and finally, we generated vocalization sound files that were all of the same length and time aligned. This standard representation was key for the subsequent analyses. The length of these vocalization segments was chosen to be 350 ms to accommodate the longest vocalizations with clear start and finish in our vocalization library (female distance calls). Vocalizations segments that were shorter than 350 ms were padded with zeros and sounds longer than 350 ms were truncated. The vocalizations were aligned by finding the *mean time* and centering this time value at 175 ms. The mean time is obtained by taking the amplitude envelope as a density function of time as described below.

This pre-processing yielded a vocalization database of 8136 calls and song syllables from 45 birds. The number of vocalizations and birds recorded varied among categories as shown on Table 1.

Acoustical Feature Spaces

To describe the acoustical properties that characterized each vocalization type we used four distinct acoustical feature spaces that were used in independent analyses. In each of these feature spaces, sounds were described by a number of parameters that were obtained from a series of nonlinear operations on the sound pressure waveform. These acoustical parameters were used to describe the information-bearing features of the sounds. These acoustical feature spaces were chosen because, as compared to the raw sound pressure waveform, these representations were closer to perceptual attributes of the sound (e.g. fundamental frequency and pitch), were more intuitively understood (e.g. RMS and intensity) and/or could have provided the non-linear transformations required for vocalization categories to be segregated with linear decision boundaries.

First, we used an acoustical feature space that summarized spectral and temporal envelopes' acoustical structure as well as fundamental frequency features since these are easily interpretable and have been widely used by bio-acousticians. Parameters in this acoustical feature space can also be directly associated with perceptual attributes (albeit human based). We called these parameters the Predefined Acoustical Features (PAFs).

Second, we used a complete and invertible spectrographic representation. This representation has multiple advantages. Being invertible, it does not make any *a priori* assumptions on the nature of the information-bearing acoustical features. It is also easily interpretable since results such as the discriminant functions obtained in linear discriminant analysis (LDA) or logistic regression can be displayed in spectrographic representation. Moreover, such results can then easily be compared to neural response functions obtained from single neurons that are frequently described in terms of spectro-temporal receptive fields (Theunissen and Elie, 2014), allowing a direct assessment of potential mechanisms for behaviorally relevant neural discrimination. The disadvantage of the spectrographic feature

space is that it is high dimensional and that it requires additional techniques in statistical regularization (dimensionality reduction) as described below.

Third, we extracted the modulation power spectrum (MPS). The MPS is the joint temporal and spectral modulation amplitude spectrum obtained from a 2D Fourier Transform of the spectrogram (Singh and Theunissen, 2003). The MPS is used to further summarize the joint spectral and temporal structure observed in the spectrogram by averaging across features that occur with different delays or frequency shifts. The MPS could therefore be a powerful representation as it offers a shift invariant of the information present in spectrograms and is able to do so with fewer dimensions by focusing on appropriate regions of temporal and spectral modulations, mostly in the lower or intermediate frequencies (Singh and Theunissen, 2003; Woolley et al., 2005). Results in the MPS feature space can also be compared to those found in auditory neurons (Woolley et al., 2009).

Fourth, we extracted the Mel Frequency Cepstral Coefficients (MFCC). The MFCC representation is similar to the MPS since cepstral coefficients are obtained from the Fourier Transform of time slices in the spectrogram. The MFCC differs from the MPS in that it starts with a spectrographic representation obtained from a Mel Frequency filter bank, reflecting the logarithmic frequency sensitivity of the vertebrate auditory system at high frequencies. As opposed to the MPS, in MFCC the temporal information is also kept in the time domain (and not transferred to the temporal modulation domain). MFCCs are commonly and successfully used in speech processing and speech recognition algorithms as they succinctly describe essential information-bearing structures in speech such as formants and formant sweeps (Picone, 1993). MFCC have also been used successfully to study animal vocalizations (Cheng et al., 2010; Mielke and Zuberbühler, 2013). We added the MFCC representation here primarily to compare the classification performance of the new classifiers we propose to those of the classifier used in these previous studies.

Predefined Acoustical Features (PAFs)—Our first sets of parameters described the shape of the frequency power spectrum (also called the spectral envelope here), the shape of the temporal amplitude envelope and features related to the fundamental frequency (Figure 1). The frequency power spectrum was estimated using the Welch's averaged, modified periodogram with a Hanning window of 42 ms and an overlap of 99%. The temporal envelope was obtained by rectifying the sound pressure waveform and low-pass filtering below 20 Hz. Note that the temporal amplitude envelope is in units of pressure amplitude while the spectral envelope (or frequency power spectrum) is in units of pressure square (power). From these envelopes we obtained 15 acoustical parameters: 5 describing temporal features, 8 describing spectral features and two describing the intensity (or loudness) of the signal.

The shapes of the amplitude envelopes (spectral and temporal) were described by treating the envelopes as density functions: the envelopes were normalized so that the sums of all amplitude values (in frequency or time) equal 1. We quantified the shape of these normalized envelopes by estimating the moments of the corresponding density functions: their mean (i.e. the spectral centroid for the spectral envelope *Mean S* and temporal centroid for the temporal envelope *Mean T*), standard deviation (i.e. spectral bandwidth *Std S* and

temporal duration *Std T*), skewness (i.e. measure of the asymmetry in the shape of the amplitude envelopes, *Skew S* and *Skew T*), kurtosis (i.e. the peakedness in the shape of the envelope, *Kurt S* and *Kurt T*) and entropy (*Ent S* and *Ent T*). The entropy captures the overall variability in the envelope; for a given standard deviation, higher entropy values are obtained for more uniform amplitude envelopes (e.g. noise-like broad band sounds and steady temporal envelopes) and lower entropy values for amplitude envelopes with high amplitudes concentrated at fewer spectral or temporal points (e.g. harmonic stacks or temporal envelope with very high modulations of amplitude).

In addition, the first, second and third quartiles (Q1, Q2, Q3) were calculated for the spectral envelope: 25% of the energy is found below Q1, 50% below Q2 (Q2 is the median frequency) and 75% below Q3. These additional parameters were calculated for the spectral envelopes as they exhibited much more structure than the temporal envelopes (see Figure 1). Nonetheless, the quartiles were highly correlated with each other and the spectral mean (see Supplementary Table 1). The average spectral envelopes and temporal envelopes for each vocalization type were estimated by first averaging the envelopes of all the vocalizations of each bird within each vocalization type, and then averaging across birds (i.e. equal weight per bird). The average spectral envelopes showed characteristic broad peaks of energy for each vocalization type that we call formants, using the acoustical definition of “formant” and thus not implying resonances in the vocal tract. Note that in Figure 5A, spectral envelopes were first normalized before average calculations to equalize the weights of the envelope shapes between vocalizations and avoid any masking effect due to differences of loudness between vocalizations.

To capture the intensity of the signal we also calculated the RMS of the signal obtained directly from the sound pressure waveform (*RMS*) as well as the peak amplitude of the temporal envelope (*Max A*).

We extracted 7 parameters describing properties related to the time varying fundamental extracted for each vocalization. This time-varying fundamental was extracted using a custom algorithm that used a combination of approaches, first identifying portions of the vocalization that had high degree of spectral periodicity (or pitch saliency) and then extracting the fundamental at these time points. To estimate pitch saliency, an auto-correlation function was first calculated using a 33.3 ms Gaussian window with a standard deviation of 6.66 ms. This window was slidden along the sound pressure waveform in 1ms steps. The largest non-zero peak in the auto-correlation function corresponding to a periodicity below 1500 Hz was found. The 1500 Hz threshold was chosen so as to avoid detection of harmonics and favor the detection of the fundamental that is known in zebra finches to mostly be below 1500Hz (Tchernichovski et al., 2001; Vignal et al., 2008; Zann, 1996). The pitch saliency was then defined as the ratio of the amplitude of that peak to the amplitude of the peak at zero delay (corresponding to the variance of the signal in that time window). The saliency was only defined for windows with a root mean square (RMS) amplitude above 10% of the maximum RMS found across all sliding windows spanning a given vocalization. The mean pitch saliency (called *Sal* in the figures) was estimated by averaging across time and is an estimate of the average “pitchiness” of the vocalization.

For all time points where the saliency was above 0.5, we extracted a fundamental frequency. An initial guess for the value of this fundamental frequency was obtained from the time delay of the nonzero peak in the auto-correlation function. We then refined the value of that estimate by fitting the frequency power spectrum of the same windowed segment of the vocalization with the spectrum of an idealized harmonic stack. This non-linear fit not only provided small corrections in the guess of our fundamental frequency but also allowed us to quantify significant deviations in the observed power spectrum from this ideal harmonic stack. In particular we detected significant peaks in the spectrum (peaks above 50% of the maxima) that were not explained by peaks corresponding to the fundamental or its harmonics. These peaks were used as evidence of an inharmonic structure occurring in a sound segment that nonetheless had high periodicity. This inharmonic structure was often the result of the presence of two sound sources (double voice phenomenon), produced by the same bird, in the same single vocalization. On the fundamental panel of Figure 1, the detected time-varying pitch is shown as a black line and the extraneous peaks of energy as green segments. The second voice parameter (*2nd V* in Figures 1 & 8) is defined as the percent of the time when a fundamental is estimated and where a second voice is found. The peak 2 parameter (*Pk 2* in Figures 1 & 8) is the average frequency of these second peaks. The other fundamental parameters describe the time varying fundamental: its mean over time (*mean F0* in Figures 1 and 8), its maximum (*Max F0* in Figures 1 and 8), its minimum (*Min F0* in Figures 1 and 8), and its coefficient of variation (*CV F0* in Figures 1and 8), which is a measure of frequency modulation.

Overall 22 acoustical parameters describing the temporal amplitude envelope, the frequency spectrum and the time varying fundamental were obtained. These PAFs were first used to qualitatively describe the defining acoustical features of each vocalization category and then used as inputs to classifiers to quantify the validity of these features to detect vocalization types.

Spectrogram—Our second feature space is the complete spectrographic representation of each vocalization (Figure 2). The spectrograms were obtained using Gaussian windows of spectral bandwidth of approximately 52 Hz (corresponding to the “standard deviation” parameter of the Gaussian). The corresponding temporal bandwidth is approximately 3 ms (or exactly: $1/(2\pi*52)$). The spectrogram had 231 frequency bands between 0 and 12 kHz and a sampling rate of 1017 Hz yielding 357 points in time for the 350 ms window used to frame each vocalization. The total number of parameters describing the sounds in this spectrographic representation was therefore 82,467. This spectrographic representation is invertible (Cohen, 1995; Singh and Theunissen, 2003) and over-complete. Thus, on the one hand, it has the potential to provide a full description of the sound – one with no *a priori* assumptions on the nature of the information-bearing features. On the other hand, these spectrograms could not be “averaged” to obtain a mean description of a vocalization type and, given the high-number of parameters, they could not be used, without further data reduction methods, as inputs to classifiers. In this study, we show how one can combine spectrograms with principal component analysis (PCA) and regularization to use them in classifiers. We also show that logistic regression can also be used to obtain a “defining” spectrographic representation for each vocalization type. In this manner, we were able to

circumvent this problem of dimensionality and use a data-driven approach to describe the defining acoustical features.

Modulation Power Spectrum—Our third feature space is the Modulation Power Spectrum (MPS; Supplementary Figure 1). The MPS is the modulus square of the 2-D Fourier Transform of the log spectrogram. Spectrograms were calculated as described above. The MPS was then obtained as follows. First, the log spectrogram $S(t,f)$ can be written as:

$$S(t,f) = \log\left|\frac{1}{\sqrt{2\pi}}\int e^{-j2\pi f\tau} s(t+\tau)h(\tau)d\tau\right|$$

where s is the sound pressure waveform and h is the Gaussian window centered at t. The MPS is obtained from the Fourier expansion of $S(t,f)$ written in discrete form as:

$$S(t,f) = A_0 + \sum_{j,k} S_{j,k}(\omega_{t,j}, \omega_{f,k})$$

where the $S_{j,k}$ are the 2D Fourier terms

$$S_{j,k}(\omega_{t,j}, \omega_{f,k}) = A_{j,k}\cos(2\pi\omega_{t,j}t + 2\pi\omega_{f,k}f + \varphi_{j,k}).$$

Here $\omega_{t,j}$ describes the modulation frequency of the amplitude envelope along the temporal dimension and has units of Hz. The parameter $\omega_{f,k}$ describes the modulation frequency of the amplitude envelope along the spectral dimension and has units of 1/Hz. In the modulation power spectrum, $A_{j,k}$ is plotted as a function of $\omega_{t,j}$ (shown on the x-axis) and $\omega_{f,k}$ (shown on the y-axis), as done in Figure Supplementary Figure 1.

The time-frequency scale of the spectrograms used in the MPS (here 3ms and 52 Hz) determines the temporal and spectral Nyquist limits of the modulation spectrum (~ 167 Hz and 9.6 cycles/kHz respectively). Because natural sounds obey power law relationships in their MPS (Singh and Theunissen, 2003), most of the relevant modulation energy is found at lower frequency modulations, for example, for zebra finch vocalizations, below 40 Hz and 4 cycles/kHz. Therefore, we chose not to represent higher frequency modulations, which greatly reduced the number of parameters. In addition, we also ignored the phase spectrum ($\varphi_{j,k}$)further reducing by half the number of parameters. Our chosen MPS feature space was ultimately based on 30 $\omega_{t,j}$ slices (between −40 and 40 Hz) and 50 $\omega_{f,k}$ slices (between 0 and 4 cycles/kHz) yielding 1500 parameters for our MPS feature space (*vs.* 82,467 for the spectrogram). Nonetheless and as for the spectrographic representation, PCA was used as a further data reduction technique before using the MPS as input in classifiers.

Mel Frequency Cepstral Coefficients—Our fourth feature space was the Mel Frequency Cepstral Coefficients (MFCCs; Supplementary Figure 2). The MFCCs are calculated from a spectrogram obtained with a frequency filter bank with varying bandwidths. We used N=25 frequency channels between 500 and 8000 Hz spanning the most sensitive region of the zebra finch hearing range (Amin et al., 2007). The Mel

frequency filters are triangular in shape with center frequencies uniformly spaced on the Mel frequency scale with 50% overlap. In natural logarithmic units, the Mel frequency scale is given by (O'Shaughnessy, 1999):

$$f_{Mel} = 1127 \times \log\left(1 + {}^{f_{Hz}}\!/_{700}\right)$$

The amplitude in these bands was estimated in a 25 ms sliding analysis window with a 10 ms frame shift. An example of such a Mel Spectrogram can be seen on Supplementary Figure 2. The cepstral coefficients were then obtained from the discrete cosine transform of the log amplitude of this Mel Spectrogram. Just as we did for the MPS, we truncated the cepstral coefficients at M=12 (out of 25 possible) since higher cepstral coefficients (corresponding to higher spectral modulations) have significantly less power. Ultimately our MFCC representation had 12 cepstral coefficients for 33 time slices for a total of 396 parameters. PCA was also used to further reduce the number of parameters before using these parameters as inputs in classifiers.

Statistical Analyses

Statistically significant differences in mean values across vocalization categories for the 22 PAFs were assessed using linear mixed-effects models. In this analysis, individual acoustical parameters (e.g. the fundamental frequency) were the predicted variable, the vocalization type was the only fixed effect (the predictor) and the bird identity was taken as the random effect. Furthermore, to prevent pseudo-replication effects, all the data for a given bird were averaged before performing the analyses for each vocalization type (Nakagawa and Hauber, 2011). In this manner, data from each bird were given equal weight. The effect size (as a main fixed effect of vocalization type) was reported as the adjusted R^2 of the model, and the statistical significance was calculated from a likelihood ratio test, which in this case is equivalent to an F-test. As post-hoc tests and to assess the differences in acoustical parameters for each vocalization type, we performed a Wald test to assess whether the estimated coefficient (corresponding to the adjusted average value for a particular acoustical parameter) for each vocalization type was significantly different from the average value obtained across all vocalization types. We also report the 95% confidence intervals for each of the coefficients. The complete statistical results from these mixed-effect models are shown in the supplementary material (Supplementary Table 1).

We also used mixed-effect models to assess the effect of sex on acoustical differences. In this model, the predicted variable was the acoustical parameter, the main fixed effects were the vocalization type, the sex and the interaction (*Type***Sex*) and the random effect was the bird identity. Statistical significance for the effect of *Sex* was then obtained from a likelihood ratio comparing the model that included *Type*, *Sex* and the interaction to the model that only included *Type*. When this test was significant, post-hoc tests were performed to determine which vocalization types were different between males and females. In the post-hoc test, data from each vocalization type were analyzed separately and a linear model was used to test the significance of the unique fixed factor *Sex*. This test was

equivalent to a t-test used to assess the differences on the average values of that acoustical parameter obtained for each male and female bird.

Classifiers

To determine the combination of acoustical features that can discriminate between vocalization types and to quantify the degree of discrimination among these categories, we compared two multinomial supervised classifiers using as input the representations in our four feature spaces (Figures 1, 2, Supplementary Figures 1 and 2). The two supervised classifiers are the Random Forest (RF) and the Fisher Linear Discriminant Analysis (FLDA). To use the nomenclature of supervised classifiers, the vocalization types will be referred to as classes in this next section. We also used an unsupervised classifier (clustering analysis): a mixture-of-Gaussians that can be used efficiently to fit multi-modal and multivariate probability density functions.

Random Forest—A random forest (RF) is a powerful supervised classifier that uses a set of classification trees in a bootstrap fashion to both prevent over-fitting and better explore potential partitions of the feature space (Breiman, 2001). RFs have been shown to often provide the best performance in classification tasks including in the field of bioacoustics (Armitage and Ober, 2010). In this study, we used RF to obtain estimates of an upper bound on classification performance. The measure of this upper bound was critical in order to validate the results obtained for the FLDA. We used Random Forests of 200 trees, with a minimum of 5 data points per leaf and a uniform prior for class probabilities to avoid any bias towards categories that would be better represented in the dataset.

Fisher Linear Discriminant Analysis—Our second classifier was the classical Fisher linear discriminant analysis (FLDA). The FLDA finds linear combination of acoustical features to maximally separate classes while taking into account the within-classes covariance matrix. These discriminant functions are the eigenvectors obtained from the ratio of the between-classes and within-classes covariance matrix. Discriminant functions are ordered by the decreasing value of the eigenvalue (i.e. the function where the ratio of the between and within variance is the greatest is first). Linear decision boundaries within the linear subspace spanned by all significant discriminant functions can then be used to classify sound into their respective classes. In our implementation, we assumed that the within-class covariance was the same for all classes. The great advantage of the FLDA over the RF is that it allows one to examine the form of the discriminant functions and thus easily interpret the acoustical factors that could be used for discrimination. The disadvantage of the FLDA is that the classes might not be linearly separable in a particular acoustical feature space.

Logistic Regression—To further facilitate the interpretation of our results, we also performed a series of logistic regression analyses, one for each vocalization type. The goal of these analyses was to find the unique linear combination of acoustical features that would allow one to separate one vocalization type (or class) from all the others. The logistic regression was only performed on the acoustical feature space based on the full spectrograms. The inputs to the logistic regression were taken to be the coordinates of each

vocalization in the subspace defined by the significant discriminant functions obtained in the FLDA.

Performance: Cross-Validation and Regularization—Just as mixed-effects modeling is required in the statistical analyses described above to potentially correct for bird dependent effects, the same care must be taken when training and testing the multivariate classifiers (Mundry and Sommer, 2007). To do so, we used a cross-validation procedure that took into account the nested format of our data. More specifically, we used a training data set where, for each vocalization type, all the data from a particular bird were excluded, and different birds could be excluded for different vocalization types. The "excluded" data were then used as our validation dataset. In this manner, the classification was assessed for vocalizations from a given bird and class that were not included in the training, allowing us to directly assess the generalization of the classifier. Two hundred (200) different permutations of excluded birds per vocalization type were obtained to generate 200 training and validation data sets. These performance data were sufficient to generate stable confusion matrices shown on Figure 10. We also used these performance data to calculate confidence intervals on percent of correct classification using a maximum likelihood binomial fit.

The cross-validation was also used as part of a regularization procedure (see Figure 2). For acoustical feature spaces that included a large number of parameters (*e.g.* the spectrogram) both the FLDA and RF classifiers generated solutions that over fitted the data. To prevent over-fitting, we used principal component analysis (PCA) as a dimensionality reduction step and tested the performance of the classifiers as a function of the number of PCs as a regularization step. As shown on Figure 2, best performances were obtained with approximately 40 PCs when using the Spectrogram as a feature space. In order to use the same number of parameters for all our large feature spaces, 40 PCs were used for both the RF and the FLDA and for the feature space based on the Spectrogram, the MPS and the MFCCs. The percent of the variance in the data explained by these 40 PCs is shown on Figures 2, Supplementary Figures 1 and 2. Using PCA as a regularization step in FLDA is equivalent to assigning a Wishart prior on the within-group covariance matrix of features (assumed to be the same for each group). This technique is called Regularized LDA or RLDA (Murphy, 2012, p. 107). Here we use both the regularization obtained from the PCA (by systematically evaluating the goodness of fit obtained by varying the number of PCs) and the dimensionality reduction obtained in the FLDA (Murphy, 2012, p. 271). We will call this technique the regularized FLDA or RFLDA.

Clustering analysis: a mixture-of-Gaussians used as an unsupervised Classifier—An unsupervised classifier (also known as a clustering algorithm) was used to further determine whether the vocalization types defined behaviorally did indeed form separate clusters in acoustical feature spaces. Unsupervised classifiers decompose the generally multi-dimensional distribution of a dataset into a sum of distributions. In the case of a mixture-of-Gaussians, the component distributions are all multi-dimensional Gaussian distributions. If the weights of these Gaussian component distributions are approximately equal and when the component distributions are well separated (e.g. separated by one standard deviation), then the joint distribution is shown to be multi-modal, suggestive of the

presence of different groups. Note that it is possible that a set of vocalization types (defined behaviorally) could form a unimodal distribution and still be separable using a supervised classifier as long as the different vocalization types are found predominantly at different ranges of this unimodal distribution. The mixture-of-Gaussians unsupervised classifier is thus stringent in that it will only generate "positive" answers for groups that are separable in the sense of being multi-modal.

In the mixture-of-Gaussians modeling, all the data points (the n vocalizations) are used as a sample of the probability density function that is modeled. Each Gaussian component is defined by its weight (one value), its mean value in the feature space (in the k-dimensional space so a k element vector) and its covariance matrix (a symmetric $k*k$ matrix). For a k dimensional feature space, each component Gaussian has therefore $1+k+(k*(k+1))/2$ parameters. For a mixture-of-Gaussians of m components, the model will have $m*(1+k+(k*(k+1))/2)$ parameters. These parameters are fitted by maximizing the likelihood using the Expectation-Maximization algorithm (EM). The EM algorithm finds a local minimum that depends on the initialization values. Thus, we ran multiple fits using different initialization values for each mixture model (i.e. for each value of m) and used the one that gave us the maximum likelihood. To compare the goodness of fit of Gaussian mixtures with different number of components (m), we used the Bayesian Information Criterion (BIC), which takes into account the negative log likelihood but penalizes for the number of parameters relative to the number of sample points (n). Since k is fixed the BIC penalizes for higher numbers of Gaussian components. The model with the smallest BIC is deemed to be the best.

In our analysis, we applied the mixture-of-Gaussians to specific call types (*e.g. Tet* calls) to determine whether they could actually be composed of multiple vocalization types or, in the contrary, to specific sets of two call types to determine whether they indeed formed separate clusters acoustically. For these analyses, we used the 22 PAFs, as described above and in Figure 1. The optimal number of Gaussian components was chosen by finding the best trade-off between having the smallest BIC and obtaining close to uniform relative weights of the Gaussian components. In the case where multiple call types were combined to be modeled by the Gaussian mixture, we could then calculate the proportion of each call type in the groups determined from the Gaussian mixture (as shown in Figures 9 and Supplementary Figure 4).

Software

All analyses were performed using custom code written in Matlab, using the following high-level functions when appropriate. The MFCCs were calculated using the *mfcc* function provided by Kamil Wojcicki to match the algorithm in the Hidden Markov Toolkit for speech processing known as HTK (Young et al., 2006). The mixed-effect modeling was performed using the Matlab function *fitlme.* The RFLDA was estimated with a custom Matlab script that used the Matlab function *manova1* to estimate the FLDA for different PCA subspaces. The random forest classifier was trained using the Matlab function *TreeBagger.* The logistic regression was performed using the Matlab function *glmfit.* The mixture-of-Gaussian modeling was performed using the Matlab function *fitgmdist*.

Results

In the following sections, we first describe qualitatively and quantitatively the different types of vocalizations produced by domesticated zebra finches. We then show how we applied multiple approaches in order to reveal the acoustic signature of each vocalization type. For this purpose, we isolated over 8000 calls and song syllables and used four independent representations for these vocalizations: various measures in the temporal and spectral domains that we call the Predefined Acoustical Features (PAF) as well as three representations in the joint spectro-temporal domain provided by the spectrogram, the modulation power spectrum and a time varying cepstrum (see methods, Figures 1–2, Supplementary Figures 1–2). We compare the results obtained from these different representations of the sounds (or feature spaces) and two distinct classification algorithms. All of our analyses emphasize the importance of the spectral shape and of the pitch saliency as the main acoustical parameters that code meaning in zebra finches' vocalizations.

The vocal repertoire of the domesticated zebra finch

On Figure 3, we show spectrograms of examples of each of the vocalization types that we recorded from our domesticated zebra finch colony. We classified these vocalizations by assessing the behavioral context in which they were emitted and by ear, having learned their acoustical signatures during previous experiments (Elie et al., 2011). Our classification also followed the grouping and nomenclature described by Zann (1996) in his fieldwork with wild zebra finches. Indeed, we found that domesticated zebra finches housed in groups and in living conditions that also promote nesting, foraging, defensive and alarming behaviors produced all vocalization types described by Zann with the potential exception of the *Stack*, a call mainly produced just before takeoff. A *Stack* call has also been described by Ter Maat *et al* (2014) in captive animals as a contact call used somewhat interchangeably with the soft and short contact call, called the *Tet* call. As we will discuss below, the call described by Ter Maat *et al* (2014) is most certainly included in our *Tet* category. Based on different behavioral contexts, we also made a distinction between the alarm call described by Zann (*Thuk* call) and another new alarm call that we named the *Tuck* call. Furthermore, because the two types of nest calls named by Zann *Arks* and *Kackles* constituted a continuum in our recordings (shown below), we decided to group all of them in the same vocalization type, *Nest* call. Note also that although we were able to record *Copulation* calls as described by Zann (1996), the quality and the quantity of recordings for this particular vocalization type were not sufficient to include it in our analysis (but see one example as Supplementary Sound File 5). Finally, we further grouped the vocalization types into calls produced only by juveniles (top row in Figure 3, blue hues), calls and song produced in affiliative contexts (second and fourth rows in Figure 3, purple/black hues) and non-affiliative calls (third row, orange hues).

In addition to qualitatively describing each vocalization type based on its characteristic spectrographic features, we quantified differences across vocalization types using the PAFs measures that described the temporal envelope, the spectral envelope and the fundamental frequency of the sounds (see Figure 1 and Methods). Those PAFs measures were aimed to fully describe the spectral, loudness, duration and pitch characteristics of the vocalizations.

Results based on those measures are shown in Figures 4–7. Statistical analyses using linear mixed effect models showed that most of these PAF measures carried some information about vocalization types as shown in Figure 8. For each PAF, the correctly weighted mean value of each vocalization type with its 95% confidence intervals obtained from the mixed effect model is shown in an additional table (see Supplementary Table 2). This table also reports the *P*-values obtained for the comparisons of each category mean value to the mean over all categories (Wald Test). In the text below, when we state that a particular acoustical measure (such as the fundamental) is significantly different between two vocalization categories, we mean that the 95% confidence intervals for the two categories do not overlap. Finally, from the average spectral envelope of each vocalization type shown in Fig. 5a, we obtained spectral peaks that we called formants using the acoustical definition of this word. The frequency of these formants is shown on Table 2.

Juvenile Calls

Begging Calls: On the fourth day after birth, chicks start to emit *Begging* calls that are described as “soft cheeping sounds”, while gaping to elicit feeding behavior in their parents (Zann, 1996). The acoustic structure of this vocalization changes along development to become the loud and noisy broadband sounds emitted in long bouts by 15-days-old to 40-days-old chicks (see ontogeny descriptions of this call in Zann, 1996 and Levrero et al., 2009). In the present study, we recorded the mature begging call of 19-days-old to 25-days-old fledglings (0 to 4 days after the chick got out of the nest; Figure 3). *Begging* calls were recorded while the chick was displaying the typical head twisted open beak posture of zebra finches. *Begging* calls were among the three vocalizations in the vocal repertoire that were the “noisiest” on average as quantified by our measure of pitch saliency (see Fig. 7B; Wald test, $P<10^{-4}$). Note however that *Begging* calls exhibited a very large range of pitch saliencies and were far from lacking harmonic structure as it can be seen in the examples of Figures 3 and Supplementary Figure 3. In addition, *Begging* calls had the highest occurrence frequency of double-voices (29% *vs* 13% average, Wald Test $P<10^{-4}$). We detected the occurrence of two voices by the presence of harmonics in the spectrogram that were not multiples of the principal fundamental. The measure of frequency of double-voices was conducted on sections of the call that had harmonic structure (defined by a pitch saliency > 0.5). On Supplementary Figure 3, we show examples of calls with two voices, including a *Begging* call. Double voices were quite common in the two juvenile vocalizations and might in these cases reflect an immature control of the bird’s vocal organ. In other avian species, double voices can generate frequency beats (Robisson et al., 1993) or rapid switching of notes in songs (Allan and Suthers, 1994), both of which might be required to produce behaviorally effective signals and provide additional information about the identity of the caller. In zebra finches, the presence of double voices could be informative to distinguish among call categories and, in particular, to further distinguish juvenile calls from adult calls. *Begging* calls had also a very distinctive spectral envelope characterized by two high frequency formants (F3 and F4) between 4 and 8 kHz (visible on the spectrogram example of Figure 3 and see Figures 4 and 5a and Table 2). These high-frequency resonances were unique to this call type. As a result, the mean frequency of their spectral envelope, or spectral mean, was significantly higher than that of all vocalization types (*Mean S* = 5430 Hz *vs.* the overall mean of 2970 Hz, Wald Test $P<10^{-4}$; Figure 7C and Supplementary Table

2). *Begging* calls also had a relatively large frequency bandwidth (Figure 7D). In terms of temporal properties, *Begging* calls displayed an average duration (Figures 6B and 7E; Wald Test P=0.48) but with a very large spread (Figure 7E), and this range of durations was observed within birds since *Begging* call bouts are often composed of calls of varying lengths (see the spectrogram shown in Figure 3). Zann (1996) noted that *Begging* calls were among the loudest in the repertoire and could be heard as far as 100m. Our measurements support that observation as *Begging* calls were on average the second loudest vocalization after the *Distance* call, although we also observed a very large spread of intensities (Figures 6A and 7F). Because of all of these unique acoustical properties, *Begging* calls were very easily classified by discrimination algorithms, as we will show below (Figures 10 and 11B).

Long Tonal Calls: The *Long Tonal* call is a contact call produced by chicks when they are about to fledge (from 15 days post-hatch). Fledglings spontaneously emit this call when they lose visual contact with members of their family and in response to the *Distance* calls of their parents or the *Long Tonal* calls of their siblings (Zann, 1996). Here we report the analysis for *Long Tonal* calls of fledglings recorded 1 to 4 days after they flew out of the nest (21–25 days-old chicks). The *Long Tonal* call is a precursor of the adult *Distance* call and starts to change slightly from 22 days after hatch (Zann, 1996). Therefore, the *Long Tonal* call shares many similarities with *Distance* calls. *Long Tonal* calls were highly harmonic as quantified by very high pitch saliency values (Fig 7B, Wald Test $P<10^{-4}$), with a range of fundamental frequencies that was very similar to that of the adult *Distance* call (Fig 7A). The average fundamental was 625 Hz for females and 671 Hz for males and although this difference was not statistically significant (P=0.15; post-hoc mixed-effect model; see also the differences in range between males and females), it showed a trend that was in accordance with the sex differences observed for the adult *Distance* call, as described below. The shape of the frequency spectrum and the location of the first two formants in *Long Tonal* calls were also very similar to that of *Distance* calls, with the juvenile call being shifted slightly towards higher frequencies and having a slightly larger bandwidth (Figures 4, 5A and 7D and Table 2). *Long Tonal* calls also had similar durations to *Distance* calls and were among the longest calls in the repertoire (measured as a temporal width, *Std T*: 43.8 ms *vs*. the overall mean of 34.7 ms; Figures 6B and 7E, Wald Test P=0.0003). Finally, the loudness of *Long Tonal* calls was middle range compared to other vocalization types and in particular those calls were significantly softer than the adult *Distance* call (see Supplementary Table 2; Figures 6A and 7F).

Juveniles also produced *Distress* calls and *Tet* calls that shared the acoustical characteristics of adult calls described below. Examples of these calls are given as additional sound files (see Supplementary Sound File 6 for a juvenile *Distress* call and Supplementary Sound File 7 for a juvenile *Tet* call). However, we did not record sufficient juvenile *Distress* and *Tet* calls to quantify any differences, were they to exist.

Affiliative calls

Tet Calls and Distance Calls: Adult zebra finches produce two contact calls: the shorter and softer *Tet* call for short-range communication and the louder and longer *Distance* call for long-range communication (Mouterde et al., 2014; Perez et al., 2015; Zann, 1996). The *Tet*

call is the most frequent vocalization as it appears to be produced in an almost automatic and continuous fashion when zebra finches move around on perches or on the ground. These "background" *Tet* calls form an almost continuous hum and do not appear to produce a particular response in the nearby birds, although in the wild Elie et al. (Elie et al., 2010) show that *Tet* calls could also be used for mate recognition in nesting birds. Increases and decreases in *Tet* call frequency might also be informative: a sudden decrease in *Tet* call frequency could signal an unusual event, and the intensity and frequency of *Tet* calls also increases before takeoffs (Zann, 1996). Note also that *Tet* calls are components along with *Nest* calls and *Whine* calls of the quiet duos that mates perform at nest sites (Elie et al., 2010). In a recent analysis of short-range contact calls produced by captive zebra finches, Ter Maat et al. (2014) distinguish *Tet* calls from *Stack* calls, where *Tet* is used to describe the slightly shorter and more frequency modulated set of contact calls and *Stack* is used to describe the calls that are presented as flat harmonic stacks in spectrographic representations. To investigate whether these two types of soft contact calls belong to a single acoustical category or to two distinct acoustical groups, we performed unsupervised clustering on all the soft contact calls we recorded and labeled as *Tets.* The unsupervised clustering was performed separately on male and female *Tets* because, as we will show below, *Tets* are also sexually dimorphic. Here the *Tets* were represented by the 10 first principal components (10 PCs) obtained from the 22 PAF (see Methods and Figure 1). As shown on Supplementary Figure 4(B and C), female and male *Tet* calls can be clustered into two groups, one with low values of coefficient of variation (CV) for the fundamental (corresponding to the description of the *Stack* by Ter Maat et al., 2014) and one with high CV values for the fundamental (corresponding to the description of the *Tet* by Ter Maat et al.). Because, in our observations, these two acoustically distinct call types were produced during the same behavioral context, we grouped them together and designate them as *Tets* from here on.

Distance calls are produced when zebra finches are out of immediate visual contact with the colony, their mate or the fledglings they care for. *Distance* calls can be produced both during flight and while perched. These loud contact calls carry individual information and elicit orienting responses and vocal callbacks both in juveniles and adults, and promote reunions: they are used for sex, mate, parent or kin recognition (Mouterde et al., 2014; Mulard et al., 2010; Perez et al., 2015; Vicario et al., 2001; Vignal et al., 2004). In our recordings, the *Distance* calls, the *Long Tonal* calls and the *Tet* calls had the highest levels of pitch saliency (Figure 7B). The sharp harmonic structure of all these contact calls can be seen in the spectrogram examples shown in Figure 3. Of the three harmonic contact calls, *Tet* calls had the lowest fundamental frequency (Figure 7A; male and female *Mean F0 Tet* = 558 Hz, *Distance* = 680 Hz, *Long Tonal* = 655 Hz; see Supplementary Table 2). *Tet*, *Distance* and *Long Tonal* calls could also be distinguished by their spectral envelope and formant frequencies: both *Distance* and *Long Tonal* calls were characterized by the highest first and second formants (see Figure 5A and Table 2), while *Tet* calls had those formants at average frequencies. As a result the spectral mean of *Tet* calls was significantly lower than that of *Distance* and *Long Tonal* calls (*Tet* = 2280 Hz, *Distance* = 3580 Hz, *Long Tonal* = 3600 Hz; see Figure 7C and Supplementary Table 2). *Tet* calls also differed in their duration and loudness. *Tet* calls were very short calls while *Distance* and *Long Tonal* calls were among

the longest (*Tet Std T* = 23.9 ms; *Distance* = 47.7 ms; *Long Tonal* = 43.8 ms; see Figures. 6B and 7E and Supplementary Table 2). *Tet* calls were also much softer than *Distance* calls (see Figures 4, 6A and 7F, and Supplementary Table 2) as one might expect given their function.

Nest Calls: Potential nest sites and nests are the scenes for particular soft calls: the *Nest* calls and the *Whine* calls. *Nest* and *Whine* calls are emitted by paired birds around reproductive activities: when they are searching for a new nest, when they are building their nest and almost each time they relieve each other at the nest during the brooding period (Zann, 1996; Elie et al. 2010; Gill et al., 2015). These calls are emitted in sequence either by one single partner (especially by the male when leading the nest search; Zann, 1996) or by both birds that are then performing soft duets using these calls in combination with *Tet* calls (Elie et al. 2010). Zann divided *Nest* calls into *Ark* and *Kackle* calls. According to Zann, *Kackles* are shorter raspy sounding loud calls produced initially around potential nesting sites. *Arks* are longer and softer sounds with a harsh sound often coming in pairs as in “ark-ark”. Although in captivity we also recorded *Nest* calls that could be acoustically classified using Zann’s descriptions as *Ark* and *Kackle* calls (see examples on Figure 3), we found that our *Nest* calls were emitted in identical behavioral context and were best described by a unimodal distribution, as revealed by the unsupervised clustering algorithm (see Supplementary Figure 5). Thus, we decided not to separate them into subcategories. *Nest* calls were among the shortest vocalizations with durations similar to that of *Tet* calls and much shorter than *Whine* calls (*Std T* = 28 ms; see Figure 6B and 7E, Wald Test P=0.015). While their level of pitch saliency was far lower than that of *Tet* calls, it was similar to the mid-range pitch saliency of *Whine* calls (Figure 7B). The two formants of *Nest* calls were lower than the formants of *Tet* calls. Compared to the *Whine* calls, the first formants were identical while the second formant was relatively higher in the *Nest* calls. As a result, Nest calls had the second lowest spectral mean after *Whine* calls (*Mean S* = 2013Hz; see Figure 7C, Wald test $P<10^{-4}$). *Nest* calls were also among the softer ones with a level of intensity between that of *Whine* calls and *Tet* calls (see Figures 4, 6A, 7F; Wald test $P<10^{-4}$).

Whine Calls: As for *Nest* calls, *Whine* calls are produced during early phases of pair bonding and around nesting activities (Elie et al., 2010; Gill et al., 2015). Birds often emit this vocalization while adopting a particular posture in the nest: they lay and slightly twist their head in the direction of their mate while fanning their tail feathers. This vocalization is often followed or preceded by chattering beak sounds. *Whine* calls are also produced during copulation but none of the *Whine* calls analyzed here were recorded during copulation (see Supplementary Sound File 5 for a *Copulation Whine*). In our recordings, *Whine* calls were harmonic sounds that had a middle level of pitch saliency (Figure 7B) but had the unique quality of having a slowly modulated pitch as shown in the example spectrogram on Figure 3. Similar to *Nest* calls, the spectral envelope of the *Whine* calls was heavily skewed towards the low frequencies (Figures 4 and 5A). With the lowest first and second formants (see Figure 5A and Table 2), they had the lowest spectral mean of the repertoire (*Mean S* = 1835 Hz; see Figure 7C, Wald test $P<10^{-4}$). *Whine* calls were also the longest (*Std T* = 55 ms *vs* 35 ms average; see Figures 6B and 7E, Wald test $P<10^{-4}$) and, with the *Nest* calls, the softest vocalizations in the repertoire (see Figures 4, 6A and 7F). It is its long duration, soft

intensity, and varying pitch with low formant frequency that gives this call its whiny quality and hence its name. *Whine* calls were also one of the adult vocalization types with the highest presence of double voices (12.6% *vs* 9.6% for all adult calls). A *Whine* call with a double voice is shown on Supplementary Figure 3.

Non-Affiliative Calls

Wsst Calls: The *Wsst* call is an aggressive call often produced right before an attack on a conspecific by the perpetrator. Both male and female zebra finches produce aggressive calls when supplanting an individual that is perching close to their nest (30–20cm), especially when they are in the nest-building or brooding phase. This aggressive call was named the *Wsst* (Zann, 1996) to describe its short noisy sound quality that can also be described as a brief cat hiss (sound examples are provided along with Figure 3). As quantified by its pitch saliency (lowest – see Figure 7B, Wald test $P<10^{-4}$) and bandwidth (largest – see Fig 7D, Wald test $P<10^{-4}$), the *Wsst* call was the noisiest vocalization in our zebra finch recordings, a quality that it shared the most with the *Begging* call and the *Distress* call. *Wsst* calls also had a very characteristic frequency spectrum: they were dominated by a low frequency formant (690 Hz) followed by a second formant that was also relatively low (1.8 kHz) but with a power spectrum that showed a long tail of significant energy at higher frequencies (see Figures 4 and 5A, Table 2, as well as the example spectrograms shown in Figure 3). From the examples shown in Figure 3, one can appreciate the similarities and differences between *Begging* calls and *Wsst* calls: both calls were clearly noisy and broadband but the *Begging* calls showed the characteristic high-frequency formants while the *Wsst* calls showed the characteristic low frequency formants. This difference was also quantified by the significantly large differences in spectral mean (*Mean S* = 2.6 kHz for *Wsst vs*. 5.4 kHz for *Begging* and *vs.* 3 kHz mean across all types, Wald test $P = 0.0001$, see Figure 7C and Supplementary Table 2). *Wsst* calls were also among the longest in duration in the repertoire (*Std T* = 52 ms *vs.* overall mean 34.7 ms; Wald test $P<10^{-4}$; see Figures 6 and 7E). *Wsst* calls had similar loudness as other non-affiliative calls, at the middle of the loudness range between the louder *Distance* calls and the softer affiliative calls (Figure 7F, Wald test $P=0.035$).

Distress Calls: Zebra finches produce *Distress* calls when they are attacked by other conspecifics, usually while they are escaping or being brutalized by their aggressor. *Distress* calls are noisy calls and share many similarities with *Wsst* calls although to our ear they sounded more tonal. The pitch saliency of *Distress* calls was measured as slightly higher than that of *Wsst* calls but a confidence interval analysis shows that this difference is not significant (Figure 7B, Supplementary Table 2). Similarly to the *Wsst* calls, the spectral envelope of *Distress* calls was characterized by low first and second formants and broad bandwidth (Figures 4, 5A, 7D and Table 2). Interestingly, the *Distress* call was the only adult call category to show a third formant as in the *Begging* call of chicks (Table 2, Figure 5A). Compared to *Wsst* calls, *Distress* calls were significantly slightly shorter in duration, with a more peaked temporal envelope (Figures 6 and 7E; see Supplementary Table 2). The unsupervised clustering applied to *Wsst* and *Distress* calls suggested that the distribution of calls is best described by two Gaussians but we also found that the two call types are equally well represented in these two groups (Figure 9A). Thus, the *Wsst* and *Distress* calls showed

a large amount of overlap and could constitute a single category with differences in pitch saliency and intensity/duration reflecting the degree of dominance in an aggressive conflict. Indeed, the performance of our classifiers revealed that *Distress* calls were often misclassified as *Wsst* calls (see below).

Alarm Calls: Adult zebra finches produced two alarm calls: the *Thuk* call, produced by parents and directed at chicks, and the *Tuck* call, a more generic alarm call. On one hand, *Thuk* calls were produced only by parenting adults when a minor sign of danger occurred (including slight noise from the hiding experimenter) while their own chicks were actively begging for food. On the other hand, *Tuck* calls were produced by our adult birds in the presence of hawks calls or the experimenter's hands in front or in the cage. The two alarm calls were the shortest vocalizations produced by zebra finches (*Std T*: 13.6 ms for *Thuk* calls and 15.0 ms for *Tuck* calls; see Figures 6 and 7E, Wald test $P<10^{-4}$ for both). The alarm calls had also a middle level of pitch saliency on a par with *Whine* calls (Figure 7A and example spectrograms in Figure 3, Wald test $P=0.34$ and $P=0.55$ for *Thuk* and *Tuck* respectively) and mid-levels of intensity on a par with other non-affiliative calls (Figure 7F). The biggest difference between *Tuck* calls *and Thuk* calls might have been in their spectral shape: *Tucks* had a higher second formant (Figures 5A. 11, Table 2), resulting in also slightly higher spectral bandwidth and spectral mean, both of which approached significance according to our conservative assessment using 95% confidence intervals (Figure 7C and 7D, see Supplementary Table 2). These spectral differences can also be observed in the examples shown on Figure 3. To further convince ourselves that these two calls were acoustically distinguishable, we also applied the unsupervised clustering algorithm to the *Thuks* and *Tucks* combined. As shown in Figure 9B, the PAF distribution for these alarm calls was best modeled with two well separated Gaussians of approximately equal weight. More importantly, one group had a much higher proportion of *Tucks* while *Thuks* dominated the second group.

Song—Beside calls, male zebra finches also emit a more complex vocalization during courtship, pair bonding, and mating behavior: the *Song*. An example of a male *Song* is shown on the last row of Figure 3. Male zebra finch *Song* has been extensively described and analyzed in previous work given the importance of the zebra finch model system for understanding the neural mechanisms underlying song production and learning (e.g. Tchernichovski et al., 2000; Tchernichovski et al., 2001; Williams, 2004). *Songs* are composed of introductory notes followed by multiple motifs each made of a stereotyped sequence of song elements or syllables. *Song* syllables vary in spectro-temporal structure and include harmonic stacks, down-sweeps, up-sweeps, high frequency tones, inverted u notes and noisy bursts. Given this variety of notes, it was not surprising on the one hand to see that *Song* exhibited intermediate average values and large ranges for all of the PAF shown in Figure 7. On the other hand, *Song* syllables appeared to share some acoustical features, such as typical spectral envelope shape and location of formants (Figures 4, 5A, Table 2). This is clearly seen on the example spectrogram of Figure 3. The formants of *Song* syllables were similar to those of the *Distance* and *Long Tonal* calls: a first high formant at 1.2 kHz and a second high formant at 3.7 kHz (see Table 2 and Figure 5A). Thus, although *Song* syllables were varied and appeared to overlap with other vocalization types, they

remained highly discriminable from calls in part because all notes shared the 3.7 kHz formant found in *Distance* calls but were otherwise shorter in duration and more modulated in time.

Sex acoustic signature—Across all vocalization types there were few but significant gender differences in acoustical parameters (Figure 8B). Post-hoc tests, taking a single vocalization type at a time, showed that sexual dimorphism was only significant for the *Tet* and *Distance* calls (see Figures 7 and Supplementary Figure 6). The greatest sexual dimorphism was observed for the *Distance* call (Vicario et al., 2001): female *Distance* calls were longer (Female *Std T* = 53.9 ms, Male *Std T* = 44.2 ms, P=0.03; Figures 7E and Supplementary figure 6B), had lower pitch (*Mean F0*: Female = 595 Hz, Male = 727 Hz, $P < 10^{-4}$; Figure 7A) and were composed of a single harmonic stack with a steadier and more salient pitch (*CV F0*: Female = 0.08, Male = 0.16, P=0.0012; mean *Sal* Female = 0.83, Male = 0.74, P=0.0065; Figure 7B). Male *Distance* calls were often composed of two parts: a short sharp harmonic stack with a very high pitch followed by a more noisy frequency down-sweep (see Figure 3). Zann and others have referred to that structure as TN for Tonal followed by Noise (Zann, 1996). The noisy down-sweep decreased the average pitch saliency of the male call. It was also during this down-sweep that instances of second voices were found. Comparatively, the sex differences for *Tet* calls were subtler: female *Tet* calls were slightly longer (*Std T* Female = 25.5 ms and *Std T* Male = 22.2 ms, P=0.001; Figures 7E and Supplementary Figure 6B) and had spectral envelopes that were slightly shifted towards higher frequencies (*Mean S* Female = 2400Hz, Male = 2150Hz, P = 0.012; Figures 7C and Supplementary Figure 6A). To further test that *Tet* calls were sexually dimorphic, we also performed the unsupervised clustering algorithm to all *Tet* calls in our data set. As shown on Supplementary Figure 4A, the PAF distribution for *Tet* calls was well described by two well-separated Gaussians with approximately equal weight; one group containing a majority of female calls and the second group containing a majority of male calls.

Summary—In summary, the zebra finch has a complex vocal repertoire of call types and song elements that are used in very specific behavioral contexts. On the one hand, all the vocalizations were broadband and showed a relatively restricted range of fundamental frequencies (at least relative to the human pitch scale) and, because of this, exhibited a characteristic zebra finch sound quality. On the other hand, the vocalization types were clearly distinct from each other both to trained ears and in our quantitative analyses. What are the acoustical parameters that are the most pertinent for the discrimination of vocalization types? We will revisit this question in a more systematic fashion below when we compare the discrimination performance using different feature spaces, but the PAF analysis was very revealing. As seen in Figure 8, parameters describing the spectral envelope (spectral mean, *Mean S*; and quartiles, Q1, Q2, Q3) varied the most across vocalization types. As described above, these differences were best understood in terms of distinctive formants (see also Figures 4,5 and Table 2). The second type of acoustical parameters that was the most distinctive across vocalization types was the set of parameters describing the pitch such as the pitch saliency (*Sal*) and pitch modulation (*CV F0*). Indeed, as shown on Fig 7A, the zebra finch repertoire is composed of vocalization types with high pitch saliency on one extreme (*Tet* and *Distance* calls) and very noisy calls on the other extreme (*Wsst* and

Distress calls). Two other PAFs were also vocalization type dependent, but to a much lesser extent than the spectral envelope and the pitch saliency: duration (*Std T*) and the fundamental (*F0*). Whereas parameters that were the most distinctive of vocalization types were mostly related to the spectral shape, parameters that were most distinctive between male and female calls were the actual fundamental frequency (*Mean F0*), its maximum value (*Max F0*), the frequency of the second voice (*Pk2*), and, to a lesser extent, the pitch saliency (*Sal*). This sex difference could be seen for the most tonal calls of the repertoire: the contact calls (*Long Tonal*, *Tet* and *Distance* calls). Male contact calls had slightly higher fundamental frequencies than female contact calls, although in post-hoc tests this distinction was only significant for the *Distance* call.

Although all vocalization types can be discriminated from each other as we quantify below, we also observed a few organizing principles that grouped multiple vocalization types into larger classes. This grouping effect was particularly evident for the pitch saliency and the spectral mean (Figures 7B and 7C). The three contact calls in the repertoire (*Long Tonal*, *Distance*, *Tet* calls) had very high and similar pitch saliency whereas the calls produced in high stress contexts *(Wsst, Distress*, *Begging* calls) had the lowest saliency. The spectral mean divided vocalization types into 4 natural groups: the *Begging* calls with very high frequencies were their own class; *Song* and the two distance contact calls *(Long Tonal* and *Distance* calls) constituted the second class; all the non-affiliative calls had intermediate values of spectral mean and formed the third class; finally all affiliative calls used in close distance communication groups together as the class of vocalization with the lowest spectral mean. This grouping by spectral means could also be seen in terms of formant frequencies as shown on Table 2 and Figure 5.

Quantifying the Classification of Vocalization Types based on Acoustical Features

Above, we described some PAFs that are distinctive of all vocalization types: we showed that these carry information about vocalization types since high fractions of the variability found in some of these measures can be accounted for by the category (Figure 8). Next, we examined and quantified how well these acoustical features can be used to discriminate each vocalization type: we identified the best combination of parameters to perform such discriminations and we investigated to what extent categories were equally well discriminated. Moreover, although envelope and fundamental parameters are easily interpretable, they remain an *ad hoc* choice of acoustical features: since they are not invertible representations of the sound, they could miss acoustical information present for example in time varying spectro-temporal patterns such as frequency sweeps. We therefore also investigated more complete acoustical feature spaces: a complete and invertible spectrogram, the time-varying Mel frequency Cepstral Coefficients and the Modulation Power Spectrum (see Methods and Figures 2, Supplementary Figures 1 and 2). Finally, we used two classification algorithms: a regularized Fisher Linear Discriminant Analysis (RFLDA) and the Random Forest (RF) (see methods).

Performances of classifiers and feature spaces—On Figure 10, we show the confusion matrices obtained for the four feature spaces and two classifiers. For all four feature spaces, the performances of the RF and the RFLDA were very similar (Figure 11A),

supporting the idea that vocalization types can be separated using linear combinations of these acoustical features. In terms of overall discrimination performance, the PAF space and the spectrogram feature space yielded similar levels of discrimination at around 60% of correct classification. As shown on Figure 11B, many vocalization types were classified well above this level. The notable exceptions were the *Tuck* call, which was confused with the other alarm call, the *Thuk* call, and the *Distress* call that was miss-classified as the aggressive *Wsst* call. As described above, given both the shared behavioral context during which these calls are emitted and the shared acoustical features, these confusions were not surprising. If *Thuk* and *Tuck* calls on the one hand, and *Distress* and *Wsst* calls on the other hand are combined to form 2 categories in lieu of 4, then the average PCC (for the RF classifier on PAFs) increases from 64 % to 76 %.

MPS and MFCC feature spaces could also be used to discriminate among call categories but not as efficiently. On the one hand, both the MPS and the MFCC captured some of the information present in the spectral envelope such as the formants, thus their good performance was not surprising. On the other hand, some acoustical properties were absent from these feature spaces. For example, the MFCC representation would not capture the pitch saliency and the MPS would not represent particular temporal sequences (for example, it cannot distinguish an upsweep followed by a downsweep from an downsweep followed by an upsweep). These two feature spaces were useful to compare our results to those obtained in previous research as well as to identify how one might design an optimal feature space. Note that the MFCC could be better exploited for the extraction of birds' formants if the coefficients were optimized for the bird's vocal tract instead of using parameters optimized for human voice. Since both the spectrogram feature space and the PAF space yielded the highest performance and provided discriminant functions that were easily interpretable, we will limit below the description of those discriminant functions to those two feature spaces.

Distinctive acoustical features given by multi-dimensional classifiers—Which acoustical features were revealed as relevant for this classification task, and how did the different vocalization types occupy that acoustical space? The top row of Figure 12 shows the first five discriminant functions (DF) obtained from the RFLDA performed on the spectrograms. The bottom two rows of Figure 12 depict how vocalization types were segregated in this discriminant space. The DF1 for the spectrogram did three things: it de-emphasized the very high formants (F3 above 4 kHz and F4 above 6.5 kHz) present in the *Begging* call, it emphasized the high frequency formant (F2 between 3.2 and 3.8 kHz) present in the *Distance* call, *Long Tonal* call and *Song* syllables and it emphasized calls with high pitch saliency by picking out the fundamental and second harmonic of a stereotypical adult vocalization. The DF2 stressed the lower frequency formants (below 3kHz) present in non-contact calls (W*sst, Distress, Nest, Whine, Thuk* and *Tuck* calls) and in the *Tet*, and deemphasized the high frequency band between 3 and 5 kHz, which corresponds to the tail of the second formants (F2) found in the *Long Tonal* call and in the *Distance* call in particular (Figure 4). DF3 deemphasized the very lower tail of the lower formants, and by doing so, separated the *Wsst, Whine* and *Distress* from other non-contact calls. DF4 and DF5 performed further analyses of the shapes of the spectral and temporal envelopes, extracting, for instance, measures of the duration of the vocalizations by differential

weighting of temporal slices that were alternatively emphasized and de-emphasized. In summary, these DF operated principally on the coarse spectral shape but also detected pitch saliency and temporal structures such as duration.

Table 3 shows the coefficients of the eight most important variables used for the first four DF in the RFLDA applied to the PAF space. As one might expect from the results of the models involving individual acoustical features (see above and Figure 8) and from the RFLDA applied to the spectrogram feature space, spectral envelope attributes dominated the first two DF and were present in all four: particular combinations of spectral means (*Mean S*), spectral skew (*Skew S*) and quartiles (Q) were used to distinguish the characteristic spectral envelope of each call. The amplitude of the call (as *Max A* or RMS) also played a role in all functions. The pitch saliency (*Sal*) in the first two DF and the fundamental (*Mean F0*) in combination with its CV (*CV F0*) in the third and fourth DF played a more minor role. The third DF also extracted temporal envelope parameters by emphasizing sounds with higher temporal modulations.

By comparing the scatterplots of the centroids in Figure 12(B and C) between the spectrogram feature space and the PAF space, one can see that the DFs 1 and 2 for both feature spaces performed a similar parsing of vocalization types (with an inverted sign in DF1). This congruence of the DF supports the conclusion that the information present in the coarse spectral envelope is highly robust to distinguish vocalization categories. Moreover, DF3, which was obtained from the PAF space and analyzed temporal modulations, performed a similar segregation as DF5 of the spectrogram feature space. Thus, both discriminant analyses uncovered similar discriminative structures in vocalization types.

Distinctive spectrographic features given by the Logistic Regression—Finally, we examined the single spectrographic dimension that would best distinguish one vocalization type from all the others. For this purpose, we performed a logistic regression in the subspace spanned by the 9 DFs (all significant with $P < 0.001$) obtained in the RFLDA of the spectrographic space (see methods). These logistic weights are shown as spectrograms on Figure 5B. Again, one can see that these functions were mostly different from each other by emphasizing different coarse regions of frequency space. Using black lines, we marked for each vocalization type the formants that were extracted from the average spectral envelope as shown on Figure 5A. One can see that these logistic functions emphasized frequencies that included the formants of the vocalization type while de-emphasizing frequencies of formants in other vocalization types often resulting in “edges” at formant frequencies: a red or positive weight at the formant next to a blue or negative weight just off the formant frequency. This organization is clearly visible for the *Whine*, *Nest*, *Tet* and *Distance* calls where the first two formants F1 and F2 are progressively higher and more separated. These logistic weights also extracted informative temporal structure and duration. In particular, all the weights for the shorter vocalization types (*Tuck, Thuk* and *Tet* calls) were shorter in duration and flanked by inhibitory side bands.

In summary, we provided multiple lines of evidence that show that behaviorally classified vocalization types can be discriminated from their acoustical properties and that spectral envelope features play a central role in this distinction. These spectral envelope features can

also be described in terms of characteristic formant frequencies. Vocalization intensity, pitch saliency and duration provide further distinguishing features.

Discussion

Domesticated zebra finches, which we raised socially and housed in enriched environments, as recommended by (McCowan et al., 2015; Olson et al., 2014), produced a range of vocalizations that could be classified based on their use in distinct behavioral contexts. These vocalizations were very similar to the ones that have been observed in wild zebra finches. By obtaining a very large database of high quality audio recordings of this complete vocal repertoire, we were able to determine the principal acoustical features that can be used to classify these vocalization types. For this purpose, we used both classical descriptions of sounds (the Predefined Acoustical Features, PAFs) and data driven approaches in combination with more modern statistical methods to extract the relevant acoustical features that could be used for this classification task. We found that zebra finch vocalizations used in different behavioral contexts are distinguishable primarily based on their spectral shape and secondarily based on their pitch saliency, which distinguishes noisy calls at one end from tonal or harmonic sounds at the other end. As we will discuss below, these results have implications for understanding the evolution of complex vocal communication signals (Fitch, 2000) and for investigating physiological and neural mechanisms involved in their perception (Elie and Theunissen, 2015; Fitch and Kelley, 2000; Woolley et al., 2009) and production (Ohms et al., 2010; Riede et al., 2006; Riede et al., 2013; Wild and Kruetzfeldt, 2012). We will first summarize and discuss the results that relate to our immediate goals of describing the zebra finch vocal repertoire before discussing in more depth the implications of our results for more general theories in animal communication.

Comparison between the vocal repertoire of domesticated and wild zebra finches

Besides the song, domesticated zebra finches have a rich vocal repertoire of communication calls that include aggressive calls, alarm calls, distress calls, contact calls, nest calls and begging calls. The repertoire of our domesticated zebra finches is similar to that of wild zebra finches as described by Zann (1996) with however 4 discrepancies: the absence of the *Stack* call, the heterogeneity (or duality) of *Tet* calls, the grouping of *Ark* and *Kackle* calls into a single *Nest* call category and the description of a new alarm call, the *Tuck* call. Zann describes the *Stack* call as “Louder, longer and higher pitched than *Tet*s, but softer, shorter and lower pitched than *Distance* call*s, Stacks* are emitted at the moment of take-off.” The unsupervised clustering analysis we conducted on *Tet* and *Distance* calls categorized those calls into two and not three groups, excluding the possibility of a “missed” category between *Tet* and *Distance* calls. Thus, our dataset does not appear to contain the *Stacks* as described by Zann. Domesticated zebra finches might produce few of these *Stack* calls because a synchronized take-off is not part of their repertoire when housed in cages. More recently, Ter Maat *et al.* (2014) also designated some soft contact calls exchanged between domesticated zebra finches as *Stack* calls because they could be described as constant harmonic stacks in the spectrogram space (see also Gill et al., 2015). These stack-looking calls could be distinguished from *Tet* calls by being slightly longer and less modulated in pitch. The results from our unsupervised clustering analyses do support the idea that, based on their acoustical

properties, soft contact calls, designated as *Tet* calls here, can be categorized into two call types: one with greater frequency modulation and one that can be described as a constant harmonic stack. This distinction was never described by Zann in wild zebra finches and could be a particularity of domesticated birds as it matches the observations of Ter Maat *et al.*. Because in our hands these two types of close distance contact calls were emitted in the same behavioral context, we decided to keep them in one group in our analysis (and labeled all of them as *Tet* calls). In the future, to avoid confusion with the take-off call described by Zann, we suggest dividing the soft contact calls into *Tet-M* and *Tet-S* for *Tet-Modulated* and *Tet-Stacks*.

Regarding the soft and short calls emitted by adults around nest activities, we did not distinguish *Arks* and *Kackles*. Here, the unsupervised clustering analysis supported a unimodal distribution of sounds, and given that the calls were produced in the same context, we could not justify additional groups and maintained a single category, the *Nest* calls. We note, however, that this category is relatively large and that calls could be classified along a continuum going from more tonal *Ark* calls to more noisy *Kackle* calls. Finally, we divided alarm calls into *Thuk* and *Tuck* calls, a distinction that had not been made until now; *Thuk calls* are alarm calls emitted by brooding parents and *Tuck* calls are alarm calls emitted by any adult (see results). The unsupervised clustering analysis supported this novel distinction. Ultimately, we described a complex vocal repertoire of 8 adult call types shared among the two sexes, 2 juvenile calls, and a song that is uniquely produced by males. In addition, the complete zebra finch vocal repertoire probably includes a *Stack* call emitted at take-off and an unmodulated *Tet* call, the *Tet-S*. Finally, *Nest* calls could be produced along a continuum of tonal to noisy sounds, which might also be produced in slightly different behavioral contexts.

Quantifying the discrimination

Our classification procedures allowed us to quantify the discrimination of vocalization types based on acoustical features of single calls and, by generating confusion matrices, to determine the potential nature of systematic errors in such classifications. The performance of the classifiers is relatively high, at approximately 60% across all categories and reaching levels above 80% in cross-validation data for *Begging*, *Long Tonal* and *Wsst* calls. The calls that are the most confused are: the *Distress* call, that is systematically confused with the *Wsst* call; the *Tuck* call, that is systematically confused with the *Thuk* call; and the song syllables. Results from our unsupervised clustering show that *Distress* and *Wsst* calls are mixed in acoustical space. This overlap might also make sense from a behavioral standpoint: *Distress* and *Wsst* calls are produced in the same context, during intraspecific conflicts where the aggressor and aggressed can change roles or express various levels of aggression and distress. If the acoustical changes are graded they might be poorly accounted by a categorization. Alternatively, *Wsst* and *Distress* calls might be acoustically distinguishable in the temporal sequence of syllables, which was not examined here. In our dataset, *Distress* calls were more often misclassified than *Wsst* calls; this asymmetry can be explained by the fact that our sample size for the *Distress* calls was much smaller than for the other call types. Given the design we chose for the cross-validation procedure of the classification algorithms (the validating dataset not only did not contain any of the sounds of the training set but also

did not contain any vocalizer that had been chosen for the training set), categories with smaller sample size were penalized. *Thuk* and *Tuck* calls are the two alarm calls that, given their observed effect on receivers, we believe are directed to juveniles and adults respectively. Our unsupervised clustering analysis shows that these two call types are indeed well separated acoustically (and are not part of a unimodal distribution that we divided in two) but with some overlap. Song syllables are also often misclassified but not systematically. Instead song syllables can be confused with many different call types, which is to be expected given the variability of song syllables (Williams, 2004). Note also that our classification procedure was based on isolated calls and song syllables. Other acoustical structure such as the temporal sequence of calls and song syllables would provide additional vocalization specific cues: *Distance* calls are often produced in pairs; *Begging* calls are produced in bursts; *Tet* calls are emitted in long intermittent streams; *Whine, Tet* and *Nest* calls are produced in synchronized duets between mates (Elie et al., 2010) and *Song* is characterized by a very stereotyped and fast temporal sequence of specific song syllables. Including such temporal sequence information in our classifiers would have certainly increased the discrimination performance. Finally, the performance of classification that we obtained should be compared with that of zebra finches. Indeed, behavioral testing of zebra finches using conditioning procedures will assess the actual behavioral discriminability of these vocalization categories based solely on acoustical cues and reveal how zebra finches hierarchically structure their own repertoire.

Acoustical features for vocalization type discrimination

Our extensive database of vocalization examples allowed us to use a data driven approach, in addition to a more classical bioacousticians' approach (using the PAFs), to determine, without making any *a priori* assumptions on the nature of the relevant acoustical features, the acoustical parameters that vary across vocalization types and could therefore be used for vocalization type classification. For this purpose, we used an over-complete representation of the sounds, their spectrograms, and used data reduction techniques (PCA) combined with cross-validated classifiers to find the relevant acoustical features. Similar approaches that rely heavily on large data sets and machine learning techniques have been used recently to classify birds' calls from different species (Stowell and Plumbley, 2014) and to cluster a primate species' calls using unsupervised algorithms (Fuller, 2014). Such data driven approaches provide unique opportunities to examine the information-bearing features in communication sounds without making *a priori* assumptions on the nature of such features or on the number of categories. Besides, our approach using the full spectrogram can also be used with sounds degraded by other signals or propagation (Mouterde et al., 2014) as it will happen in normal communication events in wild species.

Formants produced by active vocal filtering—To validate our approach and to facilitate the interpretation of our results, we used 3 sound representations, besides the spectrogram: the MFCC and the MPS as well as predefined but more classical sound features extracted from the spectral and temporal envelopes and the time-varying fundamental (the PAFs). The results for all these analyses led to the same conclusion: zebra finch vocalization types are primarily distinguishable based on the coarse shape of their spectral envelope and secondarily based on the saliency of their periodicity structure or pitch

saliency. In contrast, the frequency of the fundamental (the pitch) varied very little across the entire repertoire (see results and Figures 7A and 8A). Moreover, examination of the average spectral shape of each vocalization type (Figure 5A) and of the spectro-temporal discriminant and logistic functions (Figures 12 & 5B), both point to characteristic spectral peaks for each vocalization type that can be used for the vocalization classification task. These formants can in part be attributed to resonances in the birds upper vocal tract: in a recent original study using X-ray cinematography, Riede et al. (Riede et al., 2013) show that singing and calling zebra finches vary their tracheal length, the size of their oropharyngeal–esophageal cavity (OEC) as well as the gape of their beak (Goller et al., 2004) to modulate the resonant peaks of their vocal tract. The OEC resonance results in a formant peak between 2 and 5 kHz while the trachea and the beak can produce other formants in the same frequency range. Our data suggest that indeed zebra finch vocalizations vary in the number and the positions of these formants. We did not directly associate resonances of a particular anatomical structure of the vocal tract to the formants we measured but, given the results in (Riede et al., 2013) obtained for *Tets* and *Distance* calls, we suspect that the formant frequencies labeled F2 (and F3 for *Distress* and *Begging* calls) are produced by changes in the OEC. This hypothesis could be tested using the experimental techniques of (Riede et al., 2013) while birds produce all the vocalizations in their repertoire.

Thus, songbirds join other birds (Fitch and Kelley, 2000), some mammals (Fitch, 1997; Reby et al., 2005; Riede and Zuberbuhler, 2003) and humans (Lieberma. Ph et al., 1969) in the use of varying spectral resonant peaks in their vocal tract to generate communication calls with distinct information. Moreover, although in most animals, the vocal tract filtering appears to be principally used in a static way (Fitch, 1994; Fitch, 2000; Fitch, 2000; Fitch, 2002; with notable exceptions such as the larynx descent in the red deer that is used for acoustic size exaggeration, Fitch and Reby, 2001) and is useful for identifying the caller and some of his anatomical or physiological attributes (Fitch, 1997; Taylor and Reby, 2010), birds and primates (Riede and Zuberbuhler, 2003) are also able to use active vocal filtering for generating vocalization types with different meanings. Active vocal filtering might therefore be a more ubiquitous feature in animal vocal communication than previously thought and differences in complexity between the human control of formants and those of birds might also not be so disparate.

The role of the syrinx and respiratory system—Zebra finches' vocalization types are not only characterized by their spectral shape but also by their pitch saliency, duration and intensity. Vocalization types vary in their pitch saliency from the very noisy aggressive (*Wsst* call) and *Distress* calls to the very tonal contact calls (*Distance* calls, *Tet* calls and *Long Tonal* calls). The pitch is produced by the birds' vocal organ, the syrinx (Fee, 2002; Goller and Larsen, 1997), and models suggest that noisy sounds could also be generated at the syrinx when high air-sac pressure drives the system into chaotic regimes (Elemans et al., 2009; Fee et al., 1998). Thus, control of the syrinx in conjunction with the respiratory system will be key for controlling the pitch (the fundamental), the pitch saliency, the duration and the amplitude of the sounds. Moreover, it is also known that the non-linear dynamics of the syrinx produce spectrally rich tonal sounds: the syrinx in isolation already generates harmonic sounds with particular spectral envelopes that are correlated with the

fundamental frequency (Fee et al., 1998; Sitt et al., 2008; Williams et al., 1989). The acoustic formants we measured might therefore depend both on the mechanical properties of the syrinx and of the upper vocal tract. Thus, the generation of a variety of vocalizations in the zebra finch repertoire requires the coordinated control of the respiratory organs, the syrinx and the upper vocal tract, just as in human speech (Riede and Goller, 2010).

The neural control of the syrinx in songbirds is relatively well explored in the context of song production in male birds (e.g. Amador et al., 2013; Hahnloser et al., 2002) but the neural control of the syrinx for other vocalization types is just beginning to be examined (Ter Maat et al., 2014) and the coordination of the respiratory system, the syrinx and the upper vocal tract to produce the sounds in the entire repertoire has not been studied. Given the wealth of research on the syrinx and its neural control and now our quantified description of the complete vocal repertoire of the zebra finch, we believe that the Zebra finch model is particularly appropriate for investigating the neural and motor control for the production of a complete and complex vocal repertoire. Such studies would not only provide insights on neural control of vocal gestures but also on the evolution of brains and vocal organs for the production of a complex vocal signaling system.

Expected consequences for the behavioral and neural discrimination of vocalization types—Our analysis opens the doors for behavioral and neurophysiological experiments on the perceptual side. The behavioral perception of sounds in zebra finches has been well studied but principally as it applies to song perception (e.g. Clayton, 1987; Clayton and Prove, 1989; Scharff et al., 1998; Sturdy et al., 1999) and the question of individual recognition (e.g. Mouterde et al., 2014; Vignal et al., 2008). Zebra finches are also known to be exquisitely sensitive to the spectral structure of harmonic sounds (Lohr and Dooling, 1998). It has also already been demonstrated that zebra finches can learn to classify their song syllables and human speech vowels as open-ended acoustic categories (Kriengwatana et al., 2015; Sturdy et al., 1999). However, it remains to be seen whether zebra finches are able to perform categorical perceptions in conditioning experiments along the lines of the vocalization categories described here. Similarly conditioning experiments, where the specific cues that we have identified here as being important for vocalization categorization are systematically manipulated, need to be performed to directly assess their actual importance from the receivers' perspective. The representation of natural sounds in the avian auditory forebrain has also been well studied. We and others have shown that the avian auditory cortex is particularly responsive to spectro-temporal structure found in natural sounds (Hsu et al., 2004; Woolley et al., 2005). Moreover, spectro-temporal receptive fields (STRFs) estimated for auditory neurons in the avian auditory cortex exhibit a range of tuning that includes neurons with coarse spectral tuning that would be useful to extract formants (and timbre) and narrow spectral tuning with long integration times that are efficient at detecting pitch saliency (Kim and Doupe, 2011; Nagel and Doupe, 2008; Woolley et al., 2009). In a recent study, we have directly measured neural responses to the entire vocal repertoire and found that approximately 50% of auditory neurons have responses that carry information about vocalization type category and a fraction of these "semantic" neurons also showed selective and invariant response properties for vocalization categories (Elie and Theunissen, 2015). We are currently investigating the nature of the non-

linear transformations between sound and neural responses that could explain these "categorical" neural responses. Given the results presented in this paper, we hypothesized that we will find neuron responses that extract formant information along the characteristic axis of specific vocalization types (e.g. in STRFs that resemble the logistic weights of Figure 5B) or that are sensitive to pitch saliency. Step-like response functions along such acoustical dimensions could then be used to categorize sounds into specific vocalization types.

Evidence for Referential coding in Zebra Finches?

Multiple species of birds (Evans et al., 1993) and mammals (Seyfarth et al., 1980) have shown to produce different alarm calls depending on the type of predator. In addition, in birds, these alarm calls elicit different behaviors in chicks and adults (reviewed in Gill et al., 2013). For example, the white-crowned scrub-wren produces a buzz in response to ground predators and a trill in response to aerial predators. In response to the trill, nestlings will stop producing begging calls while adults will scan the environment and fly to cover. Here we have found that zebra finches produce two types of alarm calls that are distinguishable based both on their behavioral context and on their acoustical structure: the *Thuk* call and the *Tuck* call. Moreover, in captivity, the *Thuk* was observed to be directed at chicks that immediately stopped begging in response, while the *Tuck* was directed at the entire group and elicited adults to stay quiet and motionless and to scan the environment. As discussed in Gill *et al.* (2013), the study of alarm calls in birds could provide further insights on the degree to which animal produce communication calls that have a functional reference and thus that are not simply the result of an internal state. Here, zebra finches appear to change their alarm call depending on whether or not danger appear in co-occurrence with their chicks emitting begging calls. Further behavioral studies identifying all the exact contexts that can systematically elicit each type of alarm call and investigating whether playback yields differential responses in adult and chicks are needed to determine whether zebra finches also produce calls with functional references.

A universal size for the core elements constituting a vocal repertoire?

The size of the repertoire of the zebra finch is of similar order of magnitude to the sizes of the repertoires that have been described principally based on spectrographic examination in other species of birds; for example, adult black-capped chickadees produce 11 calls plus two in chicks (Ficken et al., 1978), 24 call types have been described in the red jungle fowl (Collias, 1987), 11 call types in the Eurasian stone-curlew (Dragonetti et al., 2013). Interestingly, similar repertoire sizes are also described in mammals: for example, spotted hyenas have a repertoire of approximately 10 calls (Kruuk, 1972), dingos produce 9 vocalizations in classes that are similar to other canids (Deaux and Clarke, 2013), chipmunks produce 13 distinct calls including 4 types of alarm calls (Brand, 1976) and 17 call types are found in Western and Mountain Gorillas (Salmi et al., 2013). Although these numbers are similar, it is clear that differences in morphology of the vocal and perceptual systems, as well as differences in social and ecological conditions across species, even closely related ones (Salmi et al., 2013), will result in distinct repertoires or distinct uses and functions of acoustically similar communication calls. Besides, this apparent lack of a universal code for communication in the animal kingdom has even been contrasted to our shared genetic code (Hauser et al., 2002). However, this idea of a lack of a universal code

can be partially refuted and our results provide additional evidence for common principles. In terms of the similarity in the numbers of communication calls produced by social birds and mammals, it might be interesting to speculate on the presumably innate capacities to produce a repertoire limited to approximately 10 call types. It is certainly possible that the number of call types has evolved separately in each species to match approximately 10 prototypical behaviors found in all social animals centered around danger (alarm calls), fighting (distress and aggressive calls), group cohesion (contact calls) and mating (nest and contact calls and song), with varying numbers in each of these categories (e.g. number of alarm calls) depending on the species dependent ecology and behavior (Wilson, 2000). Or perhaps this approximately similar limit across species in the vocal repertoire is driven by common mechanisms of production or perception. For example, it is interesting to note that human languages use between 3 and 20 (and mostly below 10) vowel sounds that are distinguishable based on their formants and spectral shape (Ladefoged, 2012). This number is within the range of the 10 to 20 vocalization categories that are found in avian repertoires and that, at least in zebra finches, differ mainly based on their spectral shape. If the use of formants to distinguish vocalization types tends to be a rule in animal communication system, then the size of the core elements constituting the repertoire could be constrained by common mechanisms of production or perception of different spectral shapes.

A Universal code for animal communication? Ecological and Motivational explanations for the structure-function of communication calls

Are there any common principles shared across species that correlate specific acoustical traits to the meaning of the sound? At a coarse level the answer to this question is yes, and the link between sound and meaning can be understood for ecological reasons, such as efficient transmission or on the contrary the need to be inconspicuous to avoid predation (Morton, 1975), or as “rules” relating motivational states to sound structure (Morton, 1977; Owren and Rendall, 2001). As summarized by Collias (1987, p.510) when describing the vocal repertoire of the red jungle fowl: “Brief, soft repetitive notes of low frequency are attraction calls. Loud harsh sounds with high-frequencies are alarm cries. Harsh sounds emphasizing low frequencies are threat sounds. These rules hold for many other birds”. These rules work because these physical properties of sounds elicit approaching or avoiding behaviors respectively from the receiver of any species. However, as very well explained by Seyfarth et al. (2010), a strict manipulative view of communication calls from the perspective of the receiver is certainly over simplistic since receivers can choose to respond differently to calls with similar acoustical features or to identical calls in different contexts and to respond similarly to calls with different features. For these reasons, Seyfarth et al. argue that animal communication is better analyzed with an information perspective and we fully agree with their point of view. However, we also found some evidence for general principles that can explain some of the physical characteristics of communication calls in terms of ecological constraints and the motivational perspective. In terms of ecological constraints, we noted that the *Distance* call of the zebra finch is the loudest allowing for long range propagation and in previous work we have also shown that its harmonic structure (high pitch saliency) as well as the modulation of the fundamental are important for transmitting the individual signature over long distances (Mouterde et al., 2014). The alarm calls are also relatively loud but very short, making them harder to localize, which is critical

in the presence of potential predators. In terms of the motivational approach, we also find support for the rules spelled out by Collias (1987): the most affiliative calls (*Nest* and *Whine*) are the softest in the repertoire and with the lowest spectral mean while the aggressive (*Wsst* call) and distress calls are the noisiest (low pitch saliency) and with the largest spectral bandwidth. It is however interesting to note that zebra finch *Begging* calls also share these physical attributes (and also have the highest spectral means), suggesting that they should be highly aversive while they clearly elicit approaching behaviors, at least in parents. These examples and apparent counter-example (the parents might also be motivated to stop the begging calls; *e.g.* Rendall et al., 2009) support the rules of the motivational theory but also illustrate that a strict manipulative view of the receivers response without taking into account the informative features in the communication will fail to explain the range of complex behaviors elicited by communication calls ("I hear begging calls: Are those the begging calls of my chicks? Have they already been fed? Are they ready to be weaned?").

Conclusions

As stressed by Marler (2004), the study of birdcalls and bird communication offers unique opportunities for behavioral neurobiology. Our quantitative analyses of the complete vocal repertoire of the zebra finch allowed us to make significant findings on the information-bearing features for vocalization type discrimination: vocalizations are mostly categorized by the shape of the spectral envelope that can be explained in terms of formants produced both by the syrinx and the vocal tract of the bird. The dynamic vocal tract shaping is therefore not unique to humans or a few mammals. In addition, we have shown how our data provides support for general principles of animal communication including, on one hand, the ecological and motivational links between physical structure and meaning and, on the other hand, the importance of an "information" approach where behavioral response to specific calls are interpreted in terms of the new specific information they provide for the receiver (e.g. the behavioral response of noisy begging calls by parents and non-parents) (Seyfarth et al., 2010). Finally, our quantitative description of complete vocal repertoire of the zebra finch will facilitate neuro-ethological research for understanding the neural basis of perception and production of communication calls.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We would like to dedicate this study to Peter Marler and Richard Zann and Allison Doupe. By his fundamental discoveries and his thoughtful contributions to the field of animal communication, Peter Marler has been a major source of guidance and inspiration for our own research efforts in the field of bird communication. In his seminal Science paper in 1967, Peter Marler said: "We are beginning to understand how the structure of animal signals relates to the function they serve". We would hope that Peter would agree that we are humbly following his footsteps. Peter Marler is the scientific great-grand-parent of FET and great-great-grand parent of JEE. Richard Zann dedicated his life to the study of wild zebra finches in his native Australia. Allison Doupe developed the Zebra Finch model in crucial seminal studies that examined the neural mechanism of vocal plasticity. She was the scientific parent of FET and grand-parent of JEE. She was an outstanding mentor and a wonderful person. We would not be able to appreciate the complexity and the relevance of our studies without their respective contributions to the field. Richard Zann died in a bushfire inferno that occurred in outskirts of Melbourne in

February 2009. Peter Marler died in July 2014 following a long illness. Allison Doupe died in September 2014 after a long battle with cancer.

This work was supported by an NIH grant CD010132 to FET, a CRCNS NSF grant IIS1311446 to FET and JEE and a fellowship from the Fyssen Foundation to JEE.

References

Allan SE, Suthers RA. Lateralization and motor stereotypy of song production in the brown-headed cowbird. J Neurobiol. 1994; 25:1154–66. [PubMed: 7815070]

Amador A, Sanz Perl Y, Mindlin GB, Margoliash D. Elemental gesture dynamics are encoded by song premotor cortical neurons. Nature. 2013; 495:59–64. [PubMed: 23446354]

Amin N, Doupe A, Theunissen FE. Development of selectivity for natural sounds in the songbird auditory forebrain. J Neurophysiol. 2007; 97:3517–31. [PubMed: 17360830]

Armitage DW, Ober HK. A comparison of supervised learning techniques in the classification of bat echolocation calls. Ecol Inform. 2010; 5:465–473.

Ballentine B, Searcy WA, Nowicki S. Reliable aggressive signalling in swamp sparrows. Anim Behav. 2008; 75:693–703.

Bennur S, Tsunada J, Cohen YE, Liu RC. Understanding the neurophysiological basis of auditory abilities for social communication: A perspective on the value of ethological paradigms. Hear Res. 2013; 305:3–9. [PubMed: 23994815]

Brand LR. Vocal repertoire of chipmunks (genus *Eutamias*) in california. Anim Behav. 1976; 24:319–335.

Breiman L. Random forests. Machine Learning. 2001; 45:5–32.

Catchpole, CK., Slater, PBJ. Biological themes and variations. Cambridge University Press; Cambridge: 1995. Bird song.

Cheng J, Sun Y, Ji L. A call-independent and automatic acoustic system for the individual recognition of animals: A novel model using four passerines. Pattern Recognition. 2010; 43:3846–3852.

Clayton N. Song tutor choice in zebra finches. Anim Behav. 1987; 35:714–722.

Clayton N, Prove E. Song discrimination in female zebra finches and bengalese finches. Anim Behav. 1989; 38:352–354.

Cohen, L. Time-frequency analysis. Prentice Hall; Englewood Cliffs, New Jersey: 1995.

Collias NE. The vocal repertoire of the red junglefowl: A spectrographic classification and the code of communication. Condor. 1987; 89:510–524.

Deaux EC, Clarke JA. Dingo (*Canis lupus dingo*) acoustic repertoire: Form and contexts. Behaviour. 2013; 150:75–101.

Dragonetti M, Caccamo C, Corsi F, Farsi F, Giovacchini P, Pollonara E, Giunchi D. The vocal repertoire of the eurasian stone-curlew (*Burhinus oedicnemus*). Wilson J Ornithol. 2013; 125:34–49.

Elemans CP, Muller M, Larsen ON, van Leeuwen JL. Amplitude and frequency modulation control of sound production in a mechanical model of the avian syrinx. J Exp Biol. 2009; 212:1212–24. [PubMed: 19329754]

Elie JE, Mariette MM, Soula HA, Griffith SC, Mathevon N, Vignal C. Vocal communication at the nest between mates in wild zebra finches: A private vocal duet? Anim Behav. 2010; 80:597–605.

Elie JE, Mathevon N, Vignal C. Same-sex pair-bonds are equivalent to male–female bonds in a life-long socially monogamous songbird. Behav Ecol Sociobiol. 2011; 65:2197–2208.

Elie JE, Soula HA, Mathevon N, Vignal C. Dynamics of communal vocalizations in a social songbird, the zebra finch (*Taeniopygia guttata*). Journal of the Acoustical Society of America. 2011; 129:4037–4046. [PubMed: 21682424]

Elie JE, Theunissen FE. Meaning in the avian auditory cortex: Neural representation of communication calls. Eur J Neurosci. 2015; 22:546–567.

Evans CS, Evans L, Marler P. On the meaning of alarm calls: Functional reference in an avian vocal system. Anim Behav. 1993; 46:23–38.

Farabaugh, SM. The ecological and social significance of duetting. In: Kroodsma, DE., Miller, EH., editors. Acoustic communication in birds. Academic Press; New York, NY: 1982. p. 85-124.

Fee MS, Shraiman B, Pesaran B, Mitra PP. The role of nonlinear dynamics of the syrinx in the vocalizations of a songbird. Nature. 1998; 395:67–71. [PubMed: 12071206]

Fee MS. Measurement of the linear and nonlinear mechanical properties of the oscine syrinx: Implications for function. J Comp Physiol A Neuroethol Sens Neural Behav Physiol. 2002; 188:829–39. [PubMed: 12471484]

Ficken MS, Ficken RW, Witkin SR. Vocal repertoire of black-capped chickadee. Auk. 1978; 95:34–48.

Fitch, WT. Vocal tract length perception and the evolution of language. Brown University; 1994.

Fitch WT. Vocal tract length and formant frequency dispersion correlate with body size in rhesus macaques. J Acoust Soc Am. 1997; 102:1213–1222. [PubMed: 9265764]

Fitch WT. The evolution of speech: A comparative review. Trends Cogn Sci. 2000; 4:258–267. [PubMed: 10859570]

Fitch WT. The phonetic potential of nonhuman vocal tracts: Comparative cineradiographic observations of vocalizing animals. Phonetica. 2000; 57:205–218. [PubMed: 10992141]

Fitch WT, Kelley JP. Perception of vocal tract resonances by whooping cranes grus americana. Ethology. 2000; 106:559–574.

Fitch WT, Reby D. The descended larynx is not uniquely human. Proc Roy Soc B Biol Sci. 2001; 268:1669–1675.

Fitch, WT. Comparative vocal production and the evolution of speech: Reinterpreting the descent of the larynx. In: Wray, A., editor. The transition to language. Oxford University Press; Oxford: 2002.

Fuller JL. The vocal repertoire of adult male blue monkeys (*Cercopithecus mitis stulmanni*): A quantitative analysis of acoustic structure. Am J Primatol. 2014; 76:203–216. [PubMed: 24130044]

Gill LF, Goymann W, Ter Maat A, Gahr M. Patterns of call communication between group-housed zebra ficnhes change during the breeding cycle. eLife. 2015; 4:e07770.

Gill SA, Bierema AMK, Hauber M. On the meaning of alarm calls: A review of functional reference in avian alarm calling. Ethology. 2013; 119:449–461.

Goller F, Larsen ON. A new mechanism of sound generation in songbirds. Proc Nat Acad Sci USA. 1997; 94:14787–14791. [PubMed: 9405691]

Goller F, Mallinckrodt MJ, Torti SD. Beak gape dynamics during song in the zebra finch. J Neurobiol. 2004; 59:289–303. [PubMed: 15146546]

Hahnloser RH, Kozhevnikov AA, Fee MS. An ultra-sparse code underlies the generation of neural sequences in a songbird. Nature. 2002; 419:65–70. [PubMed: 12214232]

Hall ML. A review of hypotheses for the functions of avian duetting. Behav Ecol Sociobiol. 2004; 55:415–430.

Hall, ML. A review of vocal duetting in birds. In: Naguib, M.Zuberbuhler, K.Clayton, NS., Janik, VM., editors. Advances in the study of behavior. Vol. 40. 2009. p. 67-121.

Hauser MD, Chomsky N, Fitch WT. The faculty of language: What is it, who has it, and how did it evolve? Science. 2002; 298:1569–79. [PubMed: 12446899]

Hsu A, Woolley SM, Fremouw TE, Theunissen FE. Modulation power and phase spectrum of natural sounds enhance neural encoding performed by single auditory neurons. J Neurosci. 2004; 24:9201–11. [PubMed: 15483139]

Kim G, Doupe A. Organized representation of spectrotemporal features in songbird auditory forebrain. J Neurosci. 2011; 31:16977–16990. [PubMed: 22114268]

Kriengwatana B, Escudero P, Kerkhoven AH, ten Cate C. A general auditory bias for handling speaker variability in speech? Evidence in humans and songbirds. Frontiers in Psychology. 2015; 6:1243. [PubMed: 26379579]

Kruuk, H. A study of predation and social behavior. Univ. Chicago Press; 1972. The spotted hyena.

Ladefoged, P. Vowels and consonants. Wiley-Blackwell; 2012.

Levrero F, Durand L, Vignal C, Blanc A, Mathevon N. Begging calls support offspring individual identity and recognition by zebra finch parents. C R Biologies. 2009; 332:579–589. [PubMed: 19520321]

Lieberma, Ph, Klatt, DH., Wilson, WH. Vocal tract limitations on vowel repertoires of rhesus monkey and other non-human primates. Science. 1969; 164:1185. [PubMed: 4976883]

Lohr B, Dooling RJ. Detection of changes in timbre and harmonicity in complex sounds by zebra finches (*Taeniopygia guttata*) and budgerigars (*Melopsittacus undulatus*). J Comp Psychol. 1998; 112:36–47. [PubMed: 9528113]

Marler P. Bird calls: Their potential for behavioral neurobiology. Ann N Y Acad Sci. 2004; 1016:31–44. [PubMed: 15313768]

McCowan LSC, Mariette MM, Griffith SC. The size and composition of social groups in the wild zebra finch. Emu. 2015; 115:191–198.

Mielke A, Zuberbühler K. A method for automated individual, species and call type recognition in free-ranging animals. Anim Behav. 2013; 86:475–482.

Morton ES. Ecological sources of selection on avian sounds. Am Nat. 1975; 109:17–34.

Morton ES. Occurrence and significance of motivation structural rules in some bird and mammal sounds. Am Nat. 1977; 111:855–869.

Mouterde SC, Elie JE, Theunissen FE, Mathevon N. Learning to cope with degraded sounds: Female zebra finches can improve their expertise in discriminating between male voices at long distances. J Exp Biol. 2014; 217:3169–3177. [PubMed: 24948627]

Mouterde SC, Theunissen FE, Elie JE, Vignal C, Mathevon N. Acoustic communication and sound degradation: How do the individual signatures of male and female zebra finch calls transmit over distance? PloS one. 2014; 9:e102842. [PubMed: 25061795]

Mulard H, Vignal C, Pelletier L, Blanc A, Mathevon N. From preferential response to parental calls to sex-specific response to conspecific calls in juvenile zebra finches. Anim Behav. 2010; 80:189–195.

Mundry R, Sommer C. Discriminant function analysis with nonindependent data: Consequences and an alternative. Anim Behav. 2007; 74:965–976.

Murphy, KP. Machine learning: A probalistic perspective. MIT Press; Cambridge, Massachusetts: 2012.

Nagel K, Doupe A. Organizing principles of spectro-temporal encoding in the avian primary auditory area field l. Neuron. 2008; 58:938–955. [PubMed: 18579083]

Nakagawa S, Hauber ME. Great challenges with few subjects: Statistical strategies for neuroscientists. Neurosci Behav Rev. 2011; 35:462–473.

O'Shaughnessy, D. Speech communication: Human and machine. Wiley-IEEE Press; 1999.

Ohms VR, Snelderwaard PC, ten Cate C, Beckers GJL. Vocal tract articulation in zebra finches. PloS one. 2010; 5(7):e11923. [PubMed: 20689831]

Olson, CR., Wirthlin, M., Lovell, PV., Mello, CV. Cold Spring Harb Protoc. 2014. Proper care, husbandry, and breeding guidelines for the zebra finch, *Taeniopygia guttata*.

Owren MJ, Rendall D. Sound on the rebound: Bringing form and function back to the forefront in understanding nonhuman primate vocal signaling. Evol Anthropol. 2001; 10:58–71.

Perez EC, Fernandez MSA, Griffith SC, Vignal C, Soula HA. Impact of visual contact on vocal interaction dynamics of pair-bonded birds. Anim Behav. 2015; 107:125–137.

Picone JW. Signal modeling techniques in speech recognition. Proc IEEE. 1993; 81:1215–1247.

Reby D, McComb K, Cargnelutti B, Darwin C, Fitch WT, Clutton-Brock T. Red deer stags use formants as assessment cues during intrasexual agonistic interactions. Proc Roy Soc B Biol Sci. 2005; 272:941–947.

Rendall D, Owren MJ, Ryan MJ. What do animal signals mean? Anim Behav. 2009; 78:233–240.

Riede T, Zuberbuhler K. The relationship between acoustic structure and semantic information in diana monkey alarm vocalization. J Acoust Soc Am. 2003; 114:1132–1142. [PubMed: 12942990]

Riede T, Suthers RA, Fletcher NH, Blevins WE. Songbirds tune their vocal tract to the fundamental frequency of their song. Proc Nat Acad Sci USA. 2006; 103:5543–5548. [PubMed: 16567614]

Riede T, Goller F. Peripheral mechanisms for vocal production in birds - differences and similarities to human speech and singing. Brain Lang. 2010; 115:69–80. [PubMed: 20153887]

Riede T, Schilling N, Goller F. The acoustic effect of vocal tract adjustments in zebra finches. J Comp Physiol A Neuroethol Sens Neural Behav Physiol. 2013; 199:57–69. [PubMed: 23085986]

Robisson P, Aubin T, Bremond JC. Individuality in the voice of the emperor penguin aptenodytes forsteri: Adaptation to a noisy environment. Ethology. 1993; 94:279–290.

Salmi R, Hammerschmidt K, Doran-Sheehy DM. Western gorilla vocal repertoire and contextual use of vocalizations. Ethology. 2013; 119:831–847.

Scharff C, Nottebohm F, Cynx J. Conspecific and heterospecific song discrimination in male zebra finches with lesions in the anterior forebrain pathway. J Neurobiol. 1998; 36:81–90. [PubMed: 9658340]

Searcy WA, Beecher MD. Song as an aggressive signal in songbirds. Anim Behav. 2009; 78:1281–1292.

Seyfarth RM, Cheney DL, Marler P. Monkey responses to 3 different alarm calls - evidence of predator classification and semantic communication. Science. 1980; 210:801–803. [PubMed: 7433999]

Seyfarth RM, Cheney DL. Signalers and receivers in animal communication. Ann Rev Psych. 2003; 54:145–173.

Seyfarth RM, Cheney DL, Bergman T, Fischer J, Zuberbühler K, Hammerschmidt K. The central importance of information in studies of animal communication. Anim Behav. 2010; 80:3–8.

Singh NC, Theunissen FE. Modulation spectra of natural sounds and ethological theories of auditory processing. J Acoust Soc Am. 2003; 114:3394–411. [PubMed: 14714819]

Sitt JD, Amador A, Goller F, Mindlin GB. Dynamical origin of spectrally rich vocalizations in birdsong. Phys Rev E. 2008; 78:011905.

Smith WJ. Animal duets - forcing a mate to be attentive. J Theo Biol. 1994; 166:221–223.

Stowell D, Plumbley MD. Automatic large-scale classification of bird sounds is strongly improved by unsupervised feature learning. CoRR. 2014 abs/1405.6524.

Sturdy CB, Phillmore LS, Price JL, Weisman RG. Song-note discriminations in zebra finches (*Taeniopygia guttata*): Categories and pseudocategories. J Comp Psychol. 1999; 113:204–212.

Taylor AM, Reby D. The contribution of source-filter theory to mammal vocal communication research. J Zool. 2010; 280:221–236.

Tchernichovski O, Nottebohm F, Ho CE, Pesaran B, Mitra PP. A procedure for an automated measurement of song similarity. Anim Behav. 2000; 59:1167–1176. [PubMed: 10877896]

Tchernichovski O, Mitra PP, Lints T, Nottebohm F. Dynamics of the vocal imitation process: How a zebra finch learns its song. Science. 2001; 291:2564–9. [PubMed: 11283361]

Ter Maat A, Trost L, Sagunsky H, Seltmann S, Gahr M. Zebra finch mates use their forebrain song system in unlearned call communication. PloS one. 2014; 9:e109334. [PubMed: 25313846]

Theunissen FE, Elie JE. Neural processing of natural sounds. Nat Rev Neurosci. 2014; 15:355–66. [PubMed: 24840800]

Thorpe WH. Duetting and antiphonal song in birds. Its extent and significance. Behaviour. 1972; 18:1–197.

Vicario DS, Naqvi NH, Raksin JN. Behavioral discrimination of sexually dimorphic calls by male zebra finches requires an intact vocal motor pathway. J Neurobiol. 2001; 47:109–120. [PubMed: 11291101]

Vignal C, Mathevon N, Mottin S. Audience drives male songbird response to partner's voice. Nature. 2004; 430:448–51. [PubMed: 15269767]

Vignal C, Mathevon N, Mottin S. Mate recognition by female zebra finch: Analysis of individuality in male call and first investigations on female decoding process. Behav Process. 2008; 77:191–198.

Wickler W, Seibt U. Vocal duetting and the pair bond. 2. Unisono duetting in the african forest weaver, symplectes-bicolor. J Comp Etholog. 1980; 52:217–226.

Wild JM, Kruetzfeldt NEO. Trigeminal and telencephalic projections to jaw and other upper vocal tract premotor neurons in songbirds: Sensorimotor circuitry for beak movements during singing. J Comp Neurol. 2012; 520:590–605. [PubMed: 21858818]

Williams H, Cynx J, Nottebohm F. Timbre control in zebra finch (*Taeniopygia guttata*) song syllables. J Comp Psychol. 1989; 103:366–80. [PubMed: 2598623]

Williams H. Birdsong and singing behavior. Ann N Y Acad Sci. 2004; 1016:1–30. [PubMed: 15313767]

Wilson, EO. Sociobiology, the new synthesis, twenty-fith anniversary edition. Harvard University Press; 2000.

Woolley SM, Fremouw TE, Hsu A, Theunissen FE. Tuning for spectro-temporal modulations as a mechanism for auditory discrimination of natural sounds. Nat Neurosci. 2005; 8:1371–9. [PubMed: 16136039]

Woolley SM, Gill PR, Fremouw T, Theunissen FE. Functional groups in the avian auditory system. J Neurosci. 2009; 29:2780–93. [PubMed: 19261874]

Young, S., Evermann, G., Gales, M., Hain, T., Kershaw, D., Liu, X., Moore, G., Odell, J., Ollason, D., Povey, D., Valtchev, V., Woodland, P. The htk book (for htk version 3.4.1). Engineering Department, Cambridge University; 2006.

Zann, RA. The zebra finch. Oxford University Press; Oxford, UK: 1996.

Zann, RA. The zebra finch: A synthesis of field and laboratory studies. Oxford University Press; Oxford: 1996.

Zuberbühler, K., Lemasson, A. Primate communication: Meaning from strings of calls. In: Lowenthal, F., Lefebvre, L., editors. Language and recursion. Springer-Verlag; New York, NY: 2014. p. 115-125.

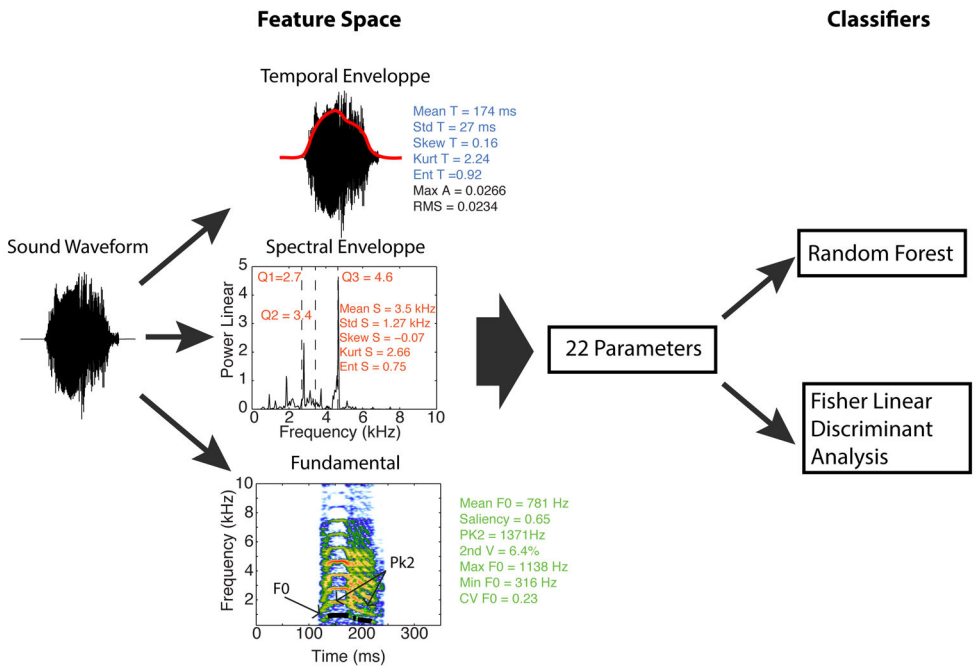

Figure 1. Extraction of the Predefined Acoustical Features (PAFs) and flow-chart showing the classification procedure using these parameters

Five acoustical parameters were obtained from the temporal amplitude envelope of the sound (middle top, in blue), two parameters characterized the amplitude of the signal (middle top, in black), eight acoustical parameters were derived from the spectral amplitude envelope (middle center, in red) and seven acoustical parameters described the time varying fundamental (bottom center, in green). The fundamental is shown as a black line on the spectrogram. The fundamental is only extracted when the pitch saliency is greater than 0.5. These 22 acoustical parameters were then used to train two classifiers in vocalization category discrimination: a Random Forest and a Fisher Linear Discriminant Analysis. Performance was assessed by cross-validation. See Methods for more details on the calculation of the parameters and on the classification procedure.

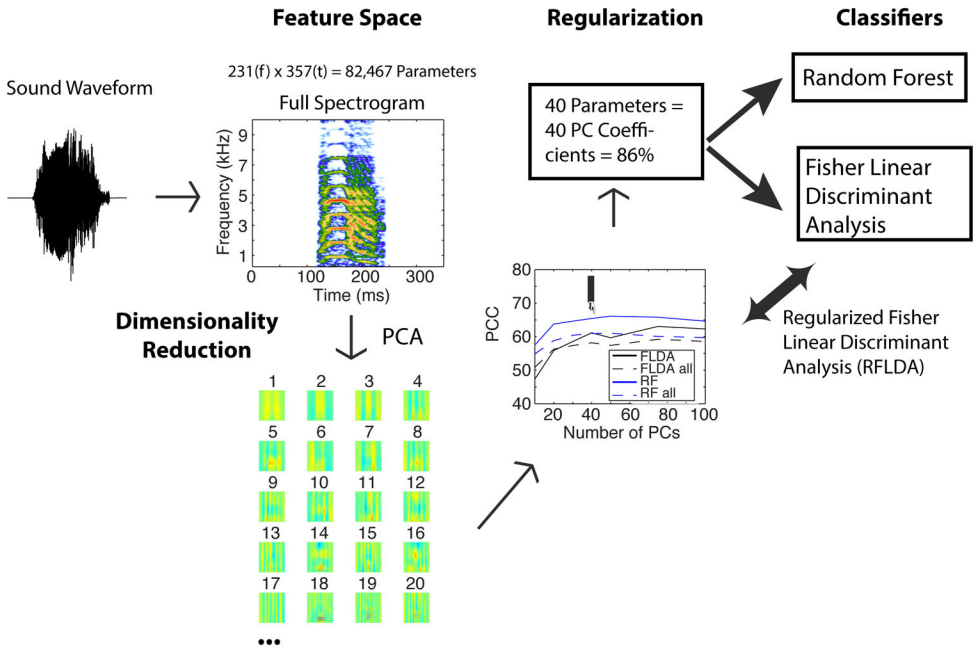

Figure 2. Flow-chart showing the regularized classification procedure using a complete and invertible spectrogram to represent the vocalizations

Here we performed a classification of vocalization category using an over-complete feature space representation of the sounds: an invertible spectrogram (top panel, Feature Space column). The invertible spectrogram had 231 frequency bands between 0 and 12 kHz (~52 Hz bandwidth) and a sampling rate of 1017 Hz yielding 357 points in time for the 350 ms window used to frame each vocalization. The total number of parameters describing the sounds in this spectrographic representation was 82,467. To prevent overfitting, we reduced the number of parameters using a principal component analysis (PCA). The first 20 PCs are shown as little spectrograms in the bottom row of the Feature Space column. The optimal number of PC coefficients was found by training the two classifiers with a varying number of PC coefficients and estimating the performance of the classifier using a cross-validation data set. The performance of the classifiers as a function of parameters is shown on the line plot in the Regularization column. RF = Random Forest, RFLDA = Regularized Fisher Linear Discriminant Analysis, PCC = Probability of Correct Classification. The solid lines correspond to the performance averaged first for each vocalization type and then average across all types. The dashed lines correspond to the average overall performance (the average across types weighted by the number of vocalizations in each category). Performance for all these measures plateaued or decreased at approximately 40 PCs. 40 PCs explained 87% of the overall variance in the spectrograms of all vocalizations. The classification results presented in detail in this paper were thus obtained by describing each sound with the coefficients of 40 PCs. See the Methods for more details on the spectrographic representation and on this regularized classification procedure.

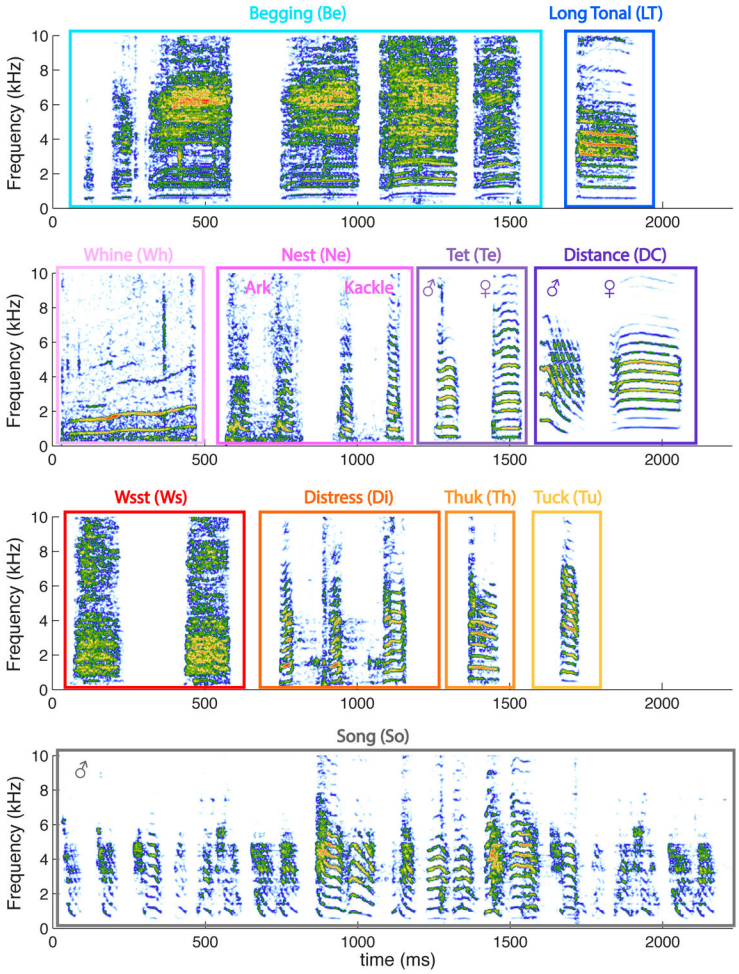

Figure 3. The zebra finch vocal repertoire

Spectrograms of examples of each vocalization type found in domesticated zebra finches. The top row shows the two types of calls produced solely by chicks: a Begging call bout and a Long Tonal call. The Long Tonal call is the precursor of the adult Distance call and functions as a contact call. The second row shows the calls produced by adults during affiliative or neutral behaviors. The Whine and the Nest calls are not only produced during early phases of pair bonding and nest building but also any time mates are relaying each other at the nest. The Tet call is a contact call produced for short-range communication while the Distance call is a contact call produced for long-range communication. Both are sexually dimorphic. The third row shows the calls produced during agonistic interactions or threatening situations by adults. The aggressive call, called the Wsst call, is made here of two syllables and is produced shortly before aggression of a conspecific. The Distress call made here of three syllables is produced by the victim during or just after the aggression. There are two alarm calls called the Thuk call, produced by parents and directed at chicks and mate, and the Tuck call, a more generic alarm call. Finally, an example of a Song, the more complex signal used by males in courtship, pair bonding and mating behavior is shown on the last row. The color code used in this figure categorizes the vocalization types into hyper classes: blue hues for chick calls, pink to deep purple hues for affiliative calls, red/

orange hues for non-affiliative calls and grey/black for song. The same color code is used in all the figures. For the spectrogram colors, vocalizations in each group (rows) where normalized to peak amplitude and a 40dB color scale was used. The sounds corresponding to these vocalizations can be found online as supplemental material (chick calls, Supplementary Sound File 1; affiliative calls, Supplementary Sound File 2; nonaffiliative calls, Supplementary Sound File 3; and song, Supplementary Sound File 4). The abbreviations used for each category in other figures are given in parenthesis.

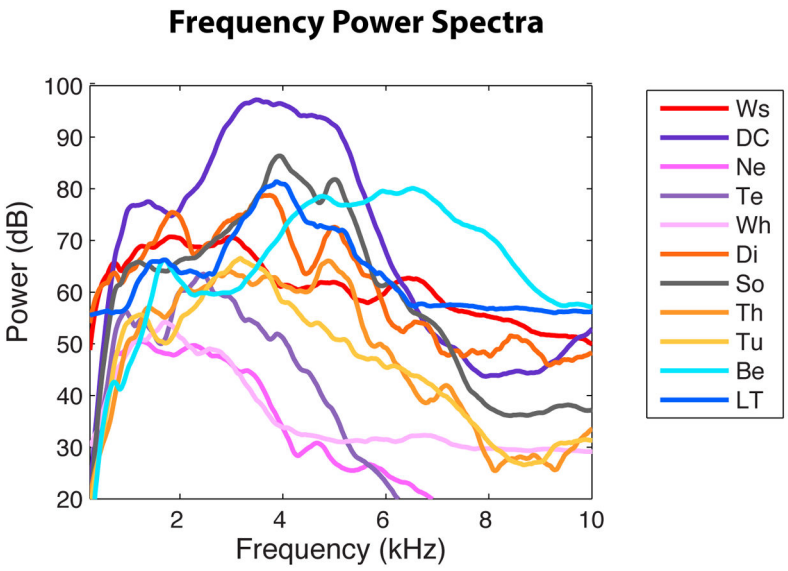

Figure 4. Average frequency power spectra

The average frequency power spectra for each vocalization type were obtained by first averaging the spectra of all vocalizations for each bird and each vocalization type, and then averaging across birds for each vocalization type. 100 dB corresponds to the peak amplitude recorded (found for Distance calls ~ 80 dB SPL at 20 cm). Abbreviations are defined in Figure 3.

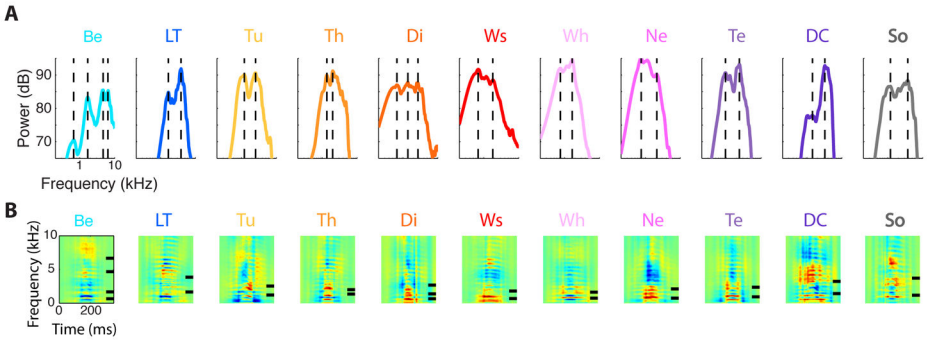

Figure 5. Formant peaks revealed by average frequency power spectra and Logistic Regression Functions (weights) for each vocalization type

The upper panel (A) shows the average power spectrum for each vocalization type in log frequency – dB scale. Compare to Figure 4, the average power spectra were calculated on normalized spectra (peak = 100 dB), by first averaging the normalized spectra of all vocalizations within the same category for each bird and then average over birds for each vocalization type. These power spectra show two and sometimes three peaks at reliable frequencies that we call formants, using its acoustical definition. The vertical dotted lines show the location of these spectral envelope peaks. (B) For each vocalization type, we performed a separate logistic regression to assess how well a particular category could be distinguished from all the others and to determine the spectro-temporal features that could best select one vocalization type over the others. The logistic regression was applied to the vocalizations in the spectrogram feature space and the weights of the regression are shown as spectrograms. On the right side of the spectrogram, short black lines indicate the formants found in the average power spectrum of each vocalization category as shown on A. Abbreviations are defined in Figure 3.

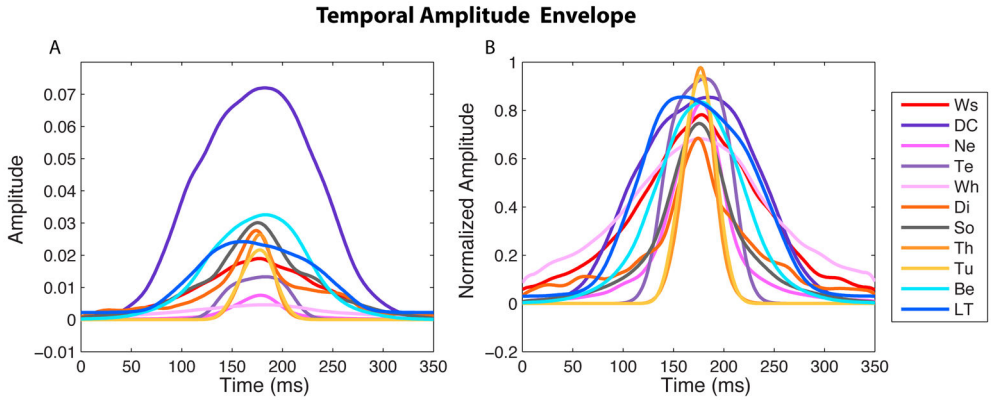

Figure 6. Temporal Amplitude Envelopes

The average temporal amplitude envelope for each vocalization type is shown raw on the left panel (A) and normalized by peak amplitude on the right (B). Note that the y-scale is linear and not logarithmic as in the frequency power spectra of Figure 4. These average envelopes were obtained by first averaging for each bird and vocalization type, and then across birds. Abbreviations are defined in Figure 3.

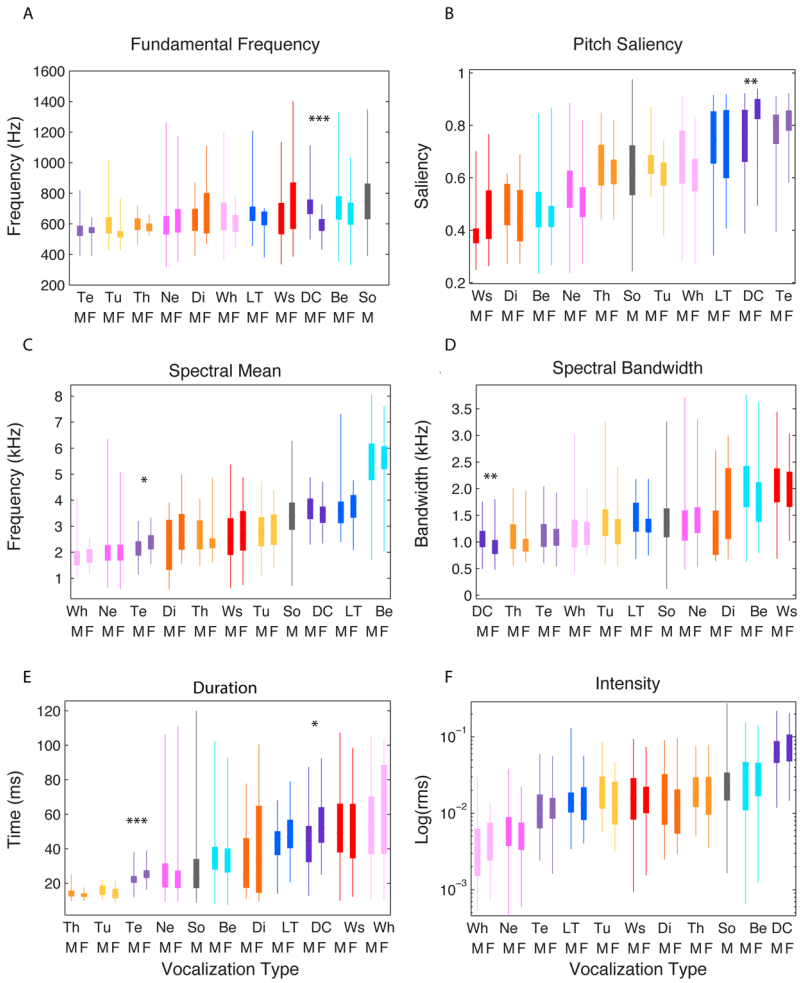

Figure 7. Box and whisker plots for 6 out of the 22 PAFs vs vocalization type

These parameters were chosen to illustrate the distinctive acoustical properties of vocalization categories. The bottom and the top of the solid rectangles correspond to the beginning and end of the 2nd and 3rd quartile and the whiskers show the entire range of values found in our data set. In all plots the vocalization types (shown in the x-axis) are ordered in increasing value of the corresponding acoustical feature to facilitate the interpretation of the results. Two acoustical properties related to the fundamental frequency are shown on the first row: the fundamental frequency F0 (A) and the saliency (B) defined as the proportion of sections of the vocalization with an auto-correlation peak amplitude at the periodicity period greater than 50% of the peak amplitude at zero. Two acoustical properties related to the spectral envelope are shown in the middle row: the spectral mean (C) and the spectral bandwidth (D). Finally, two additional properties, the duration (E) and the sound intensity (F) are shown in the third row. The *, **, *** indicate significant differences between male and female vocalizations for specific types in post-hoc tests with $p < 0.05$, <0.01 and < 0.001 correspondingly. Vocalization abbreviations are defined in Figure 3.

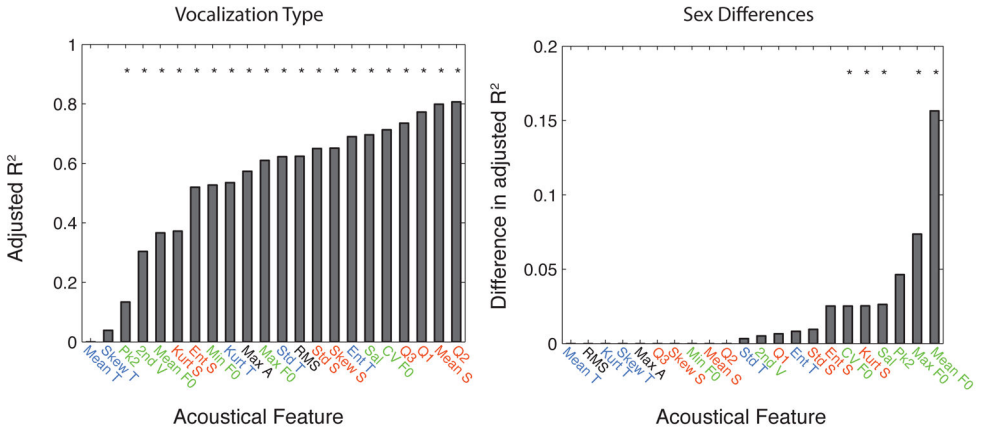

Figure 8. Variance explained by the Vocalization Type and additionally by Sex for each of the 22 PAFs

(A) The adjusted R2 is the fraction of the variance explained in a linear mixed-effects model with vocalization type as a predictor and bird identity as a random factor. (B) The difference in adjusted R2 is the difference in adjusted R2 obtained from the model that includes vocalization type and sex (including interactions) and the adjusted R2 obtained from the model that only includes vocalization type as a predictor. The color code is used to distinguish acoustical parameters that characterize the spectral envelope (red) from those that characterize the temporal envelope (blue), those that characterize the pitch of the sound (green) and those that characterize the intensity of the sound (black). The * indicate the values that were significantly different from zero with $p<0.05$. Note that a different y-scale is used in the two graphs.

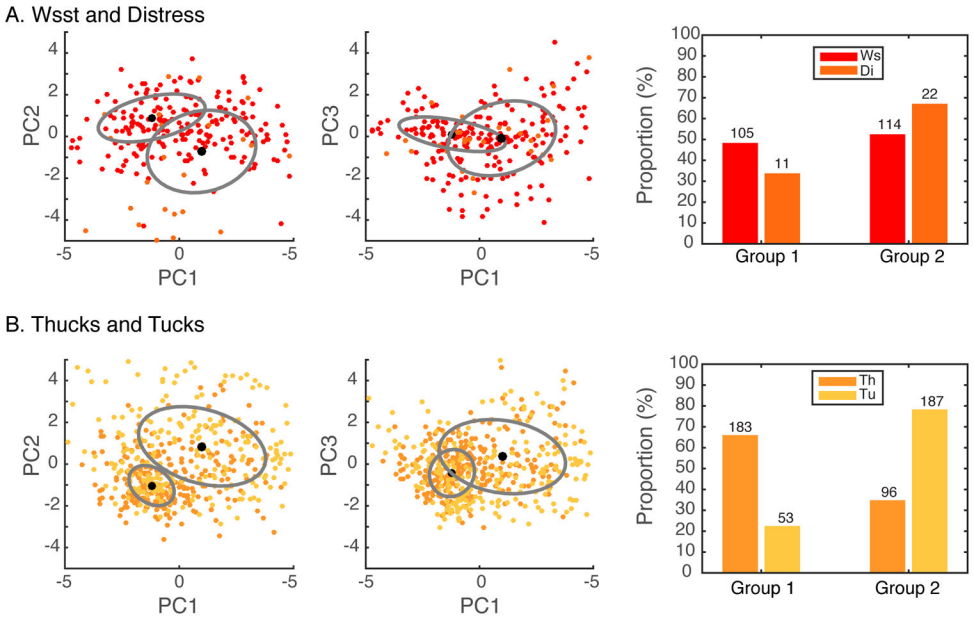

Figure 9. Unsupervised Clustering: Are Thuk acoustically distinct from Tuck calls, and Distress from Wsst calls?

A "mixture-of-Gaussians" model was fitted to the probability density distribution of the 10 first principal component coefficients derived for the 22 PAFs as defined on Figure 1 (see Methods). Each point in the raster plots on the left and middle column correspond to one vocalization. The points are color-coded according to call type but the mixture-of-Gaussians algorithm is blind to this information. The ellipses and center black dot show the covariance (at one standard deviation) and mean of the fit obtained from a mixture of two Gaussians model. The size of the center dot is proportional to the weight of the Gaussian. A. Distress and Wsst calls. We analyzed the shape of the distribution of Wsst and Distress calls as one group. Two Gaussians with similar weights (w1=0.4603, w2 = 0.5397) provided a good description of the distribution but the two call types were equally assigned to each of these two groups (z=1.57, p=0.12) suggesting that, at the level of a single call syllable, Distress and Wsst are acoustically similar. B. The second row shows the results of the same analysis for Thuk and Tuck calls. Here also two Gaussians with similar weights (w1=0.4591, w2 = 0.5409) fit the data well. The bar graph on the right panel shows the proportion (and raw number) of Thuk and Tuck that would be assigned to each of these two groups. The proportions are different in the two groups (z=9.92 p<10-4).

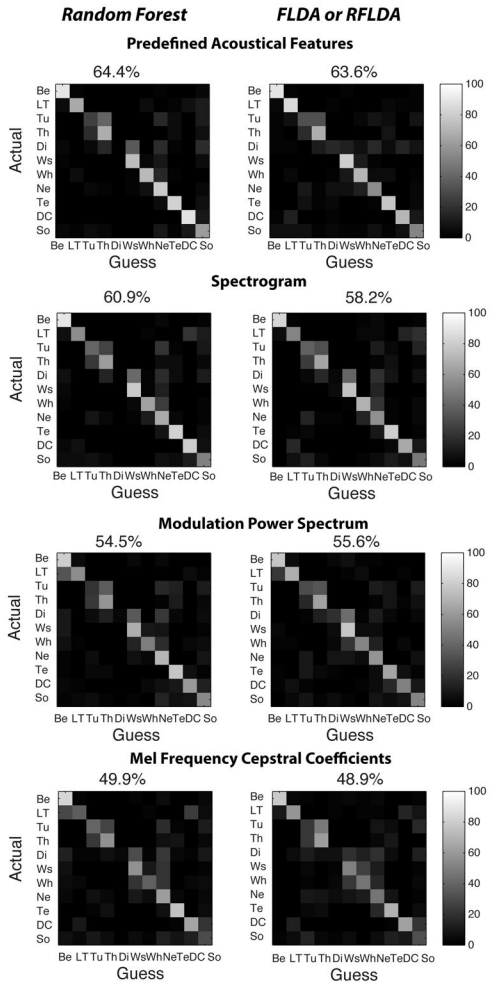

Figure 10. Confusion Matrices

The figure shows all the confusion matrices obtained from the Random Forest (left column) and Regularized Fisher Linear Discriminant Analysis (right column) for the four feature spaces used here and described on Figures 1, 2, Supplementary Figures 1 and 2. In a confusion matrix each row shows how exemplars from a particular vocalization category were classified into the categories shown in the columns. The color code is used to show the probability of that classification: the conditional probability of classifying a vocalization as type x (column x) when it is actually type y (row y). The classification is performed on a cross-validation dataset as explained in the methods. The average percentage of correct classification, obtained by averaging the diagonal of each matrix, is shown on the top of each confusion matrix. These numbers are used in the plot of Figure 11A. Abbreviations are defined in Figure 3.

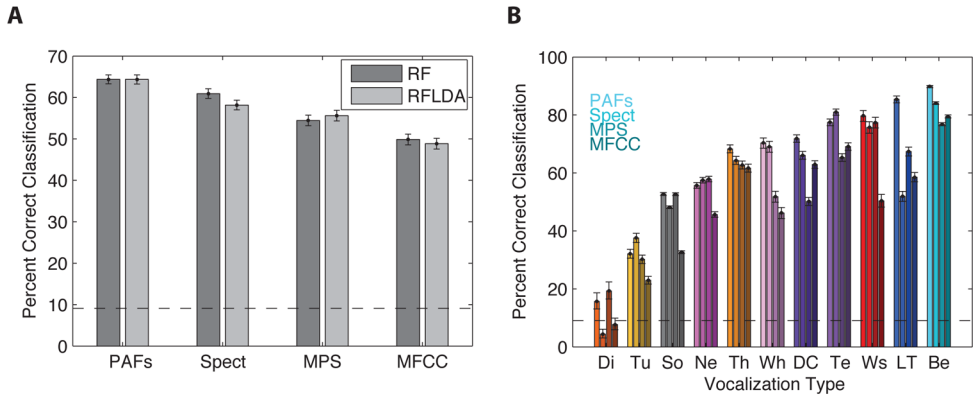

Figure 11. Performance of the Random Forest (RF) and Regularized Fisher Linear Discriminant Analysis (RFLDA) for all vocalization types and for each feature space

A. Average performance of each classifier across all vocalization types. The error bars are confidence intervals obtained from a binomial fit of the classification performance on cross-validated data. The dotted horizontal line is the chance level (1/11). B. Performance of the RFLDA for each vocalization type. A gradient of darkness (from light to dark) is used to represent the four feature spaces: Predefined Acoustical Features (PAFs), Spectrogram (Spect), Modulation Power Spectrum (MPS) and Mel Frequency Cepstral Coefficients (MFCC). The vocalization types on the x-axis are sorted in ascending order according to the percent of correct classification obtained with the Spectrogram feature space. Abbreviations are defined in Figure 3.

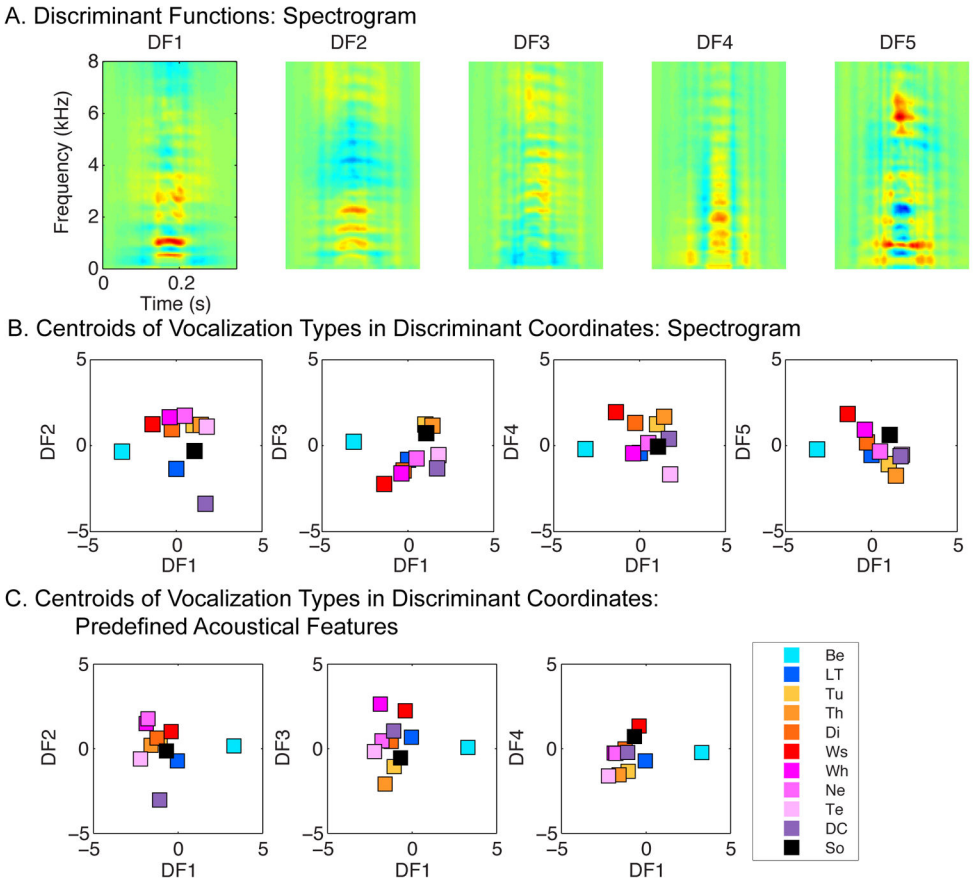

Figure 12. Discriminant Functions (A) and positions of vocalization types in Discriminant coordinates (B and C)

A. The first 5 discriminant functions (DF) obtained in the RFLDA applied to the spectrogram feature space. These discriminant functions are displayed in a spectrographic representation. Each vocalization was then represented in RFLDA coordinates by projecting its spectrogram onto these Discriminant functions (using a vector dot product). B and C. The average position of each vocalization type (centroid) is shown as a colored rectangle in coordinate-pair scatter plots. The DF have been scaled so that the within vocalization type variance along each discriminant dimension is equal to 1. In B the positions of the centroids obtained from the RFLDA applied on the spectrogram feature space are shown. In C the positions of the centroids obtained from the RFLDA applied on the Predefined Acoustical Features are shown. Vocalization abbreviations are defined in Figure 3.

Table 1

Vocalization names and number of calls or syllables and birds recorded in our zebra finch Vocalization Database

Vocalization Type	Abbreviation	# Sounds	# Birds
Wsst	Ws	235	23
Begging	Be	1824	15
Distance	DC	630	26
Distress	Di	51	11
Long Tonal	LT	217	13
Nest	Ne	1063	23
Song	So	2776	13
Tet	Te	613	24
Thuk	Th	290	13
Tuck	Tu	240	13
Whine	Wh	197	15

Table 2

Frequency of spectral peaks (formants, in kHz) in the average spectral envelope of each vocalization type

The location of these peaks is shown as dotted lines on Figure 5a.

	F1	F2	F3	F4
Begging	0.67	1.69	4.67	6.53
Long Tonal	1.67	3.86		
Tuck	1.22	2.55		
Thuk	1.36	2		
Distress	0.67	1.38	2.69	
Wsst	0.69	1.81		
Whine	0.76	1.65		
Nest	0.76	2.12		
Tet	0.95	2.38		
Distance	1.43	3.24		
Song	1.19	3.72		

Table 3

Description of the first four discriminant functions (DF) obtained from RFLDA in the PAF space

Coeff: coefficients of the 8 most important acoustical factors for each discriminant function; CumVar: cumulative between group variance explained by the discriminant functions. Note that the colors of the variables reflect the classification between spectral (red), temporal (blue), pitch (green) and amplitude (black) parameters.

DF	Cum Var	Coeff
1	55.5	2.859(Mean S) + 0.610(Skew S) - 0.594(Q2) - 0.585(Max A) +0.531(Q3) + -0.451(Q1) + -0.429(Sal) + 0.358(Std T)
2	73.9	1.335(Mean S) -1.332(RMS) + 1.055(Skew S) - 0.869(Q3) -0.768(Ent T) + 0.644(Kurt S) + 0.616(Max A) -0.482(Sal)
3	84.5	1.361(Ent T) + 0.712(RMS) + 0.692(Skew S) + 0.658(Q1) +0.654(Kurt T) -0.632(Max A) + 0.514(CV F0) + 0.495(Kurt S)
4	92.2	-0.885(Q3) + 0.841(Max A) + 0.757(Mean F0) - 0.648(Skew S) -0.623(RMS) + 0.617(CV F0) + 0.432(Std S) + 0.419(Ent S)