Locally Periodic Signal Estimation and Pitch Detection, and Acoustic Communication in American and Northwestern Crows

Exu Anton Mates

A dissertation
submitted in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

University of Washington
2016

Reading Committee:
Renee Ha, Chair
James Ha
Michael Beecher

Program Authorized to Offer Degree:
Psychology

University of Washington

Abstract

Locally Periodic Signal Estimation and Pitch Detection, and Acoustic Communication in American and Northwestern Crows

Exu Anton Mates

Chair of the Supervisory Committee:

Research Associate Professor Renee Robinette Ha, Ph.D.
Psychology

Crows and other members of the genus *Corvus* employ a complex system of acoustic communication, based largely on the calls known as "caws." These are locally periodic signals with rapidly varying pitch, and are difficult to analyze with conventional methods of pitch detection and periodic signal estimation. In this dissertation, I present novel methods of reconstructing locally periodic signals from noisy data, improve the performance of autocorrelation and the Pseudo-Maximum Likelihood Estimator as period indicator functions, and introduce averaged autocorrelation as a new period indicator which combines the advantages of autocorrelation and the PMLE. Using these signal analysis tools, I identify acoustic correlates of sex, behavioral context and individual identity in the caws of the American crow.

Finally, I demonstrate that mobbing behavior in the Northwestern crow can be explained by the adaptive functions of antipredator surveillance and communication, and that the "funeral" behavior exhibited by crows toward dead conspecifics may simply be another facet of mobbing behavior.

Dedication

For Robin, my childhood sweetheart, best friend, research colleague, editor, inspiration, frequent funding source, and once and future wife.

Acknowledgments

I would like to express my deep appreciation and gratitude to my primary advisors at the University of Washington, Drs. Jim and Renee Ha, who have mentored me since well before I actually applied to their department. Jim and Renee have been consistently supportive in both personal and academic areas, and demonstrated an uncanny ability to connect me with essentially any other professional on the planet whose acquaintance might possibly be beneficial. I am very glad to know them, and very fortunate to have worked with them.

Many thanks are also due to my coadvisor and committee member Dr. Mike Beecher, whose encyclopedic knowledge of the history and current state of the field of animal behavior, along with his commitment to maintaining the health of this program at UW, mean that he will unfortunately never be allowed to completely retire.

I am grateful to my senior labmate and coauthor Adrienne Sussman, who served as a role model and guide to both the city of Seattle and the intricacies of this graduate program, and to my other labmates, Carly Loyer, Saethra Fritscher, Kelsey McCune, Lindsey Nietmann, and Carol Xu, for providing a wealth of constructive criticism and intellectual stimulation.

I would also like to recognize:

The University of Washington's Graduate Opportunities and Minority Achievement Program (GO-MAP) for providing first-year and dissertation-year funding, networking opportunities, and advice and support toward the completion of my dissertation;

The National Institutes of Health, for providing me with two years of funding on an institutional training grant to the University of Washington?s department of Speech and Hearing Sciences;

Drs. Rebecca Heiss, Anne Clark and Kevin McGowan, of the Ithaca Crow Project at the State University of New York at Binghamton and Cornell University, for providing me with field research training and access to the most valuable research population of wild crows in the Western Hemisphere;

My advisor Dr. Ed Overman, and senior colleagues Drs. Karen Hallberg, Mitch Masters and John Wenzel, at the Ohio State University, for overseeing my mathematical training and assisting my first tentative forays into biology;

And Apollo, for turning the crow from white to black. It looks much handsomer that way.

Contents

CHAPTER 1

Introduction

1. Biographical details

I began my academic career in mathematics and physics, as an undergraduate at the University of California, Berkeley. In my third year, I participated in in the Oceanic Society's Midway Atoll Spinner Dolphin Research Project, under Susan Rickards, which led me to seek more research opportunities in animal behavior; other researchers suggested bioacoustics as a fruitful area to apply knowledge of linear algebra and harmonic analysis. I therefore applied for an internship in the Biosonar subprogram of the US Navy Marine Mammal Program, overseen by SPAWAR, under Patrick W. Moore and Dorian S. Houser. The objective of the program was to better understand the search strategies used by trained bottlenose dolphins in identifying underwater mines, so that an automated sonar-based system could be developed for the same purpose.

When applying to graduate school, I searched for a mathematics program at an institution which also did strong work in bioacoustics. I chose the Ohio State University, where I studied real and numerical analysis. Here I began independent research on detection and approximation of periodic signals, as well as research on the methodology of cladistic analysis with John Wenzel (Associate Professor, Departments of Entomology and Evolution, Ecology and Organismal Biology). Simultaneously, and in conjunction with my wife (Robin Tarter, a graduate student in animal behavior), I searched for a research organism with a native habitat that would be

slightly more accessible from a Midwestern university than is the bottlenose dolphin's. We eventually chose the American crow, and worked with the Ithaca Crow Project, headed by Kevin McGowan of Cornell and Anne B. Clark of SUNY Binghamton. I assisted with the capture, banding and wing-tagging of American and fish crows, censused family territories and foraging grounds, and made extensive acoustic recordings and behavioral observations.

Over time my research interests became centered on biology, and in 2007 I decided to pursue a Ph.D. in animal behavior instead of mathematics. In 2009, I applied for and was accepted to the animal behavior program in the University of Washington's psychology department, with advisors James C. Ha, Renee R. Ha, and Michael Beecher. At the University of Washington, I have continued my mathematical research on periodic signals, applying the resultant techniques to characterize the calls of American crows recorded in Ithaca. This latter work was published in *Bioacoustics* as "Acoustic profiling in a complexly social species, the American crow: caws encode information on caller sex, identity and behavioural context" (Mates et al. 2015). I have also studied alarm behavior in the Northwestern crows of the Puget Sound region, using natural and synthesized crow calls as stimuli.

Finally, I have worked with Dr. James Ha and Adrienne Sussman, studying the effects of sex, age, and tenure of captivity on temperament traits in captive pig-tailed macaques. This work was published in *Animal Behaviour* as "Tenure in current captive setting and age predict personality changes in adult pig-tailed macaques" (Sussman et al. 2014). My contribution to the study centered on statistical analysis, in particular the design and construction of linear mixed models to fit highly clustered data. This experience proved invaluable in designing models to study the acoustic data I gathered in my own fieldwork.

2. Dissertation structure

This dissertation is concerned with two primary topics: The analysis of locally periodic and quasi-periodic signals, and vocal communication in Northwestern and American crows. As my graduate research spans multiple fields, I have ordered the dissertation conceptually rather than chronologically.

In Chapters 2 and 3, I present a variety of novel results in the analysis of locally periodic and quasi-periodic signals, with particular applications to bioacoustics. Chapter 2 is concerned with estimation of locally periodic and quasi-periodic signals; here I present accelerated methods for least-squares approximation of such signals, and define a minimal notch filter (MNF) which has zero cross-correlation with all signals made up of a given frequency set. Chapter 3 contains research on pitch indicator functions, including improvements to autocorrelation and to the Pseudo-Maximum Likelihood estimator; the introduction of averaged autocorrelation; and the interpolation of autocorrelation between integer-valued periods.

In Chapter 4, I reproduce Mates et al. (2015), concerning acoustic variation in the "caw" calls of the American Crow. I demonstrate that certain call parameters indicate caller sex, identity, and behavioral context.

In Chapter 5, I report my field research on the effects of visual and auditory stimuli on alarm behavior in the Northwestern crow. The results of two playback studies suggest that mobbing behavior in this species functions largely to inform the mobbing bird about the location and threat level of potential predators, and to communicate this information to conspecifics. I argue that mobbing and "funeral" behavior are two variants of a single behavior pattern, and I find that a decoy of a dead conspecific can actually elicit more intense mobbing behaviors than a decoy of a live predator.

Finally, the appendices include supplementary mathematical results relevant to chapters 2 and 3, and a copy of Sussman et al. (2014).

References

Mates, Exu Anton et al. (2015). "Acoustic profiling in a complexly social species, the American crow: caws encode information on caller sex, identity and behavioural context". In: *Bioacoustics* 24.1, pp. 63–80. ISSN: 0952-4622. DOI: 10.1080/09524622.2014.933446. URL: http://www.tandfonline.com/doi/abs/10.1080/09524622.2014.933446.

Sussman, Adrienne F. et al. (2014). "Tenure in current captive setting and age predict personality changes in adult pigtailed macaques". In: *Animal Behaviour* 89, pp. 23–30. ISSN: 00033472. DOI: 10.1016/j.anbehav.2013.12.009. URL: http://dx.doi.org/10.1016/j.anbehav.2013.12.009.

CHAPTER 2

Least-Squares Approximation of Locally Periodic Signals

1. Introduction

A signal, in the context of signal processing, has been described very generally as "a function that conveys information about the behavior of a system or attributes of some phenomenon." (Priemer 1991) An acoustic (sound) signal, for instance, is a function that describes the pressure, displacement or particle motion in a medium as a function of time and space. Such signals can describe human speech, animal calls, and mechanical vibrations from nonbiological processes. For the remainder of this chapter and the next, I will be discussing signals that are functions of one dimension– typically time, although none of the mathematical results require time to be the independent variable.

An important and diverse category of such signals is that of the periodic and locally periodic signals. A periodic signal consists of precisely repeated cycles of a particular pattern, such as a sine or sawtooth wave. A *locally* periodic signal appears to be periodic when brief segments of it are viewed, but the length and pattern of its cycles may gradually change over time. In the case of a sound, these factors are largely responsible for the experiences of pitch and timbre. Locally periodic signals are very common in nature; almost any cyclic process, such as rotation or vibration, will produce them. Many biological sound generators, such as the mammalian larynx, the avian

syrinx, and insectile stridulatory structures, generate acoustic signals that are in large part locally periodic.

To characterize a (locally) periodic signal or estimate it from noisy data, one must establish its period of repetition and the pattern of the signal within each cycle. (If the signal is periodic, the period and pattern will be constants; if it is only locally periodic, they will be smooth functions of time.)

In this chapter, I will address the second of these tasks: estimating the signal once the period is known . This is typically done by breaking the signal into short frames and applying methods such as Fourier decomposition, bank filtering or least-squares approximation to each frame. Here, I will focus on least-squares approximation, which can be used for signals with periods that drift significantly within even a single frame. While the accuracy of least-squares approximation is high, particularly if the noise contaminating the signal is known to be independent and identically distributed, it is one of the slowest methods of signal estimation because it generally involves Gaussian elimination or matrix inversion.

The goal of this chapter is to provide faster algorithms for least-squares approximation of locally periodic or quasi-periodic signals, adapted from the Lanczos algorithm to find the eigendecomposition of a matrix (Lanczos 1950). Additionally, I define and compute a small filter that can be used to test for periodicity or quasi-periodicity, the Minimal Notch Filter.

2. Useful definitions: What is a (locally) periodic signal?

For continuous signals, defining a periodic signal $s(t)$ is simple: the signal must take the same value for any two values of its argument which differ by some value p: $s(t) = s(t + p)$. By this definition, a periodic signal will have multiple periods; for instance, a signal that repeats every 5 seconds also

repeats every 10 seconds, and every 15 seconds, and so forth. If the signal is non-constant and piecewise continuous, it will have a smallest positive period; this is called the signal's *fundamental* period. F_0, the reciprocal of the fundamental period, is called the signal's *fundamental frequency*; it satisfies $s(t) = s(t + \frac{1}{F_0})$ and is, of course, the largest possible value for which this holds.

While real-world signals are usually generated by continuous processes, such a signal is never *measured* on a continuous domain; it is sampled at a set of discrete points, and its value is known only at those points. For the purpose of this discussion, we'll confine ourselves to discrete signals that are sampled *uniformly*–that is, at equally-spaced points–in one dimension, which we'll call time. The *sampling period* of such a signal is the time difference between successive samples; the *sampling rate* is its reciprocal, the number of samples per unit time.

If a continuous periodic signal's fundamental frequency happens to be a unit fraction of the sampling rate (or, equivalently, if the signal's period is a multiple of the sampling period), then the above definition of periodicity can still be used for the discretely sampled version of that signal. But if this is not the case, the that definition becomes unusable, because the discrete signal's value will never be known or defined at both time t and time $t + p$. In this case, we must use a more elaborate set of definitions to capture periodicity.

Let us define a *quasi-periodic* signal as a finite linear sum of sinusoids or complex exponentials, called partials or harmonics. (The closure of the class of quasi-periodic signals, under uniform convergence, gives the class of almost periodic signals *sensu* Bohr (1925).) Define the signal's *partial frequency set* as the set of frequencies of the complex exponentials in question. If the signal is sampled on a discrete and finite domain, we will additionally

require that its partial frequency set have a cardinality less than that of the domain: e.g., if a quasi-periodic signal is sampled at 512 points, it must be decomposable into less than 512 partials. Without this latter requirement, all signals on finite domains would be quasi-periodic.

A quasi-periodic signal $s(t)$ has a *frequency bandwidth* B, if it can be represented as a sum of complex exponential harmonics with frequencies that do not exceed B. That is, $s(t) = \sum_{k=-n}^{n} c_k \psi_k(t)$, where $\psi_k(t) = e^{2ki\pi F_0}$ is the kth harmonic of the fundamental and $n = \lfloor \frac{B}{F_0} \rfloor$.

A quasi-periodic signal is *effectively periodic* if its partial frequency set is commensurable, i.e., all frequencies are integer multiples of some F_0. The largest such F_0 will be called the signal's *effective fundamental frequency*. And a signal is *antiperiodic* with respect to F_0 if its partial frequency set consists of *odd half-integer* multiples of F_0, e.g. $\pm\frac{1}{2}F_0, \frac{1}{3}F_0, /dots$.

On a continuous domain, a signal is effectively periodic iff it is periodic and has a finite bandwidth. However, on discrete domains, an effectively periodic signal is not necessarily periodic; instead, it may fall under the case already described, of a continuous signal that has been discretized at a sampling rate which is not an integer multiple of the fundamental frequency. Thus, effective periodicity is the appropriate generalization of periodicity for discrete signals sampled from real-world data. Henceforth, I will treat "periodic" and "effectively periodic" as synonymous when discussing such signals.

It will also prove useful to define the property of *nil-periodicity*. A signal is nil-periodic, relative to a given partial frequency set, if it is orthogonal to all partials in that set. (Orthogonality, in turn, must be defined with reference to a particular domain and inner product.)

Lastly, we will define a *locally periodic* signal as one for which, at any point in time, a short frame about that time can be chosen over which the

signal has a periodic (or effectively periodic) approximation with low error. Whereas an effectively periodic signal has a single fundamental frequency, a locally periodic signal has a *pitch trace*, a function relating fundamental frequency to time. The value of the pitch trace at any point in time is equal to the fundamental frequency of the best periodic approximation to a frame of the signal centered on that point. We can also define locally quasi-periodic signals in an analogous fashion, although they will not have a definable pitch trace.

3. Least-squares projection of signals to effectively periodic subspaces

3.1. Review: methods of least-squares projection. Having defined locally periodic and locally quasi-periodic signals, we now consider how to estimate such a signal from noisy data when its partial frequency set is already known. If our goodness-of-fit criterion is minimal mean squared error, then this is a least-squares linear projection problem, from the space of all signals on the (discrete) time domain, to the subspace of all quasi-periodic signals with a given set of partial frequencies.

In the simplest case, suppose we wish to estimate a single frame of a periodic signal with integer period p; that is, a signal which repeats itself every p samples. If our recorded signal is overlain with independently and identically distributed additive noise, then the signal's least-squares projection onto the period-p subspace is simply the *periodic average* of the recording: that is, the value of every sample is replaced with the average of that sample and the samples separated from it by kp samples, for all integer k. This estimator was introduced by David Slepian in an unpublished memorandum (Noll 1970).

If the noise level is not identically distributed but instead varies predictably across the frame, or if we expect the error in the measured signal to vary in a predictable fashion for some other reason, then we can use a *window function* to weight the periodic average:

$$\hat{s}[t] = \frac{\sum_k w[t+kp]s[t+kp]}{\sum_k w_{t+kp}} \tag{2.1}$$

The resultant signal will be the periodic approximation which minimizes the sum of the squared difference between it and the original signal, multiplied by the window function. If the value of the window function is inversely proportional to the expected RMS of the error of each sample, then this periodic average is also the *maximum likelihood estimate* of the periodic signal (Friedman 1977; Wise, Caprio, and Parks 1976). An additional advantage of using a window function is that formulae involving individual frames of the signal no longer need to refer directly to the frame length or interval; these values are implicit in the support of the window function.

What if the signal is effectively periodic or locally periodic, but its period is not an constant integer number of samples? In this case, we do not have the necessary data values to compute a periodic average. However, least-squares projection is still mathematically straightforward, provided the bandwidth B of the signal does not exceed the Nyquist frequency.

Suppose we believe the true (denoised) signal $\mathbf{s}$ to have instantaneous frequency $f(t) = \frac{1}{p(t)}$; let $\phi(t)$ be an antiderivative of $f(t)$. (If the signal is effectively periodic, $f[t]$ is a constant and $\phi[t]$ is a linear function.) Then the signal can be written as $\mathbf{s}[t] = \sum_{k=-n}^{n} c_k \psi_k(t)$, where $\psi_k[t] = e^{2ki\pi\phi[t]}$ is the kth harmonic of the fundamental and $n = \lfloor \frac{B}{max(|\phi'(t)|} \rfloor$). (Note that

n is chosen so that no harmonic attains a frequency higher than B over the frame.)

Estimating the true signal from the recorded signal s is then a matter of finding the harmonic coefficients c_k. Suppose the recorded signal is sampled at integer values of t, and we have chosen a window function $w(t)$ with support $\{0, 1 \cdots L-1\}$.

Let $\mathbf{W} = \mathrm{diag}(\mathbf{w}) = \begin{pmatrix} w_1 & 0 & \dots & 0 \\ 0 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \dots & 0 & w_L \end{pmatrix}$ be the diagonalized matrix of window coefficients, which we will use as a weighting matrix, and $\mathbf{\Psi}$ be the $L \times (2n+1)$ matrix of complex harmonics, with elements $\Psi_{jk} = e^{2(k-(n+1))i\pi\phi(j)}$. Then the normal equation for weighted least squares approximation is

$$(\mathbf{\Psi}^*\mathbf{W}\mathbf{\Psi})\,\mathbf{c} = \mathbf{\Psi}^*\mathbf{W}\mathbf{s} \tag{2.2}$$

This can be simplified by defining $\sqrt{\mathbf{W}}$ as the element-wise (positive) square root of $\mathbf{W}$. Then define $\tilde{\mathbf{s}} = \sqrt{\mathbf{W}}\mathbf{s}$ and $\tilde{\hat{\mathbf{s}}} = \sqrt{\mathbf{W}}\hat{\mathbf{s}}$, and define the matrix $\mathbf{A} = \sqrt{\mathbf{W}}\mathbf{\Psi}$, the matrix of weighted partials. Then we may rewrite the normal equation as $(\mathbf{A}^*\mathbf{A})\,\mathbf{c} = \mathbf{A}^*\tilde{\mathbf{s}}$, and its solution as $\mathbf{c} = \mathbf{A}^+\tilde{\mathbf{s}}$ and therefore $\tilde{\hat{\mathbf{s}}} = \mathbf{A}\mathbf{A}^+\tilde{\mathbf{s}}$, where $\mathbf{A}^+$ represents the Moore-Penrose pseudoinverse. Multiplying both sides of the former equation by $\mathbf{W}^{-1}$ gives the following equation for the estimated signal:

$$\hat{\mathbf{s}} = \mathbf{W}^{-1}\mathbf{A}\mathbf{A}^+\tilde{\mathbf{s}} \tag{2.3}$$

3.2. A rapid algorithm for orthonormalizing a basis of windowed partials. Now a practical consideration intervenes. To calculate $\hat{\mathbf{s}}$ according to Equation 2.3, it is necessary to calculate $\mathbf{A}^+$–or, if it is faster,

to directly calculate some larger factor of the right-hand product which contains $\mathbf{A}^+$. Most such computations are costly, with complexity on the order of Ln^2 flops, which is extremely high compared to the case of constant integer period. However, there is one exception.

$\mathbf{A}\mathbf{A}^+$ is an orthogonal projection matrix, with range spanned by the columns of $\mathbf{A}$. If an orthonormal basis $\mathbf{Q}$ can be found for this range, then $\mathbf{A}\mathbf{A}^+ = \mathbf{Q}\mathbf{Q}^*$.

Now, $\mathbf{A}$ is a rank $2n+1$ Krylov matrix generated by the vector $\mathbf{W}\psi_{-n}$, and the square diagonal matrix $\mathbf{D} = \begin{pmatrix} \psi_1[0] & 0 & \dots & 0 \\ 0 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \dots & 0 & \psi_1[L-1] \end{pmatrix}$. The range of $\mathbf{A}$ is the corresponding Krylov subspace. Because $\mathbf{D}$ is both diagonal and unitary, an orthonormal basis of this subspace can be found via a Lanczos-like algorithm of complexity Ln flops. I describe this algorithm (3.2) below.

The original Lanczos algorithm is a method of computing an orthogonal basis B for the Krylov subspace generated by a vector and a Hermitian matrix (Lanczos 1950). Like the more general Gram-Schmidt algorithm, it works by taking a sequence of vectors spanning the subspace, recursively computing the component of each vector which is orthogonal to all current members of the set B, and adding that component to B. However, whereas the Gram-Schmidt algorithm uses a predetermined set of spanning vectors, the Lanczos algorithm generates each successive spanning vector by taking the product of the generating matrix and the most recently calculated member of B. Because the generating matrix is Hermitian, each new spanning vector is already orthogonal to almost all of the basis vectors previously

computed, and no more than two such basis vectors need to be considered on each step.

In the current case, we wish to compute an orthonormal basis of $\mathbf{A}$. The original Lanczos algorithm does not apply, because $\mathbf{D}$ is not Hermitian. However, $\mathbf{D}$ is both diagonal and unitary, so it remains the case that most of the previously computed basis vectors will be orthogonal to the new spanning vector; it is simply that the few exceptions are the basis vectors that were computed most recently, rather than the ones that were computed first. Algorithm 3.2 exploits these properties.

The conceptually simplest form of the proposed algorithm would multiply successive basis vectors by $\mathbf{D}$ directly. However, $\mathbf{D}$ is complex-valued, and complex products are generally at least three times slower to compute than are real products. Since, for acoustic processing, we are generally working with real-valued data, I present a form of this algorithm that uses only real values.

Let $\boldsymbol{\phi}$ be a length-L vector of real phase values. Let $\mathbf{c}$ be a length-L complex vector, and n be an integer that does not exceed $(L-1)/2$. Let $\mathbf{S}_t = \left\{\mathbf{c} \circ e^{\circ ti\boldsymbol{\phi}} | t = -m \dots m\right\}$, where $\circ$ denotes the Hadamard (elementwise) product or power of vectors. In other words, $\mathbf{S}_t$ is the space of trigonometric polynomials in ϕ with degree at most m, multiplied elementwise by $\mathbf{c}$. We wish to construct an orthonormal basis for the subspace $\mathbf{S}_n$.

We wish to construct an orthonormal basis $\mathbf{Q}$ of S_n, which would also be an orthonormal basis of $\mathbf{A}$. This will consist of vectors $\hat{\mathbf{a}}_0 \dots \hat{\mathbf{a}}_N$ and $\hat{\mathbf{b}}_1 \dots \hat{\mathbf{b}}_N$.

Each round of this iteration requires order L flops to perform inner products and element-wise multiplications and divisions on length-L vectors, so the entire algorithm requires order LN flops to produce $\mathbf{Q}$. Now we

Algorithm 3.1 Pseudocode for constructing orthonormal basis for a locally periodic subspace

$\mathbf{a}_0 \leftarrow \mathbf{c}$ ▷ Generating $\hat{\mathbf{a}}_0$
$\alpha_0 \leftarrow |\mathbf{a}_0\|$
$\hat{\mathbf{a}}_0 \leftarrow \mathbf{a}_1/\alpha_1$; ▷ Generating $\hat{\mathbf{a}}_1$ and $\hat{\mathbf{b}}_1$
$\mathbf{p}_1 \leftarrow \mathbf{x} \circ \hat{\mathbf{a}}_0$;
$\mathbf{q}_1 \leftarrow \mathbf{y} \circ \hat{\mathbf{a}}_0$;
$\gamma_1 \leftarrow \mathbf{p}_1 \cdot \hat{\mathbf{a}}_0$;
$\delta_1 \leftarrow \mathbf{q}_1 \cdot \hat{\mathbf{a}}_0$;
$\mathbf{a}_1 \leftarrow \mathbf{p}_1 - (\gamma_1 \hat{\mathbf{a}}_0)$;
$\alpha_1 \leftarrow \|\mathbf{a}_1\|$
$\hat{\mathbf{a}}_1 \leftarrow \frac{\mathbf{a}_1}{\alpha_1}$;
$\zeta_1 \leftarrow \mathbf{q}_1 \cdot \hat{\mathbf{a}}_1$;
$\mathbf{b}_1 \leftarrow \mathbf{q}_1 - (\delta_1 * \hat{\mathbf{a}}_0 + \zeta_1 \hat{\mathbf{a}}_1)$;
$\beta_1 \leftarrow \|\mathbf{1}_1\|$
$\hat{\mathbf{b}}_1 \leftarrow \frac{\mathbf{b}_1/}{\beta_1}$; ▷ Generating remaining $\hat{\mathbf{a}}_k$ and $\hat{\mathbf{b}}_k$
for $k = 2 \rightarrow N$ **do**
  $\mathbf{p}_k = \mathbf{x} \circ \hat{\mathbf{a}}_{k-1}$;
  $\mathbf{q}_k = \mathbf{x} \circ \hat{\mathbf{b}}_{k-1}$;
  $\gamma_k = \mathbf{p}_k \cdot \hat{\mathbf{a}}_{k-1}$;
  $\delta_k = \mathbf{q}_k \cdot \hat{\mathbf{a}}_{k-1}$;
  $\epsilon_k = \mathbf{q}_k \cdot \hat{\mathbf{b}}_{k-1}$;
  $\mathbf{a}_k = \mathbf{p}_k - (\gamma_k \hat{\mathbf{a}}_{k-1} + \delta_k \hat{\mathbf{b}}_{k-1} + \alpha_{k-1} \hat{\mathbf{a}}_{k-2} + \zeta_{k-1} \hat{\mathbf{b}}_{k-2})$;
  $\alpha_k = \|\mathbf{a}_k\|$
  $\hat{\mathbf{a}}_k = \mathbf{a}_k/\alpha_k$;
  $\zeta_k = \mathbf{q}_k \cdot \hat{\mathbf{a}}_k$;
  $\mathbf{b}_k = \mathbf{q}_k - (\delta_k * \hat{\mathbf{a}}_{k-1} + \epsilon_k \hat{\mathbf{b}}_{k-1} + \beta_{k-1} \hat{\mathbf{b}}_{k-2} + \zeta_k \hat{\mathbf{a}}_k)$;
  $\beta_k = \|\mathbf{b}_k\|$
  $\hat{\mathbf{b}}_k = \mathbf{b}_k/\beta_k$;
end for

replace $\mathbf{A}\mathbf{A}^+$ with $\mathbf{Q}\mathbf{Q}^*$ in Equation 2.3 to yield the final equation for the estimated signal:

$$\hat{\mathbf{s}} = \mathbf{W}^{-1}\mathbf{Q}\mathbf{Q}^*\tilde{\mathbf{s}} \tag{2.4}$$

The left side of Equation 2.4 can be calculated in order Ln flops by calculating the matrix-vector multiplication from right to left.

4. Introduction to the Minimal Notch Filter (MNF)

In the section above, I discussed how to estimate a locally periodic signal by constructing a basis for the subspace of all locally periodic signals with a given pitch trace, then projecting the measured signal onto that subspace. A second approach, if the signal is quasi-periodic, would be to to construct a basis for the subspace of all signals that are *nil*-periodic with respect to a given partial frequency set. Since, for any such set, any signal can be decomposed into a quasi-periodic and a nil-periodic component, this amounts to estimating the error term of the quasi-periodic approximation instead of estimating the approximation itself. The error term can then be subtracted from the original signal to give the quasi-periodic approximation.

Given a set of signed partial frequencies (not exceeding the Nyquist frequency) and inner product, we define its *minimal notch filter*, or MNF, as the signal F with the narrowest support (on [0,1,...N] for some positive integer N) that is orthogonal to all partials and has F(0)=1.

Since translating (time-delaying) any complex exponential is equivalent to multiplying it by a scalar, all translates of the MNF are also orthogonal to the partials; that is, the cross-correlation of the MNF with the partials is zero everywhere. Accordingly, the MNF can also be thought of as the impulse response for the lowest-order filter which has notches at each frequency in the partial set.

This property continues to hold in the case of finite signals; on a finite interval, a signal is quasi-periodic with a given partial frequency set iff its cross-correlation on that interval with the MNF is zero.

4.1. Construction of the MNF. The MNF for an arbitrary signed frequency set can be constructed by recursive convolution, using the following rules:

(1) The MNF for a single signed frequency f is the 2-element sequence $[1, -e^{-2i\pi f}]$;

(2) The MNF for a pair of signed frequencies $f, -f$ is the 3-element sequence $[1, \frac{-2}{\cos 2\pi f}, 1]$.

(3) The MNF for the union of two disjoint signed frequency sets S_a and S_b is the convolution of the MNFs for each frequency set.

(4) The MNF for a shifted frequency set $\{f_k+c\}$ is equal to the MNF for the original frequency set $\{f_k\}$, multiplied by the complex sinusoid $e^{-2i\pi ct}$

In general, a set of N signed frequencies will have an MNF with support $N+1$ in length. The support cannot be shorter than this, because any N complex exponentials with frequencies below the Nyquist frequency are linearly independent on the domain [0,1,...N-1]. (See Appendix A for proof.) Thus, it is impossible for a signal with support on that domain to be orthogonal to all such complex exponentials.

4.2. The Inverse MNF. As described in the previous subsection, MNFs for disjoint partial frequency sets can be combined through convolution. The identity function for convolution of discrete signals is the discrete impulse function $\delta[n]$, which is equal to 1 for $n = 0$ and vanishes for all other values of n. Two signals are convolutive inverses of each other if their convolution is equal to $\delta[n]$. Given a partial frequency set, we will define its canonical *inverse MNF* as the *particular* convolutive inverse of its MNF which is zero for all negative values of n.

Again, given a set S of signed partial frequencies and inner product, let $f[n]$ be its MNF. We will define its canonical *inverse MNF*, $g[n]$, as the *particular* convolutive inverse of $f[n]$ which is zero for all negative values of n. That such a function exists and is unique can be shown as follows. Let $p[n]$

be the quasi-periodic signal with partial frequency set S, such that $p[n] = 0$ for $n \in \{-(|S| - 1) \cdots - 1\}$ and p[0] = 1. $p[n]$'s existence and uniqueness is assured by the linear independence of $|S|$ complex exponentials with distinct frequencies for any $|S|$ consecutive values of n, as mentioned above. Then define $g[n] = p[n]$ for all $n > -|S|$, and $g[n] = 0$ elsewhere.

If we evaluate the convolution of f and g, $\sum_{t=-\infty} \infty f[t]g[n-t]$, we find that $(f*g)[n] = 0$ for n ¡0, because $f[t]$ or $g[n-t]$ is zero for all t $(f*g)[n] > 0$ for n ¿ 0, because $f[t]$ is equal to the quasi-periodic $p[t]$ for all t where $g[n-t]$ is nonzero, and $(f * g)[0] = 1$, because $f[t]$ or $g[0-t]$ is zero for all nonzero t, and f[0]=g[0] = 1.

Thus, $(f * g) = \delta[n]$ and g is a convolutive inverse of f, as desired.

There are other convolutive inverses of the MNF: namely, the sums of g and any quasi-periodic signal with partial frequency set S. However, g is the only inverse of the MNF which is zero for all negative values of n. The practical importance of this property is that it allows inverse MNFs to be numerically convolved with one another, and with finite-length signals such as MNFs. If the supports of inverse MNFs were not bounded on one side, computing each value in their convolution products would require adding and multiplying an infinite number of terms.

As is the case for the MNF, the inverse MNF can be constructed by recursive convolution, using the following rules:

(1) The inverse MNF for a single signed frequency f is the signal $g[n]$ with $g[n] = e^{-i\pi fn}$ for $n \geq 0$, $g[n] = 0$ otherwise;

(2) The inverse MNF for a pair of signed frequencies $f, -f$ is the signal $g[n]$ with $g[n] = \frac{\sin 2\pi f(n+1)}{\sin 2\pi f}$ for $n \geq 0$, $g[n] = 0$ otherwise;

(3) The inverse MNF for the union of two disjoint signed frequency sets S_a and S_b is the convolution of the inverse MNFs for each frequency set.

Finally, convolution with MNFs can be used to remove frequencies from the partial frequency set of an inverse MNF, and vice versa:

(1) Given two signed frequency sets $S_a \subset S_b$, the MNF for the frequency set $S_b \setminus S_a$ is the convolution of the MNF for S_b with the inverse MNF for S_a. Given two signed frequency sets $S_a \subset S_b$, the inverse MNF for the frequency set $S_b \setminus S_a$ is the convolution of the inverse MNF for S_b with the MNF for S_a.

4.3. MNFs and inverse MNFS for the special case of uniform frequency stacks. For a symmetric, uniformly spaced stack of N signed frequencies, the terms of the MNF can be expressed as simple products. If the frequencies are spaced by F_0, then the MNF is the sequence

$$f[t] = \begin{cases} 0 & \text{for } t < 0 \\ 1 & \text{for } t = 0 \\ (-1)^t \dfrac{\prod_{x=1}^{N} \sin x\pi F_0}{\prod_{x=1}^{t} \sin x\pi F_0 \prod_{x=1}^{N-t} \sin x\pi F_0} & \text{for } t \in 1 \ldots N-1 \\ (-1)^N & \text{for } t = N \\ = & \text{for } t > 0 \end{cases} \tag{2.5}$$

In this case, the MNF is symmetric on its support, with even (odd) symmetry if N is odd (even). The reader may notice that the nonzero terms of $f[t]$ closely resemble binomial coefficients, except that factorials are replaced with products of sine terms. This is not accidental; it can be demonstrated that the MNFs for these stacks are closely related to Gaussian or q-binomial coefficients.

The terms of the inverse MNF can be expressed in a similar way. For a symmetric, uniformly spaced consisting of N signed frequencies spaced by F_0, the inverse MNF is the sequence

$$g[t] = \begin{cases} 0 & \text{for } t < 0 \\ \dfrac{\prod_{x=1}^{N-1} \sin \pi F_0 (t+x)}{\prod_{x=1}^{N-1} \sin \pi F_0 x} & \text{for } t \geq 0 \end{cases} \tag{2.6}$$

Both the MNF and inverse MNF for symmetric uniform frequency stacks can be calculated as the ratios of two cumulative products of sine terms, which is a considerably faster method of constructing them than performing repeated convolutions. This is particularly useful because effectively periodic signals with fundamental frequency F_0, and signals which are anti-periodic with respect to the same, have partial frequency sets which are symmetric uniform stacks.

4.4. The shape of the MNF for effectively periodic signals. For an effectively periodic, Nyquist-limited signal with period P, the partial frequency set is a symmetric stack of frequencies including 0, and possibly also including the Nyquist frequency (for which the sign is meaningless) if P is an even integer. The shapes of the MNFs for such signals show a cyclical pattern, as seen in Figure 1. For values of P between any two consecutive even integers, the corresponding MNF and inverse MNF are both continuous (vector-valued) functions of P. However, at even integer values of P, there is a two-sided finite discontinuity in the size of the partial frequency set, and the values of the MNF and inverse MNF change discontinuously as

well. These discontinuities will become significant when discussing pitch detection methods in the next chapter.

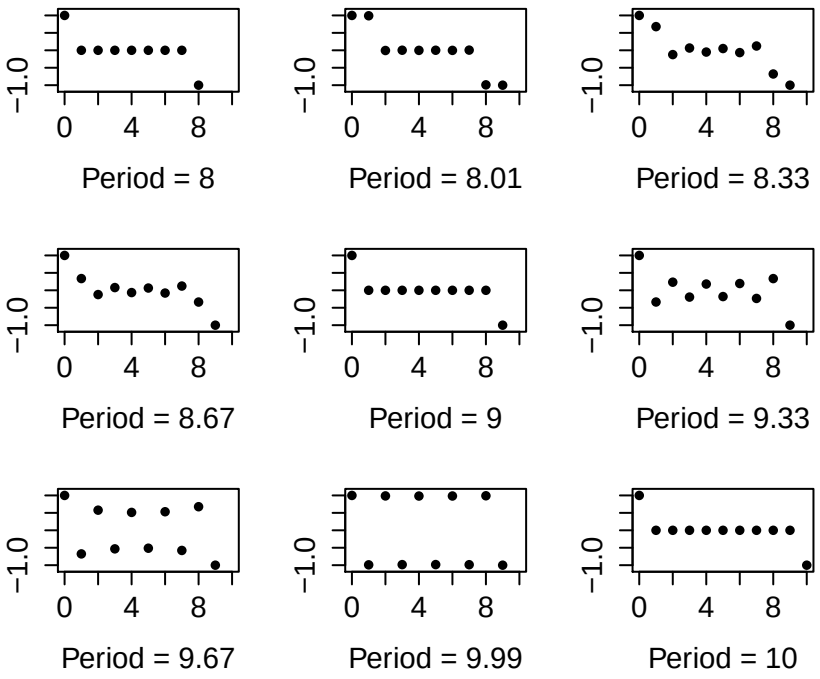


FIGURE 1. Values of the Minimal Notch Filter for the partial frequency sets of effectively periodic signals. Note the discontinuous change in the shape of the MNF as the period passes through an even integer value such as 8.

4.5. Projection via the MNF. Given the compact support of the MNF, any finite set of its translates (time-delayed copies) is linearly independent. Therefore, the set of all translates of the MNF with support inside a finite interval form a basis for the subspace of all signals on that interval which are nil-periodic with respect to the MNF's frequency set. If we can orthonormalize this basis, we can project any given signal onto the nil-periodic subspace, accomplishing the goal of this section. This can be done particularly efficiently if we employ a rectangular (i.e. unit) window function.

Suppose an MNF with support width L has K translates $\{\mathbf{x}_0, 1 \ldots \mathbf{x}_{K-1}\}$ with supports contained inside the interval $[0, 1 \ldots L+K-2]$. Let $\mathbf{X}$ be the $(L+K-2)\times(K)$ matrix of translates, with elements $\mathbf{X}_{j,k} = (\mathbf{x}_k[j]$. We wish to generate signals $\{\mathbf{q}_0, 1 \ldots \mathbf{q}_{K-1}\}$ where $\mathbf{q}_k$ is spanned by $\{\mathbf{x}_0, 1 \ldots \mathbf{x}_k\}$ and $\mathbf{q}_k$ is orthogonal to $\{\mathbf{x}_0, 1 \ldots \mathbf{x}_{k-1}\}$.

Algorithm 4.1 Pseudocode for constructing orthonormal basis for nilperiodic subspace

$\mathbf{q}_0 \leftarrow \mathbf{x}_0$
$\gamma_0 \leftarrow \|\mathbf{q}_0\|^2$
for $k = 1 \rightarrow K-1$ **do**
  for $t = 0 \rightarrow k$ **do**
    $\mathbf{a}_k[t] \leftarrow \mathbf{q}_{k-1}[t-1]$
    $\mathbf{b}_k[t] \leftarrow (-1)^L \mathbf{q}_{k-1}[k-2-t]$
    $\mathbf{q}_k[t] \leftarrow \mathbf{a}_k[t] - \frac{<\mathbf{a}_k, \mathbf{x}_0>}{\gamma_{k-1}} \mathbf{b}_k[t]$
  end for
  $\gamma_k \leftarrow \|\mathbf{q}_k\|^2$
end for

Then the set $\mathbf{N} = \{\frac{\mathbf{q}_k}{\sqrt{\|\gamma_k\|}}\}$ is the desired orthonormal basis of nilperiodic signals. This algorithm is again order L^2. If we project the original signal onto this basis and subtract the projection from the original signal, the difference is the quasi-periodic approximation:

$$\hat{\mathbf{s}} = \mathbf{s} - \mathbf{N}\mathbf{N}^*\mathbf{s} \tag{2.7}$$

5. Comparing projection methods

Above, I have described two algorithms for projecting a signal onto the locally periodic subspace of a given pitch trace, or the nil-periodic subspace of a partial frequency set. Both projection algorithms can be used to estimate a locally periodic or quasi-periodic signal from noisy data.

Each algorithm has its advantages and disadvantages. The nil-periodic projection algorithm is generally faster than the periodic projection algorithm, not only because new spanning vectors are generated through translation rather than multiplication, but also because each new spanning vector has to be orthgonalized against only one of its predecessors rather than four. The nil-periodic projection algorithm can also be applied to partial frequency sets that are not uniformly spaced, allowing it to be used for estimation of quasi-periodic or band-limited periodic signals. On the other hand, the periodic projection algorithm can be used with arbitrary window functions, and can be applied to cases where the fundamental frequency varies within the signal frame by allowing the fundamental phase vector $\boldsymbol{\phi}$ to have a nonlinear relationship with time. For this reason, the periodic projection algorithm is generally more accurate than the nil-periodic projection algorithm when used to estimate locally periodic signals with extremely rapidly varying fundamental frequencies.

References

Bohr, Harald (1925). "Zur theorie der fastperiodischen funktionen". In: *Acta Mathematica* 46.1-2, pp. 101–214.

Friedman, David H (1977). "Pseudo-Maximum-Likelihood Speech Pitch Extraction". In: *Acoustics, Speech and Signal Processing, IEEE Transactions on* 25.3, pp. 213–221. ISSN: 0096-3518. DOI: 10.1109/TASSP.1977.1162940.

Lanczos, C (1950). "An iteration method for the solution of the eigenvalue problem of linear differential and integral operators". In: *Journal of Research of the National Bureau of Standards* 45.4, p. 255. ISSN: 0091-0635. DOI: 10.6028/jres.045.026.

Noll, A Michael (1970). "Pitch Determination of Human Speech by the Harmonic Product Spectrum, the Harmonic Sum Spectrum and a Maximum Likelihood Estimate". In: *Proceedings of the Symposium on Computer Processing in Communications*. Ed. by Jerome Fox and Polytechnic Institute of Brooklyn. Microwave Research Institute, pp. 779–796.

Priemer, Roland (1991). *Introductory signal processing*. Vol. 6. World Scientific.

Wise, James D, James R Caprio, and Thomas W Parks (1976). "Maximum Likelihood Pitch Estimation". In: *IEEE Transactions on Acoustics, Speech, and Signal Processing* 24.5, pp. 418–423.

CHAPTER 3

Pitch Detection and Period Indicator Functions

1. Introduction

A common task in signal analysis is pitch detection: the estimation of the fundamental frequency of a periodic, or locally periodic, signal. Most pitch detection algorithms are based on *period indicators*, functions which provide a numerical assessment of how well a given fundamental frequency candidate fits a given signal. The choice of period indicator is a tradeoff between several criteria, including computational complexity, frequency resolution, and robustness against noise. One of the most noise-robust indicators is the pseudo-maximum likelihood estimator (PMLE)Friedman (1977); however, this indicator is rarely used, possibly because of its high computational complexity. The autocorrelation function is much more commonly used in research and industry.

In this chapter, I introduce refinements to both period indicators mentioned above. I develop a novel method for normalizing the short-term autocorrelation which improves its accuracy, and extend the PMLE to apply to hypothetical periods which are not integer multiples of the sampling period. The *approximate pseudo-maximum likelihood estimator* (APLE) approximates the PMLE closely, but is much simpler to compute. The *extended pseudo-maximum likelihood estimator* (EPLE) extends the PMLE to assess a larger set of fundamental frequencies, providing exceptional frequency resolution and allowing identification of fundamental frequencies that vary

within a frame. These indicators may be used in tandem to provide precise pitch detection of a very large class of periodic and locally periodic signals.

1.1. Definition and desirable properties of a period indicator function. A *pitch detector* is a program or numerical algorithm which estimates the fundamental frequency of a periodic signal, or the pitch trace of a locally periodic signal. Most pitch detectors rely on a *period indicator function*, a function which takes a signal frame and a period as arguments, and attains its maximum value if the signal is in fact periodic and possesses that fundamental period. Typically, the graph of a period indicator will actually show several peaks of comparable height, corresponding to multiples of the fundamental period; the pitch detector writes the locations of these peaks into a list, then chooses the shortest value from among them. There are a wide variety of extant period indicator functions, including autocorrelation, comb-filtered energy, and the maximum and pseudo-maximum likelihood estimators (Friedman 1977).

I propose that an ideal period indicator function should have the following properties:

P1. The function is bounded above and below by known values. (Without loss of generality, we may use a linear transformation to assure that these values are ± 1)

P2. The function attains its maximum value (WLOG, 1) for a given signal and period iff the signal is effectively periodic and that period is a multiple of its fundamental period.

P3. The expected value of the function for a signal consisting of independently distributed noise and any nonzero period is 0.

P4. The function is robust to broadband noise; that is, the addition of noise to the tested signal it changes the location of its peaks by relatively little.

P5. The function is defined for all periods on a continuous interval, continuous with respect to period, and ideally differentiable as well.

P6. The function is reasonably fast to compute.

2. Improving approximations to autocorrelation

One of the most commonly used period indicator functions is the autocorrelation function. The autocorrelation of a signal, for a given delay, is the correlation between that signal and a delayed copy of itself. If a signal is nearly periodic, it will be highly correlated with any copy delayed by a multiple of its fundamental period. Normalized by signal energy, the *long-term* autocorrelation of an infinite signal on a continuous domain satisfies properties P1, P2, and P3, and partially satisfies P4. If computed using a pair of Fast Fourier Transforms (FFTs) for each frame, it satisfies P6 as well. (It cannot satisfy P5, as the autocorrelation is only defined for delays that are integer multiples of the sampling period.)

In practice, only *short-term* approximations of the autocorrelation can be used, since real-world signals are defined on a discrete domain and for a finite span of time. The simplest method of performing a short-term autocorrelation is to window the signal, treat it as an infinitely long signal with finite support, and take the long-term autocorrelation. Unfortunately, this form of the short-term autocorrelation will not show peaks located precisely at multiples of a signal's period, because the signal and its delayed copy have different supports. It must therefore be normalized to provide a decent approximation. One normalization method, found for instance in MATLAB's xcorr function, divides the short-term autocorrelation by the

number of samples on which the truncated signal's support overlaps with the support of its delayed copy (MathWorks Documentation 2015):

$$R(\tau) = \frac{\sum_{t=\max(0,\tau)}^{N-1+\min(0,\tau)} \overline{s}_t s_{t+\tau}}{N - |\tau|}$$

A generalization of this method, allowing for the use of an arbitrary window function w_t, was proposed by Boersma (1993). In the Boersma method, the autocorrelation of the windowed signals is normalized by the autocorrelation of the window function itself, and that ratio is again normalized by its own value for $\tau = 0$:

$$R_B(\tau) = \frac{\left(\sum_{t=-\infty}^{\infty} [w_t \overline{s}_t]\,[w_{t+\tau} s_{t+\tau}]\right)\left(\sum_{t=-\infty}^{\infty} w_t^2\right)}{\left(\sum_{t=-\infty}^{\infty} w_t^2 |s_t|^2\right)\left(\sum_{t=-\infty}^{\infty} w_t w_{t+\tau}\right)} \tag{3.1}$$

If w is a rectangular window function with unit height, this formula reduces to that used in MATLAB's xcorr function.

When using any of the windows typically employed in signal analysis, such as Gaussian, Kaiser, Hamming, etc., the short-term autocorrelation with Boersma normalization is a much more accurate estimate of the true (long-term) autocorrelation than is the xcorr function or the non-normalized short term autocorrelation. Furthermore, Boersma normalization requires only two additional FFTs for the autocorrelation of the window frame, no matter how many frames are analyze, so it satisfies P5. However, the Boersma-normalized autocorrelation function still does not satisfy properties P1, P2 or P4 above. It may slightly exceed the bounds of [-1, 1]; it is not guaranteed to attain the value of 1 only for multiples of the signal's fundamental period; and it is only defined for integer periods.

I propose an alternate method of normalizing the short-term autocorrelation, namely:

$$R_M(\tau) = \frac{2\sum_{t=-\infty}^{\infty} [w_t s_t]\,[w_{t+\tau}\overline{s}_{t+\tau}]}{\sum_{t=-\infty}^{\infty} w_{t+\tau} w_t |s_t|^2 + \sum_{t=-\infty}^{\infty} w_{t-\tau} w_t |s_t|^2} \tag{3.2}$$

This is equivalent to taking the inner product of the sequences $a = \sqrt{w_t w_{t+\tau}}s_t$ and $b = \sqrt{w_t w_{t+\tau}}s_{t+\tau}$, and dividing by $\frac{\|a\|^2 + \|b\|^2}{/}2$. Since $< a, b > \leq \|a\|\|b\|$ by the Cauchy-Schwarz inequality, and $\|a\|\|b\| \leq \frac{\|a\|^2 + \|b\|^2}{/}2$ by the non-negativity of $\|a+b\|^2$, $R_M(\tau)$ is always bounded by ± 1. Moreover,

$R_M(\tau) = 1 \iff a = b \iff s_t = st + \tau \forall t\colon w(t)w(t+\tau) \neq 0$

$\iff$ s is periodic with period τ on the interval where the supports of $w(t)$ and $w(t+\tau)$.

Thus, $R_M(\tau)$ shares properties P1 and P2 with the true (long-term) autocorrelation, while other normalization methods, such as that of Boersma 1993, do not.

Moreover, as proved in Appendix A, the error bounds for this normalization method are particularly attractive. If a signal s is periodic with period τ, the error of the short-term approximation to the autocorrelation for delay n is bounded by

$|E_M(n)| \leq \frac{\left(\|\vec{s}\|^4 - a_s(n)^2\right)}{\|\vec{s}\|^4} \frac{(p-1)c_n}{a_w(n) - c_n}$

where $c_n = w[\lceil n/2 \rceil] w[\lfloor n/2 \rfloor]$.

A few things are immediately apparent from this error bound. First, the error varies with $\frac{\left(\|\vec{s}\|^4 - a_s(n)^2\right)}{\|\vec{s}\|^4}$. Thus, the method is most accurate when the (unnormalized) autocorrelation magnitude approaches the signal energy.

This occurs for any signal when the delay is close to an integer multiple of p, so the method should return autocorrelation peaks particularly accurately. It also tends to be the case for relatively narrow-band signals, since the uncertainty principle implies a wide dispersion of the autocorrelation function. Conversely, the method will be poorest for a delta-like impulse train, for which the autocorrelation is near-zero for almost all delays.

Second, the error will decrease as $\frac{a_w(n)}{c_n}$ increases. The error will therefore be smaller for small n, (since c_n is a single term in the sum making up $a_w(n)$, and the number of nonzero terms in that sum increases with decreasing n) and for slowly-decaying windows with wide support.

Short-term autocorrelation values for no normalization, $R_B(\tau)$, and $R_M(\tau)$ are graphed for various signals of fundamental period 10π in Figure reffigure:autoexamp. Note that these autocorrelations were calculated using a frame spanning two cycles of the period, so their values become unreliable beyond delays of 10π because they are no longer based on a complete cycle of the signal. It can be seen that $R_B(\tau)$ and $R_M(\tau)$ are generally more accurate than unnormalized autocorrelation until the 10π reliability cutoff, and that the two normalization methods yield quite similar values for waveforms that are not strongly peaked in time, such as sine and square waves. However, for strongly peaked waveforms such as sawtooth waves and impulse trains, $R_B(\tau)$ can severely overshoot or undershoot the true (long-term) value of the autocorrelation for delays near the signal's true period, whereas $R_M(\tau)$ does not.

The orders of computational complexity of $R_M(\tau)$ and $R_B(\tau)$ are equal; however, $R_M(\tau)$ requires roughly twice as many operations does $R_B(\tau)$,

because the denominator in Equation 3.2 includes a cross-correlation that requires two additional FFTs per frame.

2.1. Rapid interpolation for non-integer periods. All autocorrelation techniques discussed so far are based on pairwise comparison of sampled values of the signal, so they are limited to testing delays that are integer multiples of the sampling period. Autocorrelation values for non-integer delays must therefore be interpolated. There exist myriad methods for interpolating between uniformly-spaced values; ideally, we would like to generate an interpolant which not only approximates the true (long-term) autocorrelation with little error, but also satisfies as many of properties P1-6 as possible.

Properties P3, P5 and P6 are relatively easy to satisfy via common interpolation methods such as linear, sinc, or spline interpolation. Properties P1-2 are more challenging. Even if the short-term autocorrelation is normalized such that its values are bounded by ± 1, sinc and spline interpolants may exceed those bounds. On the other hand, interpolation using non-negative kernels such as a triangle function (which results in linear interpolation) or a sinc-squared function will yield an interpolant which is bounded by the maximum and minimum of the interpolated values. This guarantees boundedness by ± 1, but also prevents the interpolant from approaching 1 more closely than any of the interpolated values do, which implies that autocorrelation peaks will be poorly reconstructed by the interpolant unless at least one such peak is centered on an integer delay value.

How can we produce an autocorrelation interpolant which is properly bounded, yet can still accommodate peaks occurring between integer delay values?

One method is to create smooth interpolations of both the original (discrete) signal and the window function, for instance by spline or sinc interpolation. We then cross-correlate the windowed signal with the product of the delayed, interpolated signal and window functions, and normalize in the fashion already outlined.

Let s and w be the original signal and window function; let $\tilde{s}(t)$ and $\tilde{w}(t)$ be their interpolants. Then, for non-integer τ, the autocorrelation interpolant will be

$$\tilde{R}_M(\tau) = \frac{2\int_{t=-\infty}^{\infty} [w_t s_t]\left[\tilde{w}(t+\tau)\overline{\tilde{s}}(t+\tau)\right]}{\sum_{t=-\infty}^{\infty} w_t\tilde{w}(t+\tau)|s_t|^2 + \sum_{t=-\infty}^{\infty} w_t\tilde{w}(t+\tau)|\tilde{s}(t+\tau)|^2} \tag{3.3}$$

It is easily seen that $\tilde{R}_M(\tau) = R_M(\tau)$ for integer τ. Moreover, by the same argument used to bound $R_M(\tau)$, $\tilde{R}_M(\tau)$ is bounded by ± 1. Unfortunately, it is *not* the case that $\tilde{R}_M(\tau) = 1$ iff s is essentially periodic with period τ. This is because the interpolated signal $\tilde{s}(t)$ will not in general share the periodicity of s. However, $\tilde{R}_M(\tau)$ is still a "good enough" interpolant for most purposes, particularly because it can be rapidly calculated with FFTs for any uniformly spaced grid of delay values. Thus, it satisfies P1 and P3-P6. In Figure 2.1, $\tilde{R}_M(\tau)$ (red line) is compared to the true long-term autocorrelation (black line); until the 1-cycle reliability cutoff, the interpolant closely matches the true value for all four signal waveforms being tested.

3. Improving the Pseudo-Maximum Likelihood Estimator

3.1. Definition of the PMLE. In Chapter 2, I discussed least-squares projection of a signal to the subspace of signals with a given effective period. If that period is an integer, this projection can be accomplished by means of

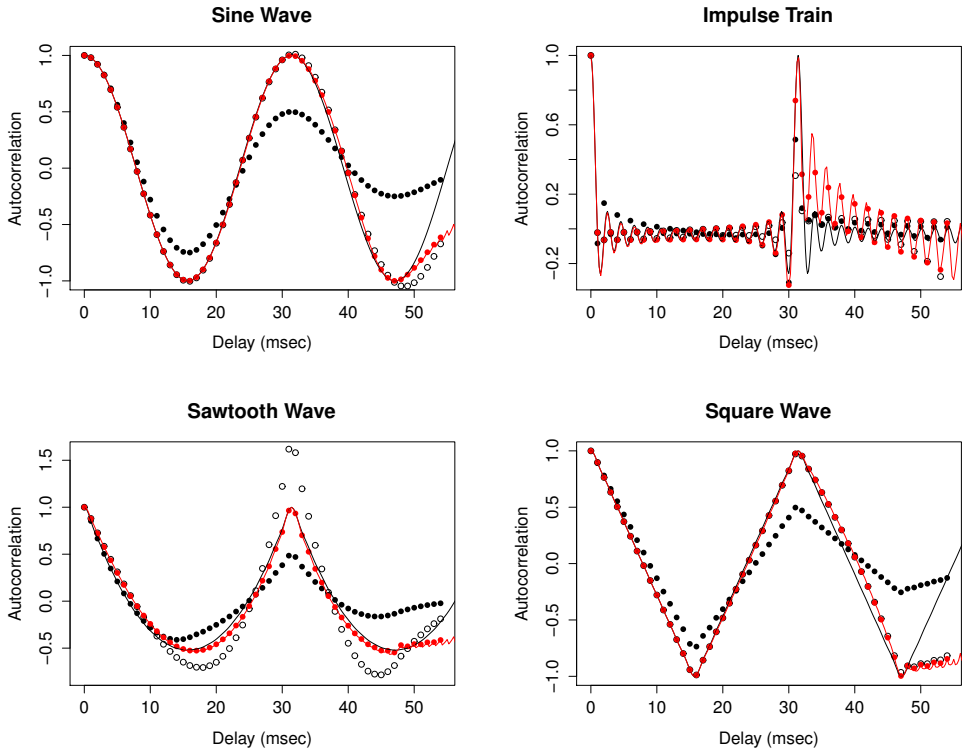


FIGURE 1. Comparisons of the true (long-term) autocorrelation with various short-term approximations, for four different waveforms with fundamental period 10π. Signal frames are two cycles long. Black line: true autocorrelation. Black circles: unnormalized short-term autocorrelation. Open circles: R_B, short-term autocorrelation normalized by the Boersma method. Red circles: R_M, short-term autocorrelation normalized by the method introduced in this chapter. Red line: rapidly computable interpolant of R_M.

a weighted periodic average of signal samples. The ratio of the energies (or weighted mean squared amplitudes) of the periodic average and the original signal can be used as a period indicator function, termed the Maximum Likelihood Estimator Noll (1970) . Because the periodic average is a linear projection, its energy cannot exceed that of the original signal, and they have equal energy only if the original signal and the periodic average are identical. Also, the periodic average is the optimal approximation (in a

least-squares sense) of a periodic signal when the original is contaminated by independently-distributed zero-mean noise. The Maximum Likelihood Estimator (MLE) function therefore satisfies desired properties P1, P2 and P4.

However, it does not satisfy property P3. Not only is the expected value of the MLE nonzero for a broadband noise input, but it varies roughly linearly with the test period. This is because, for longer test periods, the signal frame contains fewer pairs of values which are separated by a multiple of that period. Thus, the number of signal elements averaged together to produce each term in the periodic average is smaller, and the expected energy of the periodic average is larger. Because of this, the MLE is biased toward favoring larger test periods, and the amount of bias is dependent on the amount of noise in the test signal.

Wise, Caprio, and T. W. Parks (1976) proposed removing this bias in the MLE by normalizing it with a denominator that was also a linear function of period. A more sophisticated correction, the Pseudo-Maximum Likelihood Estimator (PMLE), was proposed by Friedman (1977), and relies on subtracting those terms from the numerator and denominator of the MLE which must be nonnegative for any test signal. The resulting formula, for a signal s, a window function w and a test period p, is:

(3.4)

$$PMLE(s,w,p) = \frac{||s_p||^2_{w_p} - ||s||^2_{\frac{w^2}{w_p}}}{||s||^2_w - ||s||^2_{\frac{w^2}{w_p}}} = \frac{\sum_{t=-\infty}^{\infty} \frac{\sum_{\delta \neq 0} w(t)s(t)w(t+\delta p)s(t+\delta p)}{w_p(t)}}{\sum_{t=-\infty}^{\infty} \frac{\sum_{\delta \neq 0} w(t)s^2(t)w(t+\delta p)}{w_p(t)}}$$

where the subscript on the squared norm refers to the weight function used to calculate it. Another way to see this formula is as the ratio between the energies of the periodic average and original signal, after each energy has been "noise-corrected" by subtracting the expected value of the energy of a noise frame with the amplitude profile of the original signal. The PMLE satisfies P3 as well as P1, P2 and P4, and shows no bias toward longer or shorter periods in the presence of noise.

The PMLE was partially rediscovered in Sethares and Staley (1999), but has seen little mathematical analysis since that time, nor experimental use beyond a few studies such as D. D. Muresan and T. W. Parks (1999) and D. Muresan and T. Parks (2003). Perhaps this is because it is substantially slower than autocorrelation, requiring $O(L^2)$ operations to calculate for a signal frame of length L.

3.2. Extending the definition of the PMLE to non-integer and non-constant periods. The formula given above for the PMLE only applies to constant, integer test periods, because periodic averaging is not possible otherwise. Here, I show how it can be extended to a wider class of test periods. We simply need to express Equation 3.4 in a manner such that $||s_p||^2_{w_p}$, $||s||^2_w$, and $||s||^2_{\frac{w^2}{w_p}}$ are well defined for non-integer and non-constant periods. This is easily done, using the notation from Equations 2.2 and2.3. $||s||^2_w$ becomes $\|\tilde{\mathbf{s}}\|^2$, $||s_p||^2_{w_p}$ becomes $\|\tilde{\hat{\mathbf{s}}}\|^2 = \tilde{\mathbf{s}}^* \mathbf{A}\mathbf{A}^+ \tilde{\mathbf{s}}$, and $||s||^2_{\frac{w^2}{w_p}}$ becomes $\tilde{\mathbf{s}}^* \mathrm{Diag}(\mathbf{A}\mathbf{A}^+)\tilde{\mathbf{s}}$, where $\mathrm{Diag}(\mathbf{A}\mathbf{A}^+)$ is the matrix with main diagonal entries equal to those of $\mathbf{A}\mathbf{A}^+$, and all other entries zero. With this formulation, the PMLE can be applied to any test period for which the matrices $\mathbf{\Phi}$ and $\mathbf{A}$ from Equations 2.2 and2.3 can be defined. Computing these matrices can be done via the Lanczos-like algorithms already described. The value

of $||s||^2_{\frac{w^2}{w_p}}$ then remains equal to the expected value of $\|\tilde{\hat{\mathbf{s}}}\|^2$ if the signal consists of independently-distributed zero-mean noise, so the PMLE continues to satisfy P3.

4. Averaged autocorrelation

4.1. Disadvantages of autocorrelation and PMLE. As period indicator functions, autocorrelation and the PMLE embody a common trade-off. The PMLE is more robust against noise, but is very slow and is not a continuous function of period. Autocorrelation is fast, and (when interpolated) a continuous and smooth function of test period, but is less robust against noise and has broader peaks

It is therefore desirable to find a period indicator which approximates the PMLE, but can be calculated as quickly as autocorrelation and interpolated smoothly. The indicator I introduce here is *averaged autocorrelation*: the weighted average of all autocorrelation values for delays that are a multiple of the test period, with weights equal to the denominators for the normalization method introduced in Section 2 of this chapter:

$$AR_M(\tau) = \frac{2\sum_{t=-\infty}^{\infty}\sum_{\delta\neq 0}\left[w_t s_t\right]\left[w_{t+\delta\tau}\overline{s}_{t+\delta\tau}\right]}{\sum_{t=-\infty}^{\infty}\sum_{\delta\neq 0} w_{t+k\tau}w_t|s_t|^2} \tag{3.5}$$

This formula is very similar to Equation **??**, with the exception that the terms of the sums in the numerator and denominator of the latter are divided by $w_p(t)$, the periodic average of the window function. As the window become broader and smoother, $w_p(t)$ becomes nearly constant, and the averaged autocorrelation value converges to the PMLE.

In Figure 4.1, averaged autocorrelation is compared to short-term autocorrelation for a noisy sawtooth signal of period 10π samples, and a frame length of approximately five periods. It can be seen that the averaged autocorrelation not only has less jitter from noise, but also has narrower peaks. This is because for any delay value τ which is near but not precisely at an autocorrelation peak, its multiples $\delta\tau$ will drift farther and farther from the peaks as δ increases. The average of the autocorrelation values at multiples of τ will therefore be smaller than the autocorrelation value at τ. By the same argument, unlike the ordinary autocorrelation, the averaged autocorrelation does not have a central peak at $\tau = 0$ and in fact vanishes there if the test signal has no DC offset, because the average value of the entire autocorrelation function is zero. This is a useful property when working with noisy signals, since the ordinary autocorrelation's central peak may house local maxima that are mistakenly identified as peaks in their own right, leading to spuriously small period estimates for the test signal.

The order of computational complexity of averaged autocorrelation is the same as for ordinary autocorrelation, namely, $O(L \log L)$ for a signal frame of length L. Averaged autocorrelation requires only one additional round of $O(L \log L)$ additions for each of the numerator and denominator in Equation 3.5.

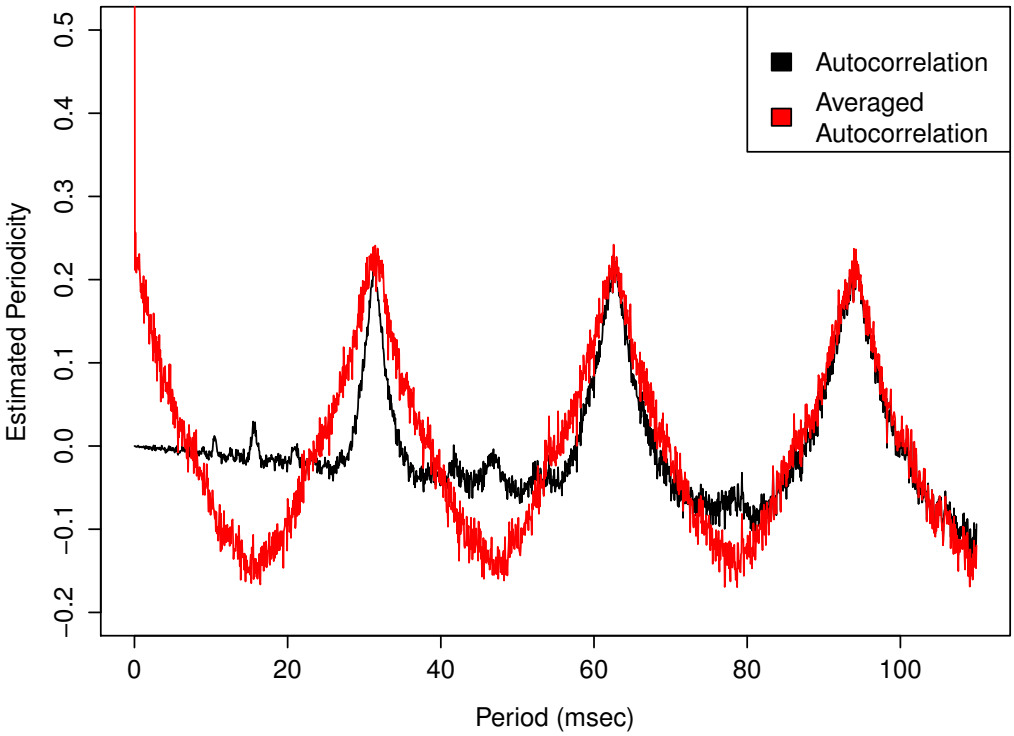


FIGURE 2. Comparison of the short-term autocorrelation (normalized with the method introduced in this chapter) and the averaged autocorrelation. The test signal is a sawtooth wave contaminated with white noise, in a 1:4 signal to noise ratio. The true peak periodicity value is 0.2.

References

Boersma, Paul (1993). "Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound." In: *Proceedings of the Institute of Phonetic Sciences* 17, pp. 97–110. DOI: 10.1371/journal.pone.0069107. URL: http://www.cs.northwestern.edu/%7B~%7Dpardo/courses/eecs352/papers/boersma-pitchtracking.pdf.

Friedman, David H (1977). "Pseudo-Maximum-Likelihood Speech Pitch Extraction". In: *Acoustics, Speech and Signal Processing, IEEE Transactions on* 25.3, pp. 213–221. ISSN: 0096-3518. DOI: 10.1109/TASSP.1977.1162940.

MathWorks Documentation (2015). *Cross-correlation - MATLAB xcorr*. URL: http://www.mathworks.com/help/signal/ref/xcorr.html (visited on 10/20/2015).

Muresan, D Darian and Thomas W Parks (1999). "Orthogonal Subspace Decomposition of Periodic Signals". In: *Conference Record of the Thirty-Third Asilomar Conference on Signals, Systems, and Computers*.

Muresan, D.D. and T.W. Parks (2003). "Orthogonal, exactly periodic subspace decomposition". In: *IEEE Transactions on Signal Processing* 51.9, pp. 2270–2279. ISSN: 1053-587X. DOI: 10.1109/TSP.2003.815381. URL: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=1223539.

Noll, A Michael (1970). "Pitch Determination of Human Speech by the Harmonic Product Spectrum, the Harmonic Sum Spectrum and a Maximum Likelihood Estimate". In: *Proceedings of the Symposium on Computer Processing in Communications*. Ed. by Jerome Fox and Polytechnic Institute of Brooklyn. Microwave Research Institute, pp. 779–796.

Sethares, William A. and Thomas W. Staley (1999). "Periodicity transforms". In: *IEEE Transactions on Signal Processing* 47.11, pp. 2953–2964. ISSN: 1053587X. DOI: 10.1109/78.796431.

Wise, James D, James R Caprio, and Thomas W Parks (1976). "Maximum Likelihood Pitch Estimation". In: *IEEE Transactions on Acoustics, Speech, and Signal Processing* 24.5, pp. 418–423.

CHAPTER 4

Acoustic Profiling in a Complexly Social Species, the American Crow [Mates et al., 2014]

RESEARCH ARTICLE

Acoustic profiling in a complexly social species, the American crow: caws encode information on caller sex, identity, and behavioral context[1]

Exu Anton Mates[2]
Department of Psychology, University of Washington, Seattle, WA, USA
Mailing Address: Box 351525, Seattle, WA 98195
Phone: 510-295-3686 Fax: 206.685.3157 Email: xamates@uw.edu

Robin R. Tarter
Department of Evolution, Ecology and Organismal Biology, Ohio State University, Columbus, OH
Mailing Address: 5600 University Way NE, Apt. 9, Seattle, WA 98105
Phone: 614-499-6781 Email: cbrachy@gmail.com

James C. Ha
Department of Psychology, University of Washington, Seattle, WA, USA
Mailing Address: Box 351525, Seattle, WA 98195
Phone: 206-543-2420, 206-543-7494 Fax: 206-685-3157
Email:jcha@u.washington.edu

Anne B. Clark
Department of Biological Sciences, Binghamton University, Binghamton, NY, USA
Mailing Address: PO Box 6000, Binghamton, NY 13902
Phone: 607-777-6228 Fax: (607) 777-6521 Email: aclark@binghamton.edu

Kevin J. McGowan
Laboratory of Ornithology, Cornell University, Ithaca, NY, USA
Mailing Address: 159 Sapsucker Woods Rd. Ithaca, NY 14850
Phone: 607-254-2452 Email: kjm2@cornell.edu

Acoustic profiling in a complexly social species, the American crow: caws encode information on caller sex, identity, and behavioral context [3]

Abstract: Previous research on interindividual variation in the calls of corvids has largely been restricted to single call types, such as alarm or contact calls, and has rarely considered the effects of age on call structure. This study explores structural variation in a contextually diverse set of "caw" calls of the American crow (*Corvus brachyrhynchos*), including alarm, foraging recruitment and territorial calls, and searches for structural features that may be associated with behavioral context and caller sex, age, and identity. Automated pitch detection algorithms are used to generate 23 pitch-related and spectral parameters for a collection of caws from 18 wild, marked crows. Using principal component analysis and mixed models, we identify independent axes of acoustic variation associated with behavioral context and with caller sex, respectively. We also have moderate success predicting caller sex and identity from call structure. However, we do not find significant acoustic variation with respect to caller age.

Keywords: acoustic feature analysis; call classification; caller identity; sexual dimorphism; vocal ontogeny; PACS Code 43.80.Ka

I. Introduction

In species with complex social structures, it is often adaptive to exchange information on identity, sex, age and other attributes with conspecifics. Accurate signaling and recognition of identity may reduce the risk of inbreeding, decrease the intensity and frequency of aggressive interactions with conspecifics, and facilitate reciprocal altruism and kin selection (Tibbetts and Dale 2007). Not only contact calls and territorial advertisements, but even alarm signals may be more useful if they are individually distinctive, as they enable the receiver to selectively attend to signals produced by more reliable or informative senders (Blumstein et al. 2004, Pollard 2011). On the other hand, individual distinctiveness can be costly to the sender, as it makes deception more difficult, and allows receivers and eavesdroppers to consistently target the sender for harmful behaviour, e.g. predation attempts, social punishment, or rejection as a potential mate (Dale et al. 2001, Tibbetts and Dale 2007).

The corvids - crows, magpies, jays and their allies - are highly social songbirds, living in groups comprising individuals of different ages and sexes. Many species engage in cooperative breeding, foraging, and defense behaviors, and consistent relationships between individuals can last for many years (Kilham 1989; Ekman and Ericson 2006). Corvids are extremely vocal, employing calls in disputes over food and territory, collective defense against predators, and recruitment of conspecifics to food resources (Chamberlain and Cornwell 1971; Baeyens 1981; Heinrich 1988; Ha et al. 2003). Corvid species are therefore good candidates for the ability to recognize other individuals and small groups of conspecifics by their vocalizations, and indeed there is evidence of this ability in several species. Mexican jays (*Aphelocoma ultramarina*) can

discriminate the "primary calls" of in-group birds from those of birds outside of their family groups (Hopp, Jablonski, and Brown 2001). Adult pinyon jays (*Gymnorhinus cyanocephalus*) recognize the calls of mates and of related nestlings, and nestlings preferentially beg in response to the calls of related adults (McArthur 1982). Juvenile rooks (*Corvus frugilegus*) can discriminate the calls of siblings from those of similarly aged non-siblings (Røskaft and Espmark 1984). Jungle crows (*Corvus macrorhynchos*) can be trained to show discrimination between the contact calls of familiar conspecifics (Kondo et al. 2010), while captive carrion crows (*Corvus corone corone*) spontaneously discriminate between the calls of familiar jackdaws (*Corvus monedula*) (Wascher et al. 2012). Jungle crows have even demonstrated the ability to recognize group members cross-modally, matching calls to visual representations (Kondo et al. 2012).

Although recognition of individuals and groups has thus been demonstrated, it is less clear what acoustic cues corvids rely on to perform such discrimination tasks. The studies that have explored this question have generally done so from a signal analysis perspective, by analyzing the calls themselves and searching for properties that correlate significantly with caller identity. These studies, however, have for the most part confined themselves to a single call type, such as the "inflected alarm caw" of the American crow (*Corvus brachyrhynchos*), the "krah" call of the Hooded crow (*Corvus corone cornix*), the "ka" contact call of the jungle crow (*Corvus macrorhynchos*), and the food call of the common raven (*Corvus corax*) (Brown 1985; Allenbacher et al. 1995; Kondo et al. 2010; and Boeckle et al. 2012, respectively). This leaves open the question of whether corvids possess consistent signatures of individual identity across their entire repertoire of species-typical vocalizations. Such consistent signatures are not always found in other taxa, even if particular subsets of their calls are individually distinct. For instance,

both alarm barks and contact barks in chacma baboons (*Papio ursinus*) show individually distinctive characteristics, but the pattern of individual variation is not consistent between bark types (Fischer et al. 2001). On the other hand, rhesus macaques (*Macaca mulatta*) and South Polar skuas (*Catharacta maccormicki*) each show individual characteristics in some call types and not others, but across all individually distinctive call types, the same parameters are relevant for individual identification (Rendall, Owren, and Rodman 1998; Charrier et al. 2001).

Moreover, only a few studies of corvids have looked for acoustic correlates of demographic attributes above the individual level, such as sex and age (Laiolo, Palestrini, and Rolando 2000; Yorzinski et al. 2006). The American crow is a promising subject for research in this area, due to its complex social life and vocal behaviour (reviewed in Verbeek and Caffrey 2002). American crows are sexually monomorphic in appearance and most behavior, although adult males are somewhat heavier on average than adult females. Crows are facultative cooperative breeders, with 0-10 nonbreeding auxiliaries residing on the territory of the breeding pair. Many of these auxiliaries are offspring from previous years, who may remain on their natal territory for up to six years before breeding themselves. Although both sexes of American crows reach physiological maturity by the end of their second year, most females are at least three years old before they establish independent breeding territories and take social mates, and most males are at least five (Verbeek and Caffrey 2002; Robinson, Jr. 2009).

Though not celebrated for their song, American crows produce a wide variety of vocalizations. One of the most common calls is the caw, produced in contexts ranging from

territorial defense to food recruitment to predator mobbing (Tarter 2008), and general mild alarm (Yorzinski et al. 2006). Caws are harmonically rich and locally tonal, varying from a few hundred milliseconds to several seconds in length, and contain rapid variations in pitch and amplitude, which produce a "harsh" quality to the human ear (Laiolo and Rolando 2003). They are acoustically distinct from the rhythmic and largely atonal rattles, constant-frequency coos, and lower-pitched begging calls, which are also produced by American crows (Kilham 1989; Tarter 2008).

Caws vary considerably in pitch, duration, cadence and timbre, and observers have long sought to decode their behavioral significance. This task has proven quite challenging, due to the diversity of caws, the continuous distribution of their acoustic properties, and the difficulty of verifying age, sex and individual identity in the field for this gregarious, monomorphic species. The earliest attempts at comprehensive caw classification were based primarily on observations of large groups of unmarked birds, and defined most caws by their apparent collective function, e.g., assembly, dispersal, scolding and alert calls (Frings et al. 1958; Chamberlain and Cornwell 1971; Richards and Thompson 1978). Later researchers, such as Brown (1985), Parr (1997), and Tarter (2008), constructed classification systems that were based more purely on the acoustic properties of the caws. Parr in particular identifies approximately ten structural types of caw (short, medium and long; harsh, “ko”s, and “koaw”s; 2-syllable, doubled short, long-medium and medium-short). Parr draws many correspondences between her categories and those of earlier studies; for instance, her “ko” call matches Brown's “inflected alarm caw” and Chamberlain and Cornwell's "simple scolding call." Nevertheless, most categories in one author's classification

system overlap with several in another, and it is not yet entirely clear how these systems will be reconciled.

No sex- or age-specific call types have been confirmed among birds past their first year, and Parr (1997) found that both sexes produced all types of "territorial" caws. However, the frequency of usage of certain call types may vary with sex and breeding status (Tarter 2008). Sexual dimorphism within caw types was reported by Davis (1958), although this was based on field observations without reliable verification of caller sex. The strongest evidence for individual distinctiveness and sex differences in caw structure has been produced by Yorzinski et al. (2006). Restricting their analysis to the "inflected alarm caw" (Brown 1985), Yorzinski et al. found that the caws of females tended to have a higher pitch, shorter duration, higher frequency, greater bandwidth, and more peaked pitch contours than those of the males, and linear classifiers were able to sort calls by sex and caller identity with rates well above chance.

In this study, we sought to extend the findings of Yorzinski et al. by searching for signatures of identity, sex, age, and behavioral context within a diverse set of caws that was not pre-sorted by type. The caws were recorded from a banded population of wild yearling and adult American crows in Ithaca, NY, which was also studied by Yorzinski et al. (2006), Tarter (2008) and Yorzinski & Vehrencamp (2009). Because caw pitch oscillates rapidly, caws are difficult to represent precisely using conventional Fourier analysis, which suffers from an inherent tradeoff between time and frequency resolution (Cohen 1989). We therefore developed a pair of novel pitch detection algorithms with superior frequency resolution over short signal frames, and used

them to represent the calls as locally periodic signals, similar to those generated by source-filter models of speech production. We then parameterized the calls according to fundamental frequency, amplitude and spectral properties. Finally, we explored the resulting distributions of call properties with respect to caller identity, sex, age and behavioral context. Specific hypotheses addressed were:

•*Individual Identity:* The acoustic parameters of individual calls vary significantly with caller identity. An automated classifier can use these parameters to predict the identity of callers with accuracy above chance.

•*Sex and Age:* The acoustic parameters of individual calls vary significantly with caller sex and age. Automated classifiers can use these parameters to predict the sex and age of callers with accuracy above chance. Interactions may exist, such that older birds show higher degrees of sexual dimorphism on acoustic parameters.

•*Behavioral context:* The acoustic parameters of individual calls vary significantly with the observed behavioral context in which the call is produced (e.g. alarm/mobbing, territorial displays or recruitment to a food source). This variation is independent of that associated with identity/sex/age, allowing calls to provide simultaneous information about multiple caller attributes.

II. Materials & Methods

Recording

Calls were recorded in Cayuga Heights, NY, June-August 2006, from wild crows marked as part of a long-term study of crows in the Ithaca (Tompkins Co.), NY area by the Ithaca Crow Research Group. The Ithaca Crow Research Group, made up of researchers from Binghamton and Cornell Universities, has marked nestlings and adult birds with leg bands and patagial tags annually since 1989. The calls analyzed for the current study were recorded from a vehicle parked on known family territories. Territories of the birds in this study were all located in fairly quiet residential-suburban neighborhoods or adjacent open managed habitats such as golf courses or cemeteries (McGowan 2001). If the resident family could not be located before the start of recording, peanuts were scattered to attract them, but there was no other interaction with the birds. Recordings were made in WAV format, at a 48 kHz sampling rate, using a Marantz PMD-670 solid-state recorder and a Sennheiser ME-67 directional microphone with wind shield. Call bouts were extracted in Amadeus II v3.8.7 (Hairersoft, 2007); all further processing was done in MATLAB R2009B and R2010B, using the Signal Processing and Statistics toolboxes, and in R 3.0.1, using the nlme and AICcmodavg package (Pinheiro et al. 2013, Mazerolle 2013).

Call Selection

1674 calls were selected for a high (>93%) signal to noise ratio, a visible caller identifiable by its leg bands and patagial tags, and a lack of overlap with other birds' vocalizations. All calls used were tonal, harmonically rich "caws," and were produced by 18 birds, 11 male and 7 female (Table 1). Birth years were known, and age from fledging was estimated using a fledging date of June 1. The typical fledging period for this population is late May through June, so ages are

expected to be accurate to within approximately 30 days. Ages were 396-4833 days, with median 816; no birds younger than 1 year were included. We included 7-337 calls per bird; only three birds had <=15 calls each. A subset of 305 calls had behavioral contexts identified by observation. Contextual categories were: 1) food recruitment, 2) territorial "counter-cawing" (defined in Parr 1997), 3) beg rebuffs and 4) alarm calls. (Categories are further described in Table 4.) For the remaining calls, context was unclear or ambiguous in some way, including the focus of the caller's attention being unknown, or the caller being accompanied by unidentified conspecifics.

Representation of Calls as Locally Periodic Signals

Calls were represented as locally periodic signals with smoothly varying fundamental frequency (pitch). They were divided into overlapping frames, and within each frame they were represented as a sum of harmonics of a single fundamental frequency. The fundamental frequency of each frame was determined using an extension of Friedman's Pseudo-Maximum Likelihood Estimator (PMLE), originally developed to estimate the pitch of human speech (Friedman 1977). We have developed two period indicator functions from the PMLE, which will be summarized here (E. A. Mates and J. C. Ha, unpublished data). The *approximate pseudo-maximum likelihood estimator* (APLE) approximates the PMLE with a weighted sum of autocorrelations that can be computed more efficiently. The *extended pseudo-maximum likelihood estimator* (EPLE) calculates the PMLE precisely, but allows for non-integer-valued periods and periods that vary over the span of a frame, and for arbitrarily chosen bandwidths. Additional details on these algorithms can be found in the supplementary materials.

To illustrate the resultant locally periodic representation, an example is displayed alongside a traditional spectrogram of the same call (Figure 1). For the calls used in this set, locally periodic approximations captured an average of 92.90% of each filtered frame's energy. Measured fundamental frequency did indeed vary smoothly, with an average RMS fundamental frequency shift of 12.67 Hz between frames.

Parameterization of calls

(Additional details on parameterization can be found in the supplementary materials.)

Each clip of a single call was automatically partitioned into a "voiced" section preceded and followed by "silent" sections, based on the estimated call energy in each frame. Only the voiced section was used for further analysis.

Twenty-three parameters were extracted from each call (Table 2): 11 based on the pitch trace; six based on the power envelope; and six based on the spectral properties. Among the pitch trace parameters, the mean, 5th percentile and 95th percentile pitch were estimated from the pitch distribution. A cubic polynomial was fit to the pitch trace, in order to represent its gradual variation over the course of the call. The cubic curve was then subtracted from the pitch trace to leave a residual. The RMS magnitude of the residual was recorded as the "pitch instability." The residual was expected to contain a "wobble" or high-frequency modulation in pitch, which is typically found in "caw"-type calls across the genus *Corvus* (Laiolo and Rolando

2003). This wobble's fundamental frequency was estimated, and two alternative estimates of its magnitude were also computed.

Among the signal power parameters, the duration of the voiced section of each clip was used as a measure of call length. The central moments of call energy with respect to time were also computed. As was the case for the pitch trace, there was expected to be a rapid "wobble" in signal power; its fundamental frequency and magnitude were estimated.

For the spectral parameters, we computed the relative energies of the first twelve harmonics in each frame. The time-averaged values and derivatives of these energies were then computed. Finally, we recorded the three largest principal components of the energy averages and the energy derivatives, respectively.

Six of the 23 parameters appeared to have distinctly non-normal distributions in terms of skewness and kurtosis, and were therefore transformed toward normality with logarithm or power transforms (Table 2).

Description & Classification

(Additional detail on the construction and evaluation of mixed models and linear classifiers can be found in the supplemental materials.)

Principal component analysis was used to reduce the number of variables from the original 23

parameters. Enough components were retained to account for >50% of the variance on each of the original parameters. Save for the largest component (which accounted for >3 times as much variance as any other component), the retained components were varimax-rotated so that each component would be more strongly correlated with one or more original parameters, improving its interpretability.

To investigate sex and age-related variation, we fit linear mixed-effect models to each component. Caller identity and call bout were specified as random effects. Fixed predictors included sex, linear and quadratic terms for log-transformed age. All distinct subsets of main effects and first order interactions were examined, under the usual constraint that no subset may contain a product or quadratic term unless it also contains the main or linear terms as well. This resulted in a total of 6 models. All models were fit with R 2.15.2, using version 3.1-108 of the "nlme" package (Pinheiro et al., 2010).

The explanatory power of models was judged by the Akaike Information Criterion with small-sample correction (AICc, Hurvich & Tsai, 1989). To extract predictions and effect estimates, we followed a model-averaging approach as described by Burnham and Anderson (2002). Model averaging was performed in R 2.15.2 with version 1.27 of the "AICcmodavg" package. Parameters were tested for significant difference from zero, using $\alpha = .0039$, the Dunn-Šidák correction to $\alpha = .05$ for 13 independently tested principal components. Only fixed effects were tested for significance, as AICcmodavg does not permit model-averaged estimation of random parameters; there exists relatively little literature on the latter question.

To investigate behavioral context-related variation, we again fit mixed-effect models to each component, specifying caller identity and call bout as random effects, and caller sex, age, and three orthogonal behavioral context variables as fixed predictors (Table 5). Eight models were compared for each component, and effects estimated using the model averaging approach described above.

Calls were classified by sex, age, and caller identity, using linear discriminant analysis (LDA) on selected sets of principal components. Continuous dependent variables are not suitable for LDA, so we dichotomized age at a threshold of 2.0 years after hatching, thus separating subjects into "juvenile" and "adult" age groups. For the sex (age group) classifier, we used all components that were found to show a significant effect of sex (age group) in the mixed-effect models. If no components were found to show a significant effect, those with a marginally significant effect (corrected $\alpha = .0081$) were used instead. For the identity classifier, we selected components differently, as identity was a random rather than fixed effect. We selected all components for which identity was estimated as accounting for at least 10% of their variance in the mixed-effect models.

Classifier accuracy was compared to the accuracy of three types of "chance" classifier: a classifier that randomly assigns classes to calls with uniform probability; a classifier that assigns all calls to the most common class; and a classifier of the same construction as our true classifiers, but applied to a data set with randomly permuted class memberships, following the recommendation of Mundry & Sommer (2007). The random permutation procedure is described

315 in more detail in the supplementary materials. Classifier accuracy was also cross-validated using a "leave one out" procedure, performed at the level of the next nested grouping variable. That is, the sex and age-group classifiers were repeatedly trained on the calls of all but one individual, then tested on the calls of the remaining individual. The individual identity classifier was repeatedly trained on all but one call bout, then tested on the remaining call bout. Unless
320 otherwise noted, all classifier accuracy values given in the text refer to cross-validated performance.

Weighting

325 Because the number of calls and call bouts available for each bird varied considerably, it was necessary to weight the calls in a non-uniform fashion for data analysis. We chose to weight them following a rule for optimal least squares weighting of group means (Isaev 1979). The resultant weights tended to favor calls belonging to bouts and/or individuals that were represented by few other calls. The total weight assigned to the calls of each individual is shown
330 in Table 1. It can be seen that individuals with more calls tend to receive more total weight, because their mean parameter values can be estimated more accurately; however, they do not receive as much weight as if all calls were weighted uniformly. Additional details can be found in the supplementary materials.

335 These weights were used for calculating all descriptive statistics, for transforming acoustic parameters, for principal component analysis and linear discriminant analysis, and for

measuring the accuracies of our linear classifiers. They were not used in our mixed-effect models, as the algorithms used in nlme allow for unbalanced designs.

III. Results

Principal Components

Thirteen principal components were sufficient to capture more than 50% of the variance of every parameter (Table S1, supplemental). Collectively, they captured 94.14% of the total variance. Principal Component 01 captured 39.58% of the total variance, over three times as much as the next largest component. The other 12 components each captured 3.3-7.4% of the total variance, after varimax rotation (or 2.5-9.2% before rotation). PC 01 was highly positively correlated (Pearson's $r > 0.71$) with three parameters associated with call length, and highly negatively correlated with six parameters associated with call pitch, pitch wobble periodicity, pitch contour concavity and the first spectral component. In other words, calls with high PC 01 values were longer, lower-pitched, with flatter pitch contours that had less regular pitch wobble, and with more energy in the third harmonic as opposed to the second. The distribution of PC 01 was bimodal, but not discrete; no discrete call cluster could be identified via this or any other component (Figure 2).

PC 02 was highly negatively correlated with the residual of various pitch parameters after subtraction of PC 01; that is, calls with high PC 02 values were unusually low-pitched, compared to other calls with similar PC 01 values but low PC 02 values (Table S1). PC 05 was associated with parameters measuring pitch contour symmetry, and PC 08 was associated with parameters

measuring the magnitude of pitch wobble. All other principal components were highly correlated with one parameter each.

The percentage of residual variance (that is, the variance not ascribed to call bout or caller identity in a null model) varied between 8.96% (for PC 01) and 77.48% (for PC 010). For seven principal components, over 50% of the estimated variance was residual, suggesting that calls are not entirely stereotyped within bouts.

Individual Identity

For all but two principal components (PC 01 and PC 03), caller identity accounted for less than 15% of their variance when a null model was used (Table S1). It accounted for 1.5-5.0% of the variance of PCs 05, 08, 09, 10, and 12, and 10.4-14.5% of the variance of PCs 02, 04, 06, 07, 11, and 13.

When all principal components with over 10% of variance accounted for by caller ID were used in an LDA classifier, it correctly classified 35.36% of calls (24.24% cross-validated) by identity, significantly above chance (Table S2, supplemental).

Sex

Only PC 02 showed a significant main effect of sex, with females averaging lower values than males (Table 4). This implies that, for calls with a given value of PC 01, the calls of females are on average higher-pitched than the calls of males. Male and female PC 02 distributions overlap

considerably, leading to a standardized β of 0.3; this predicts that female calls will be approximately 22 Hz higher in mean frequency, and 26 Hz higher in maximum pitch, than male calls with the same value of PC 01. The overlap is apparently due to variance between males, as well as within the call sets of individual males; three males (AS, BF and RE) had mean PC 02 values in the female region.

An LDA classifier using PC 02 correctly classified 66.65% of calls (66.19% cross-validated) by sex, significantly above all "chance" classifiers (Table S2). Two males (BF and RE) and one female (ZU) had calls that were categorized by sex with below-chance accuracy.

Age

No components showed significant main effects of age, although for four components (PCs 08, 11, 12 and 13), the models with linear and/or quadratic age terms had the lowest AICc value (Table 3). Marginal main effects ($p < 0.0081$, equivalent to $\alpha = 0.1$ with Dunn-Šidák correction) were found for PCs 11 and 12, associated with harmonic linear trends and amplitude wobble frequency, respectively.

An LDA classifier using PCs 11 and 12 correctly classified 60.35% of calls (57.52% cross-validated) by age group. This did not significantly exceed the accuracy of a "chance" classifier based on randomly permuted data (Table S2).

Interactions between Sex and Age

Significant age/sex interaction effects were found on PC 09, associated with energy wobble magnitude (Table 3). This interaction is difficult to interpret, as our individuals included males of ages 11 and 13 but no females older than 6. It appears to reflect the fact that PC 09 values were particularly low for females aged 2-5, relative to younger females and males of all ages.

Behavioral Context

Six principal components varied significantly with at least one contextual contrast (Table 5). PC 01 in particular was almost parallel to the first linear discriminant for context (Pearson's r = 0.96). It effectively separated food recruitment calls (which tended to be brief and high-pitched) from alarm calls and beg rebuffs (which tended to be longer and lower-pitched), with counter-cawing falling somewhere in the middle (Figure 3). Certain beg rebuffs were acoustically more similar to food recruitment calls, and may actually have been such calls, since they occurred in a context where a breeding female was near begging offspring and scattered food items simultaneously.

PCs 04, 09, 10 and 13 also varied significantly with certain contrasts, although the effect sizes were not large enough to separate contextual classes of calls using these components. No significant effects of sex or age were observed in models that also included context; in particular, the sex effect on PC 02 did not reach significance (p = .017). However, this p-value was still smaller than the p-values for the context effects on PC 02 (p > .06 in all cases). Furthermore, the

model-averaged, unstandardized effect of sex on PC 02 was not significantly affected by whether the models were fit to known-context or unknown-context calls ($B = 0.46$, $SE = 0.19$ for the known-context subset; $B = 0.28$, $SE = 0.06$ for the unknown-context subset; $z = 0.91$, $p = 0.36$).

IV. Discussion

The primary source of variation in the acoustic structure of "caw"-type calls appears to be behavioral context. A single principal component (PC 01) accounted for approximately 40% of the variance in the parameters we recorded, chiefly those measuring call length, mean pitch, and the wobble and peakedness of the pitch contour. The value of PC 01 did not vary significantly with age or sex, and the majority of its variance was not explained by caller identity. However, it did vary significantly with context, particularly the contrast between food recruitment calls on one hand, and alarm calls and beg rebuffs on the other. While this statistical relationship could only be tested for that subset of calls with known contexts (roughly 18% of the entire call set), context is a plausible explanatory factor for PC 01's variation over the entire set as well, given the component's statistical independence from other caller properties. Unfortunately, the four contextual categories that we were able to identify excluded common behaviors such as aerial mobbing, territorial fights, or calling in chorus, because birds engaging in these behaviors were more difficult to identify and record individually. It is probable that, if detailed contextual information were available for all calls, additional components of acoustic variation would turn out to be context-associated as well. This could be confirmed by studies similar in breadth to the current one, but conducted on calls that have been derived from other contexts, even if produced by unmarked birds.

The calls used for this study span multiple call types, as defined in other studies, where they were quantified mostly on length and the comparative intensities of various harmonics (Chamberlain and Cornwell 1971; Parr 1997; Tarter 2008). We found a pattern of continuous variation in all acoustic parameters (e.g. Figure 2), and hence did not attempt to classify our calls into discrete call types. Calls with low PC 01 values (short and high-pitched) are acoustically similar to Parr's "regular short caws" and "double caws," and to Tarter's "short call" and "doubled short call." Parr associates these calls with a variety of contexts and functions, including "generalized alarm" and "calls-to-arms for family members." In our observations and those of Tarter (2008), they were given in the context of food provisioning on family territories, and family members responded by approaching the caller and foraging. Thus, it appears that these calls can be used for familial recruitment even outside of an "alarm" context.

Interestingly, although the short-duration “ko” or “inflected alarm caw” is commonly observed as an alarm call (Brown 1985; Parr 1997), none of the alarm caws in our known-context subset had low PC 01 values. This may support Brown's contention that “inflected alarm caws” are only given toward a particular class of threat; in our subset, all of the potential predators scolded by crows were mammalian, moving across the ground at some distance from trees, and not obviously pursuing or watching the crows. Brown states that “inflected alarm caws” were given preferentially toward soaring raptors, but Parr objects that they are given toward climbing humans and perched raptors as well. Parr's observations, and an experimental

study on our population by Yorzinski & Vehrencamp (2009), suggest that inflected alarm caws do not denote a particular type of predator. Nevertheless, they could conceivably denote a particular predator location (airborne or arboreal) or threat level, which might explain why we did not record them. Alternatively, they may require body postures or movements (such as flight) which would have prevented us from reliably identifying the caller or establishing context.

Calls with high PC 01 values (long and low-pitched) resemble the "long caw" and "harsh caw" of Parr (1997), and the “rough call” and “scream call” of Tarter (2008). Such calls are typically associated in the literature with mobbing, conspecific aggression and predator alarm. However, as reported in Tarter (2008) and observed by us, some of these calls are also produced in the context of rebuffing begging offspring, and are not accompanied by obviously agonistic behavior.

Calls with moderate PC 01 values were observed during "counter-cawing" between territories and are likely equivalent to Parr's "medium caw" and Tarter's "fading call," calls of moderation duration and pitch that were also observed in that context. Parr found that crows responded strongly to and often approached playbacks of medium caws, providing additional evidence that they are associated with territorial advertisement.

The identity of American crows can be inferred from their calls, with accuracy significantly above chance (24% versus 6%), using a common set of parameters across all "caw"-

type calls. This implies that callers can be identified even when the behavioral context of their calls is unknown, at least if the set of possible candidates is small. Crows may therefore be able to recognize conspecifics at a distance, without first approaching and assessing their behavioral state. This would be a valuable ability for a bird that must maintain long-term family bonds despite frequently traveling miles between home territories, foraging sites and communal roosts (Verbeek and Caffrey 2002). However, Yorzinski et al. (2006) were able to identify callers with considerably higher accuracy than we achieved, using alarm calls alone. This suggests that patterns of inter-individual variation are not very consistent between call types, and that contextual information can therefore significantly improve the accuracy of caller recognition.

Call properties, principally pitch, varied with the sex of the caller. Sex and behavioral context appeared to have independent effects on pitch, as expressed via PC 02 and PC 01; male calls tended to be lower-pitched regardless of context. Here a note of caution is warranted; since the analysis restricted to the known-context subset of calls did not return a significant sex effect on PC 02, it remains possible that the significant sex effect shown in the *full* call set was actually mediated by context. However, we think that this is unlikely to be the case, given the similarity in sex effect sizes between the known-context and unknown-context call subsets, which suggests that the inclusion of context did not attenuate the effect of sex. The reduced significance of the sex effect in the known-context subset was probably due to the smaller size of that subset, which reduced our power for testing all fixed effects.

We are not the first to identify a sex effect on call pitch in crows, although this study

provides new evidence that this difference persists across behavioral contexts. Sexual dimorphisms in call pitch and duration were also found by Yorzinski et al. (2006) when examining alarm calls alone, and pitch dimorphism was found by Laiolo et al. (2000) in another corvid, the red-billed chough *(Pyrrhocorax pyrrhocorax)*. Males commonly produce lower-pitched calls in many avian species (Ballintijn and Cate 1997; Herting and Belthoff 2001).

Much of the acoustic variation we observed, particularly in pitch and frequency properties, may follow from variation in the dimensions of the syrinx and vocal tract. Syringeal size is usually correlated with overall body size, and in many avian species, larger individuals produce lower-pitched calls (Ballintijn and Cate 1997). Body size data were not available for our study subjects at adulthood, as they were banded in the nest, but American crows do generally show a slight sexual size dimorphism (Clark, James, and Morari 1991).

Thanks to this variation, the sex of an American crow can be inferred from a single call with above-chance accuracy: in our study, sexing accuracy was 66.65%, cross-validated. This level of sex classification accuracy was relatively low, compared to that achieved in other studies (65 - 100%) using a combination of body size metrics measured from birds in the hand (Clark, James, and Morari 1991; Yaremych et al. 2004; Ludwig, Begras-Poulin, and Lair 2009). It is also lower than that achieved by a prior study on this population (87% non-cross-validated, excluding one particularly atypical male), using the acoustic properties of single alarm caws (Yorzinski et al. 2006). This suggests that acoustic sexing, like caller identification, would be more accurate when applied to calls from a single behavioral context. Sexing error was largely

due to inter-individual variance between males; three males had mean discriminant values that fell within the range spanned by the set of female means. Of these three, two were yearlings and one was a breeding male.

It should be noted that methods of sexing crows at a distance are generally unreliable. The sexes are not visually dimorphic, save for a brood patch in breeding females, and there are no known visually distinctive sex-specific behaviors except during copulation and egg-laying. Even the tail-quivering precopulatory display, reported from females of many corvid species, is performed by both sexes of American crow (Verbeek 1972; Kilham 1989). Using calls, therefore, may be the most effective way for both conspecifics and human researchers to sex crows at a distance, even if its accuracy is also limited.

Our attempts at age prediction met with little success. A linear classifier did not distinguish yearlings from older birds with sufficient accuracy to permit confidence that the classifier was not merely capitalizing on chance variation. Given that most yearling crows do not breed, it is notable and somewhat surprising that their vocal behavior is not readily distinguishable from that of adults.

Thus, as we predicted, the sex and individual identity of a crow can be inferred from its call across multiple contexts; but, contrary to our prediction, its age cannot. This suggests that it may be adaptive for a crow to communicate sex and identity, but less so to communicate age, perhaps because younger birds make less desirable mates.

With the aid of a PMLE-based pitch estimator, we have demonstrated that American crow caws are locally periodic and can be therefore decomposed into fundamental frequency and spectral envelope components for subsequent analysis. The calls exhibit rapid oscillations in fundamental frequency and amplitude, which are relatively difficult to capture via traditional Fourier-based methods; we suggest that "pitch-first" analysis methods may be helpful for the calls of other corvid species as well. We have also shown that behavioral context, and the sex and identity of the caller, can be inferred from the structure of individual calls. This suggests that American crows could themselves make such inferences. Further improvements in automated classification along these lines may also point to practical methods for acoustic censusing of endangered and vulnerable corvid populations.

Acknowledgments

Financial support for this study comes principally from XXXXX, and from the National Institutes of Health (XXXXX).

References

Allenbacher R, Böhner J, Hammerschmidt K. 1995. Individuelle Merkmale im "krah"-Ruf der Nebelkrähe *Corvus corone cornix*. Journal für Ornithologie 136(4): 441-446.

Baeyens G. 1981. The role of the sexes in territory defence in the magpie (*Pica pica*). Ardea 69:69–82.

Ballintijn MR, Cate C ten. 1997. Sex differences in the vocalizations and syrinx of the collared dove (*Streptopelia decaocto*). The Auk 114:22–39.

Blumstein, DT, Verneyre L, Daniel JC. 2004. Reliability and the adaptive utility of discrimination among alarm callers. Proceedings of the Royal Society of London. Series B: Biological Sciences, 271(1550):1851–1857.

Boeckle M, Szipl G, Bugnyar T. (2012). Who wants food? Individual characteristics in raven yells. Animal Behaviour, 84(5):1123–s1130.

Brown ED. 1985. Functional interrelationships among the mobbing and alarm caws of Common Crows (*Corvus brachyrhynchos*). Zeitschrift für Tierpsychologie 67:17-33.

Chamberlain DR, Cornwell GW. 1971. Selected vocalizations of the common crow. The Auk 88:613–634.

Charrier I, Jouventin P, Mathevon N, Aubin T. 2001. Individual identity coding depends on call type in the South Polar skua *Catharacta maccormicki*. Polar Biology 24:378–382.

Clark AB, Robinson DA Jr., McGowan KJ. 2006. Effects of West Nile Virus Mortality on Social Structure of an American Crow (*Corvus brachyrhynchos*) Population in Upstate New York. Ornithological Monographs 60:65-78.

Clark RG, James PC, Morari JB. 1991. Sexing adult and yearling American crows by external measurements and discriminant analysis. Journal of Field Ornithology 62:132–138.

Cohen L. 1989. Time-frequency distributions--a review. Proceedings of the IEEE DOI - 10.1109/5.30749 77:941–981.

Dale J, Lank DB, Reeve HK. 2001. Signaling individual identity versus quality: a model and case studies with ruffs, queleas, and house finches. The American Naturalist, 158(1):75-86.

Davis LI. 1958. Acoustic evidence of relationship in North American crows. The Wilson Bulletin, 70(2):151–167.

Ekman J, Ericson PGP. 2006. Out of Gondwanaland; the evolutionary history of cooperative breeding and social behaviour among crows, magpies, jays and allies. Proceedings of the Royal Society B: Biological Sciences 273:1117 –1125.

Fischer J, Hammerschmidt K, Cheney DL, Seyfarth RM. 2001. Acoustic features of female chacma baboon barks. Ethology 107:33–54.

Frings H, Frings M, Jumber J, Busnel RG, Giban J, & Gramet P. 1958. Reactions of American and French species of *Corvus* and *Larus* to recorded communication signals tested reciprocally. Ecology 39(1):126-131.

Ha RR, Bentzen P, Marsh J, Ha JC. 2003. Kinship and association in social foraging northwestern crows (*Corvus caurinus*). Bird Behavior 15:65–75.

Heinrich B. 1988. Winter foraging at carcasses by three sympatric corvids, with emphasis on recruitment by the raven, *Corvus corax*. Behavioral Ecology and Sociobiology 23:141–156.

Herting BL, Belthoff JR. 2001. Bounce and double trill songs of male and female western screech-owls: Characterization and usefulness for classification of sex. The Auk 118:1095–1101.

Hopp SL, Jablonski P, Brown JL. 2001. Recognition of group membership by voice in Mexican jays, *Aphelocoma ultramarina*. Animal Behaviour 62:297–303.

Isaev, AB. 1979. Applicability of the generalized least-squares method for processing correlated observations. Measurement Techniques 22(8):924–926.

Kilham L. 1989. The American Crow and the Common Raven. 1st ed. College Station: Texas A&M University Press Available from: http://books.google.com/books?id=GDYXjr2Pk-8C

Kondo N, Izawa EI, Watanabe S. 2010. Perceptual mechanism for vocal individual recognition in jungle crows (*Corvus macrorhynchos*): Contact call signature and discrimination. Behaviour 147:1051–1072.

Kondo N, Izawa EI, Watanabe S. 2012. Crows cross-modally recognize group members but not non-group members. Proceedings of the Royal Society B: Biological Sciences, 279(1735):1937–1942.

Laiolo P, Palestrini C, Rolando A. 2000. A study of choughs' vocal repertoire: variability related

to individuals, sexes and ages. Journal für Ornithologie 141:168–179.

Laiolo P, Rolando A. 2003. The evolution of vocalisations in the genus *Corvus*: effects of phylogeny, morphology and habitat. Evolutionary Ecology 17:111–123.

Ludwig A, Begras-Poulin M, Lair S. 2009. Morphological description of American crow, *Corvus brachyrhynchos*, populations in southern Quebec. The Canadian Field-Naturalist 123:133–140.

Mazerolle, M. 2013. AICcmodavg: Model selection and multimodel inference based on (Q)AIC(c). R package version 3.0.1.

McArthur PD. 1982. Mechanisms and development of parent-young vocal recognition in the piñon jay (*Gymnorhinus cyanocephalus*). Animal Behaviour 30:62–74.

McGowan KJ. 1997. Reproductive and social behavior of two crow species in New York. U.S. Department of Agriculture. Available from: http://www.birds.cornell.edu/crows/hatchrep.html

McGowan KJ. 2001. Demographic and behavioral comparisons of suburban and rural American Crows, p. 365–381. In J. M. Marzluff, R. Bowman, and R. Donelly [EDS.], Avian ecology and conservation in an urbanizing world. Kluwer Academic Press, Norwell, MA.

Mundry R, Sommer C. 2007. Discriminant function analysis with nonindependent data: consequences and an alternative. Animal Behaviour 74(4):965–976.

Parr CS. 1997. Social behavior and long-distance vocal communication in Eastern American crows [doctoral dissertation]. University of Michigan. Available from http://mirlyn.lib.umich.edu/Record/003940533 by subscription.

Pinheiro J, Bates D, DebRoy S, Sarkar D, and the R Development Core Team. 2013. nlme: Linear and Nonlinear Mixed Effects Models. R package version 3.0.1.

Pollard KA. 2011. Making the most of alarm signals: the adaptive value of individual discrimination in an alarm context. Behavioral Ecology, 22(1):93–100.

Rendall D, Owren MJ, Rodman PS. 1998. The role of vocal tract filtering in identity cueing in rhesus monkey (*Macaca mulatta*) vocalizations. J. Acoust. Soc. Am. 103:602–614.

Richards DB, Thompson NS. 1978. Critical properties of the assembly call of the common American crow. Behaviour 64(3-4): 184-203.

Robinson Jr. DA. 2009. The relationship of nestling qualities to survival and breeding strategies of cooperatively breeding American crows in Ithaca, NY [dissertation]. State University of New York at Binghamton. Available from: http://alephprod.binghamton.edu by subscription.

Røskaft E, Espmark Y. 1984. Sibling recognition in the rook (*Corvus frugilegus*). Behavioural Processes 9:223–230.

Tarter RR. 2008. The vocal behavior of the American crow, *Corvus brachyrhyncos* [master's thesis]. Ohio State University. Available from: http://rave.ohiolink.edu/etdc/view?acc_num=osu1204876597

Thompson NS. 1982. A comparison of cawing in the European carrion crow (Corvus corone) and the American common crow (*Corvus brachyrhynchos*). Behaviour 80(1-2):106–117.

Tibbetts, EA, Dale J. 2007. Individual recognition: it is good to be different. Trends in Ecology

& Evolution, 22(10), 529–537.

Verbeek NAM, Caffrey C. 2002. American Crow (*Corvus brachyrhynchos*), Issue No. 647. Birds of North America Online [Internet]. Available from: http://bna.birds.cornell.edu /bna/species/647/articles/introduction

Verbeek NAM. 1972. Comparison of displays of the yellow-billed magpie (*Pica nuttalli*) and other corvids. Journal of Ornithology 113:297–314.

Wascher CAF, Szipl G, Boeckle M, Wilkinson A. 2012. You sound familiar: carrion crows can differentiate between the calls of known and unknown heterospecifics. Animal Cognition 15(5):1015-1019.

Yaremych SA, Levengood JM, Novak RJ, Mankin PC, Warner RE. 2004. Gender determination and lack of sex-specific West Nile virus mortality in American crows. Wildlife Society Bulletin 32:893–899.

Yorzinski JL, Vehrencamp SL, Clark AB, McGowan KJ. 2006. The inflected alarm caw of the American crow: Differences in acoustic structure among individuals and sexes. The Condor 108:518–529.

Yorzinski JL, Vehrencamp SL. 2009. The effect of predator type and danger level on the mob calls of the American crow. The Condor 111:159–168.

Bird ID	Sex	Age	Family	# of Calls Used	# of Call Bouts Used	# of Days Recorded	Fraction of Total Call Weight
MH	f	1	NEEL	145	41	3	6.82%
UB	f	1	CLAR	15	5	2	3.71%
ZU	f	1	WWCK	120	38	4	6.62%
AZ	f	2	NEEL	39	12	6	5.12%
OY	f	4	WWCK	337	87	10	7.61%
8Z	f	5	SEPG	117	25	7	6.30%
N1	f	6	WKAY	175	44	10	6.91%
AS	m	1	WWCK	178	35	7	6.71%
IL	m	1	NEEL	80	28	2	6.35%
KJ	m	1	NEEL	29	12	3	5.09%
RE	m	1	CLAR	24	6	3	4.01%
XW	m	1	WWCK	169	40	6	6.83%
FT	m	2	CLAR	58	10	3	5.07%
33	m	3	ORHA	49	27	5	6.10%
0E	m	4	NEEL	25	6	2	4.10%
0O	m	4	SEPG	7	5	1	3.45%
BF	m	11	ROWA	10	3	2	2.96%
AP	m	13	WKAY	97	25	6	6.25%

Table 1. Demographic and recording data for subjects of this study. Ages are given in years since fledging; all birds were at least one year old.

	Parameter	M	σ	γ_1: Skew	γ_2: Excess Kurt.	R^2 by Canon. Vars.
Pitch Contour	01. Mean F0	604.34	75.82	-0.19	-0.99	0.98
	02. 95th Pctl F0	696.63	66.08	-0.35	-0.46	0.95
	03. 5th Pctl F0	476.60	84.26	0.58	-0.81	0.84
	04. F0 Peak Location*	-0.13	0.19	0.44	1.38	0.86
	05. F0 Peak Value	668.56	77.95	-0.50	-0.08	0.93
	06. F0 quadratic term	187.71	83.46	0.00	-0.60	0.90
	07. F0 cubic term*	50.67	35.25	-1.60	4.70	0.90
	08. Wobble Frequency	50.15	12.04	0.04	-0.85	0.99
	09. Wobble Periodicity 1*	0.52	0.13	0.46	0.07	0.97
	10. Wobble Periodicity 2	0.38	0.15	0.67	0.31	0.98
	11. Pitch Instability	28.16	12.88	0.80	1.57	0.72
Power Envelope	12. Call Length*	-2.56	0.75	0.23	-0.76	0.96
	13. 2nd central moment*	-4.59	0.78	0.32	-0.61	0.97
	14. 3rd central moment	0.02	0.04	0.19	1.50	0.94
	15. 4th central moment*	-4.29	0.79	0.28	-0.74	0.97
	16. Wobble Frequency	47.97	15.67	0.55	-0.84	0.99
	17. Wobble Magnitude	0.38	0.17	0.26	-0.45	0.97
Spectral Properties	18. Time-averaged harmonic powers, 1st principal component	0.60	0.99	-0.46	-1.13	0.89
	19. Time-averaged harmonic powers, 2nd principal component	-7.62	1.01	2.68	10.89	1.00
	20. Time-averaged harmonic powers, 3rd principal component	4.36	1.00	-0.11	12.80	0.98
	21. Linear trend in harmonic powers, 1st principal component	0.19	0.98	-1.19	3.50	0.99
	22. Linear trend in harmonic powers, 2nd principal component	-0.29	1.03	-1.69	9.59	0.99
	23. Linear trend in harmonic powers, 3rd principal component	-0.11	0.99	-0.48	31.28	0.98
*Log- or power transformed. See text for details.						

Table 2. Descriptive statistics of the acoustic call parameters used. The last column contains the coefficient of multiple determination for each parameter, when regressed on the 13 canonical variables used for call classification.

Principal Component	β: Sex	β: Age (Lin.)	β: Age (Quad.)	β: Sex:Age (Lin.)	β: Sex:Age (Quad.)	**Lowest AICc Model**	**Akaike Weight of Lowest Model**
01	0.13	0.03	0.14	-0.15	0.19	Null	0.37
02	0.30**	0.07	0.09	-0.08	-0.19	Sex	0.45
03	0.00	-0.14	0.05	-0.08	0.28	Null	0.31
04	0.00	-0.11	0.15	0.01	0.07	Null	0.3
05	0.07	-0.05	0.00	-0.02	0.18	Null	0.3
06	0.01	-0.06	-0.10	0.01	-0.01	Null	0.4
07	0.01	0.09	0.29	-0.12	-0.44†	Null	0.28
08	0.04	0.08	0.02	-0.07	0.02	Age (Lin.)	0.21
09	0.00	0.11	0.21†	-0.15*	-0.23**	Sex * Age (Quad.)	0.61
10	-0.02	-0.02	-0.01	0.08	-0.07	Null	0.42
11	0.10	0.23†	-0.02	0.05	0.16	Age (Lin.)	0.35
12	0.01	-0.07	-0.15†	-0.03	-0.05	Age (Quad.)	0.58
13	0.10	-0.19†	-0.03	0.03	-0.02	Age (Lin.)	0.31
							†p < 0.0081 ***p < 0.0039** ****p < 0.00077**

Table 3. Sex and age related variation in the principal components of acoustic parameters, on the full set of 1674 calls. β coefficients are averaged across mixed-effect models containing various subsets of sex and linear/quadratic age as fixed effects. Reported *p*-values correspond to *t*-tests that each β coefficient differs from zero.

Behavioral Context	Observational Criteria	#of Caws Included in Set	#of Male Callers	# of Female Callers
Food Recruitment	Given by perched or standing birds on their territory, after bait was provided. Family members responded by approaching and foraging or begging.	155	1	1
Territorial Counter-Cawing	Given by perched birds facing out of territory, while birds on neighboring territories responded with a similar cadence.	13	1	2
Alarm Call	Given by perched birds watching a mammalian predator or potential predator (cat, human, skunk or squirrel) and calling with head downward.	48	3	1
Beg Rebuff	Given by adult birds in response to begging juveniles while on the ground. They did not feed the juveniles or coo at them, and almost always moved away while calling. Several times, they took flight and were pursued by juveniles at a low altitude.	89	0	3

Table 4. Observed behavioral contexts associated with a subset of calls (305 out of 1674).

Principal Component	β: Sex	β: Age (Lin.)	β: C1	β: C2	β: C3
01	0.15	-0.10	-0.02	-0.97**	0.19
02	0.46	0.32	-0.43	-0.27	0.05
03	0.34	0.31	0.22	0.03	0.28
04	-0.22	-0.88	0.30	-1.07**	-0.31
05	0.16	0.05	-0.15	0.00	0.03
06	0.25	0.07	0.41	-0.14	0.28
07	-0.37	0.49	-1.35	0.57	1.13*
08	0.10	0.01	0.25	0.18	-0.05
09	0.15	-0.14	-0.40	-0.33**	0.11
10	0.09	0.13	-0.40	0.34**	-0.06
11	-0.10	0.21	-0.35	-0.07	-0.08
12	-0.14	0.10	-0.34	-0.02	-0.03
13	-0.30	0.35	-0.20	0.33*	0.78*
***p < 0.0039** ****p < 0.00077** **C1: Counter-Cawing vs. Food Recruitment, Alarm Call and Beg Rebuff** **C2: Food Recruitment vs. Alarm Call and Beg Rebuff** **C3: Alarm Call vs. Beg Rebuff**					

Table 5. Behavioral context-related variation in principal components of acoustic parameters, on a subset of 305 calls. β coefficients are averaged across mixed-effect models containing various subsets of sex, linear age and behavioral context as fixed effects. Context is coded with three orthogonal contrasts. Reported *p*-values correspond to *t*-tests that each coefficient differs from zero.

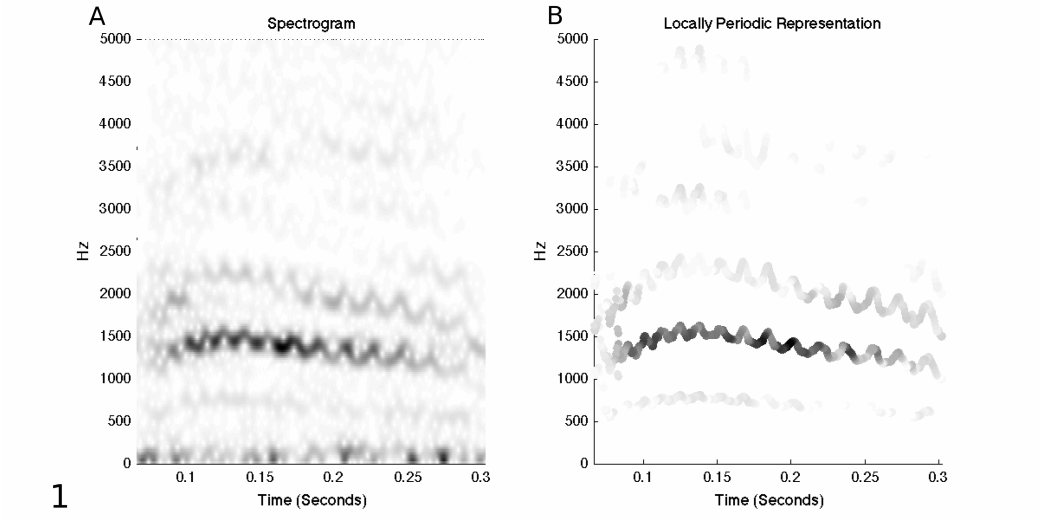


Figure 1. Traditional and locally periodic representations of a crow call. The call is from AS, a yearling male. (A) Traditional spectrogram, using 512-sample Hamming-windowed frames and a 6-sample step size. (B) Locally periodic representation, using 240-sample frames and a 24-sample step size. Note that this representation assigns energy only to frequencies corresponding to harmonics of the estimated fundamental for each frame.

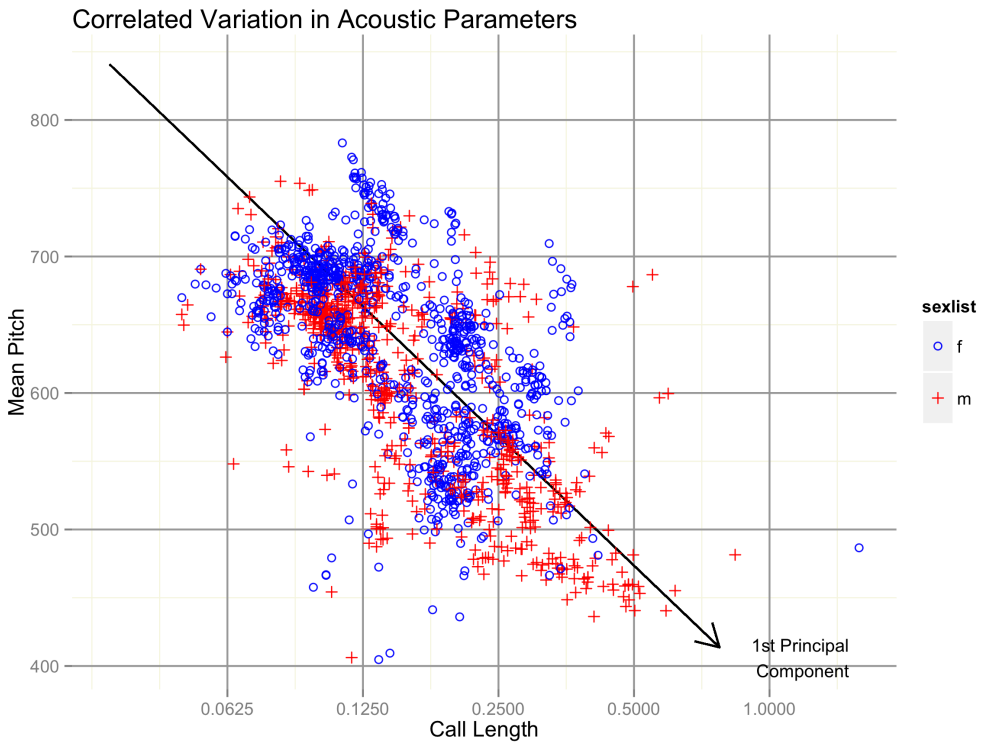


Figure 2. Distribution of call lengths and mean pitches among all calls, with first principal component of acoustic parameters projected onto this surface.

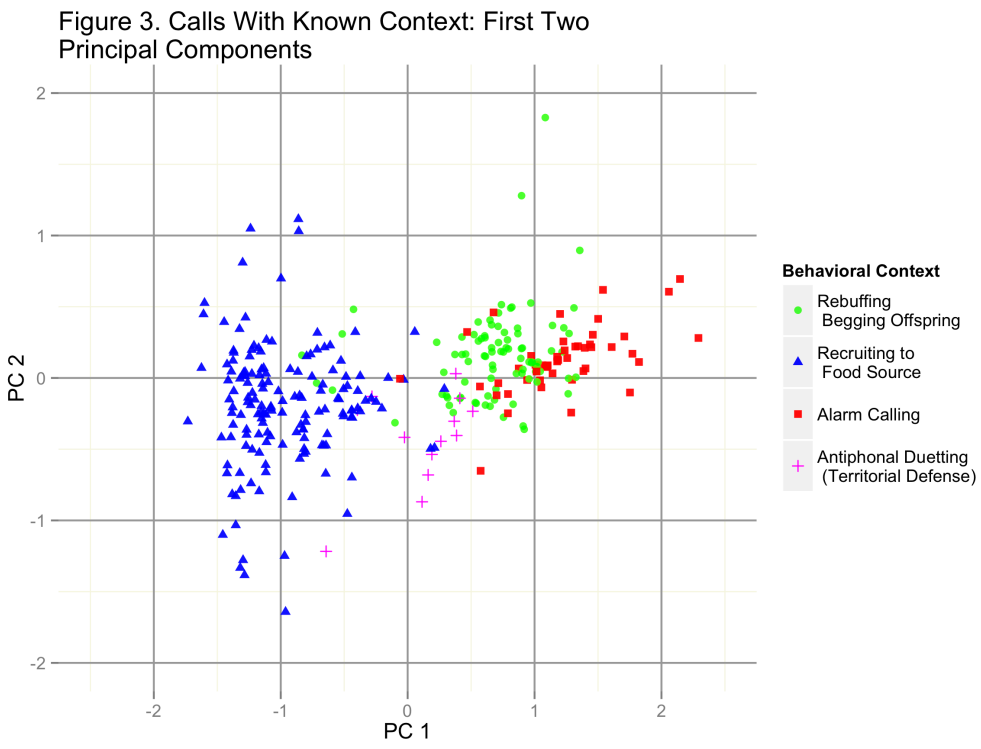


Figure 3. Distribution of calls with known behavioral context, on two principal components.

Supplementary Details on Methodology

Representation of Calls as Locally Periodic Signals

Calls were represented as locally periodic signals with smoothly varying fundamental frequency (pitch). They were divided into overlapping frames, with a time step of 0.5 milliseconds, and within each frame they were represented as a sum of harmonics of a single fundamental frequency. Frame length varied between 4 and 7 milliseconds, being chosen for each call to accommodate two cycles of the lowest plausible fundamental frequency, as judged by a human observer visually assessing the spectrogram; in almost all cases, the frame length allowed for a minimum frequency of 325-350 Hz.

The fundamental frequency of each frame was determined using an extension of Friedman's Pseudo-Maximum Likelihood Estimator (PMLE), developed to estimate the pitch of human speech (Friedman 1977). The PMLE measures the goodness-of-fit of the frame's signal to a given integer-valued period, by making a periodic approximation to that signal using a weighted least squares criterion, using the frame's windowed function as a weight function. The ratio of the energies of the periodic approximation and original signal is then taken, after a term representing the contribution of broadband noise has been subtracted from both energies.

We have developed two period indicator functions from the PMLE, which will be summarized here (E. A. Mates and J. C. Ha, unpublished data). The *approximate pseudo-maximum likelihood estimator* (APLE) approximates the PMLE with a weighted sum of

autocorrelations that can be computed more efficiently. Whereas the PMLE for a set of periods of (1,2...L) samples is computed by an algorithm of order L^2 flops, the APLE for the same set of periods can be computed by a faster, FFT-based algorithm of order L log (L) flops. The *extended pseudo-maximum likelihood estimator* (EPLE) calculates the PMLE precisely, but allows for non-integer-valued periods and periods that vary over the span of a frame, and for arbitrarily chosen bandwidths. It requires the solution of a least-squares optimization problem; we have developed a Lanczos-like algorithm to efficiently solve the normal equations for this system. These two period indicators can be used in tandem, with fundamental frequency candidates rapidly approximated by APLE, then refined using peak-finding algorithms based on EPLE.

To process a call for this study, we bandpass-filtered each of its frames for frequencies between 300-450 Hz and 10 kHz, and used APLE to identify a small number of fundamental frequency candidates per frame. These candidates were then concatenated across frames into up to 50 pitch traces, chosen by a greedy algorithm that attempted to maximize average APLE value, average pitch, and smoothness (quantified as similarity of the pitch trace to a low-pass-filtered copy of itself). The optimal pitch trace was selected from this set, and EPLE was used to fine-tune its fundamental frequency estimate and produce a periodic approximation to the signal for each frame. The resultant locally periodic approximations captured an average of 92.90% of each filtered frame's energy, for the calls used in this set. Fundamental frequency did indeed vary smoothly, with an average RMS fundamental frequency shift of 12.67 Hz between frames.

Parameterization of calls

Each clip of a single call was automatically partitioned into a "voiced" section preceded and followed by "silent" sections. This was done by computing the fifth root of call energy for each frame, and collecting the frames into voiced and silent groups, such that the ratio of between-group variance to within-group variance in this quantity was maximized. Only the voiced section was used for further analysis. The time variable for each call was mean-centered, using frame energy as a weight function. Frame energy was also used as a weight function in all other quantities that required integrating or averaging across frames.

Twenty-three parameters were extracted from each call (Table 2): 11 based on the pitch trace; six based on the power envelope; and six based on the spectral properties. Among the pitch trace parameters, the mean, 5th percentile and 95th percentile pitch were estimated from the pitch distribution. A cubic polynomial was fit to the pitch trace, and the position in time of the cubic curve's local maximum was recorded, as well as the value of that maximum and the values of the curve's second and third derivative at the point. The cubic curve was then subtracted from the pitch trace to leave a residual. The RMS magnitude of the residual was recorded as the "pitch instability." The residual was expected to contain a "wobble" or high-frequency modulation in pitch, which is typically found in "caw"-type calls across the genus *Corvus* (Laiolo and Rolando 2003). The overall fundamental frequency of this wobble was estimated using the APLE method described above, as confined to the region 30-90 Hz. Two estimates of the magnitude of the pitch wobble were computed: the peak value of the APLE, and the fraction of the wobble's mean square amplitude that was contained in a periodic

approximation.

Among the signal power parameters, the length in time of the voiced interval in the clip was used as a measure of call length. The second through fourth central moments of call energy with respect to time were also computed. As in the pitch trace, there was expected to be a rapid wobble in signal power. Its fundamental frequency was estimated using the APLE method, and its magnitude was estimated as the peak value of the APLE.

For the spectral parameters, we computed the power of the first twelve harmonics in each frame, normalized by the total power in that frame. The time averages (means) and average derivatives of power were then computed. Finally, we recorded the three largest principal components of the harmonic power averages, and the three largest principal components of the power derivatives.

Six of the 23 parameters appeared to have distinctly non-normal distributions in terms of skewness and kurtosis, and were transformed accordingly. Parameters 12, 13, and 15, were transformed with a logarithm base two (Table 2). Parameters 04, 07, and 09 were subjected to a shifted power transform of the form $Y= \mathrm{Sign}(X - \alpha) \cdot |X - \alpha|^{\lambda}$. For parameter 04, $\alpha = 0.005$ and $\lambda = 0.4$; for parameter 07, $\alpha = 0$ and $\lambda = 1/3$; and for parameter 09, $\alpha = 0$ and $\lambda = 0.5$.

Description & Classification

To investigate sex and age-related variation, we fit linear mixed-effect models to each component. Random, intercept-only effects of caller identity and call bout were specified (assumed to be Gaussian in distribution.) Fixed predictors included sex, linear and quadratic terms for log-transformed age, and products of the former with the latter. All distinct subsets of effects were examined, under the usual constraint that no subset may contain a product or quadratic term unless it also contains the main or linear terms as well. This resulted in a total of 6 models. All models were fit with R 2.15.2, using version 3.1-108 of the "nlme" package (Pinheiro et al., 2010).

The explanatory power of models was judged by the Akaike Information Criterion with small-sample correction (AICc, Hurvich & Tsai, 1989). To extract predictions and effect estimates, we followed a model-averaging approach as described by Burnham and Anderson (2002). Each model is assigned an Akaike weight proportional to the inverse of the exponential of its AICc value, and predictions are averaged across all models using these weights. Model parameter values and effect sizes are also averaged, across the subset of models that contain that parameter and do not contain product or quadratic parameters. This approach avoids overly privileging the single "best" model if it has very close competitors in AICc values, and has repeatedly shown superior predictive performance to the best-model approach (Burnham & Anderson, 2004). Model averaging was performed in R 2.15.2 with version 1.27 of the "AICcmodavg" package. Parameters were tested for significant difference from zero, using $\alpha = .0039$, or the Dunn-Šidák correction to $\alpha = .05$ for 13 independently tested principal components.

Weighting

Because the number of calls available for each bird varied dramatically, and call parameter values were expected to depend on caller identity and call bout, it was necessary to assign the calls non-uniform weights for data analysis. We chose to weight them following the rule for optimal least squares weighting of group means: the row/column sum of the precision matrix, or inverse of the expected covariance matrix between observations (Isaev 1979). Note that the entries of the latter matrix contain the *expected* covariance in the values of a single variable, for a given pair of observations, rather than the *observed* sample covariance between a pair of variables.

Accordingly, for each parameter, we fit the observations to a null model with random effects of caller identity and call bout, and no fixed effects. We then extracted the estimated observation covariance matrices, inverted them and took row sums. Finally, we normalized these sums to be weights, and then averaged the weights across all 29 parameters. The resultant weights tended to favor calls from bouts or individuals with few other calls. The total weight assigned to the calls of each individual is shown in Table 1. It can be seen that individuals with more calls tend to receive more total weight, because their mean parameter values can be estimated more accurately; however, they do not receive as much weight as if all calls were weighted uniformly.

These weights were used for calculating all descriptive statistics (such as means and higher moments), for scaling and centering parameters, and for generating covariance matrices between variables for purposes of principal component analysis and linear discriminant analysis; and for measuring the accuracies of linear classifiers. They were not used in our mixed-effect models, as the algorithms used in nlme allow for unbalanced designs.

It should be noted that our results were not particularly sensitive to the exact weighting scheme used; when we applied other weighting schemes such as taking the inverse square root of the number of calls for that individual, the principal component analysis breakdown and the estimated sex/age effects were very similar in size and significance.

Assessing Classifier Accuracy: Correcting for Capitalization on Chance Variation

If classifiers can capitalize on random variation associated with a grouping variable nested inside the class of interest, there is a risk that they will exhibit spuriously high accuracy. For instance, if there is real variation between individuals but not between sexes, a sex classifier may still achieve high accuracy on its training set, simply by exploiting any overall differences that randomly occur between the male and female individuals in that set. Likewise, an individual identity classifier may achieve spuriously high accuracy by exploiting variation between particular call bouts in its training set.

To avoid this problem, we assessed classifier performance in two ways beyond raw

accuracy. First, we cross-validated performance using a "leave one out" procedure, performed at the level of the next nested grouping variable. I.e., for sex and age-group classifiers, each individual's calls were classified using a classifier trained on the calls of all other individuals. For the individual identity classifier, calls in a given bout were classified using a classifier trained on all other bouts. Unless otherwise noted, all classifier accuracy values given in the text refer to cross-validated performance.

Second, as alluded to above, we performed a permutation test as proposed by Mundry & Sommer (2007). Class membership was repeatedly and randomly permuted at the level of the next nested grouping variable. I.e., the individuals were reshuffled between sexes/age groups for the sex and age-group classifiers, and the call bouts were reshuffled between individuals for the individual identity classifier. We then estimated *p* values for the classifiers' accuracy on the true data sets, according to the empirical probability of achieving higher accuracy on a randomly permuted data set. 250 permutation runs were performed for sex and age group, and 1000 runs were performed for individual identity.

170 Supplementary Tables and Figures

Principal Component	Fraction of total variance	% of variance due to ID	% of variance due to bout	Residual % of variance	Positively correlated with parameters ($r > 0.7071$)	Negatively correlated with parameters ($r < -0.7071$)
01	39.58	32.24	58.81	8.96	12,13,15	1,3,5,6,10,18
02	7.43	13.48	61.99	24.53		Residual 1,2,5
03	3.7	18.77	42.80	38.44	14	
04	4.37	14.54	50.98	34.49	20	
05	4.72	2.13	33.82	64.05	Residual 7	Residual 4
06	4.48	11.36	36.63	52.01	22	
07	4.44	14.32	68.24	17.44		19
08	5.14	1.65	37.27	61.07		9, Residual 10
09	4.11	3.61	27.96	68.43	17	
10	4.62	2.28	20.24	77.48	8	
11	4.52	10.60	17.58	71.81		23
12	3.77	5.00	22.20	72.80	16	
13	3.26	10.43	42.94	46.63		21

175 **Table S1.** Principal Components derived from a PCA of all acoustic parameters. These 13 variables were sufficient to capture >50% of the variance of every parameter. Variables 2-13 were varimax rotated for superior interpretability. Fractions of variance due to caller ID, bout, and residual variance, were estimated by mixed-effect models using ID and bout as random effects and no fixed effects.

180

	Individual Identity, PCs 01, 02, 03, 04, 06, 07, 11, 13	**Sex PC 02**	**Age Group, PCs 11,12**
Training Set Accuracy	35.36%***	66.65%*	60.25%
Cross-validated Accuracy	24.24%	66.19%	57.52%
Expected "Uniform Random Change" Accuracy	5.88%	50.00%	50.00%
Expected "Most Frequent Group" Accuracy	7.35%	50.00%	50.00%
Nested Grouping Variable	Call Bout	Caller Identity	Caller Identity
Number of Random Permutation Runs	1000	300	300
Empirical Average Accuracy under Random Permutation	19.10%	57.23%	55.34%
			***p < 0.05 ***p < 0.001**

Table S2. Performance of linear discriminant analysis (LDA) classifiers for individual identity, sex and age group (immature yearlings versus mature birds of ages 2 and up) on all calls. Reported *p*-values correspond to deviation from the empirical average accuracy, under the random permutation method described in the supplementary text. E.g., $p < 0.05$ indicates that when class membership was randomly permuted, the resulting classifier performed less accurately than it had performed on the original data set at least 95% of the time.

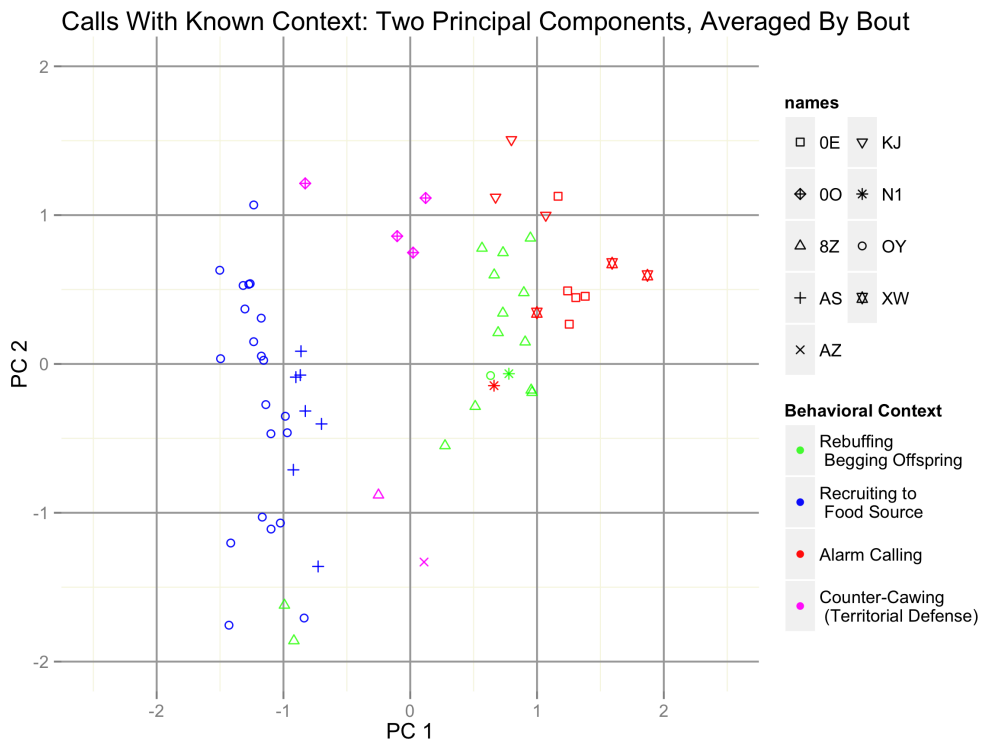


Figure S1. Distribution of call bout averages with known behavioral context, on two principal components. Colors represent behavioral context; marker shapes represent caller identity.

CHAPTER 5

Alarm Behavior of Northwestern Crows in response to Visual and Auditory Stimuli

Anton Mates

Dissertation Chapter

Alarm Behavior of Northwestern Crows in response to Visual and Auditory Stimuli

Introduction

Definition of mobbing behavior

Collective defense against predation is found in a variety of vertebrate taxa, such as songbirds, social carnivores, primates, and colonially nesting fish. In birds, it often takes the form of "mobbing," which involves exceptionally close approaches to, and even attacks on, the predator in question. Shields (1984) defines mobbing in this way:

"Avian mobbing is usually defined as an approach towards a potentially dangerous predator (whether it is actively hunting or not), followed by frequent position changes with most movements centred on the predator. Relatively stereotyped visual displays and loud and localizable vocal displays usually accompany the locomotion. Often mobbing includes swoops or runs at a potential predator and it may involve direct attack, with physical contact by the mobber."

Mobbing is clearly costly, not only because it consumes time and energy, but also because mobbing animals may place themselves at higher risk of predation. [Sordahl (1990)] What, then, are the compensatory benefits of this costly behavior?

Survey of functional hypotheses for mobbing

Dugatkin and Godin (1992) identify four general categories of benefits that could be provided by mobbing. I summarize them below:

- **Deterrence:** A reduction in the probability that the predator will attack the mobbing bird or its kin/social allies. This may occur if the predator fears that the mobbing bird will injure it, if it prefers to attack from ambush, or if it conversely pursues the mobbing bird rather than more vulnerable kin.
- **Surveillance:** An opportunity to gather additional information about the predator's location and threat level.
- **Conspecific Communication:** Informing kin/social allies about the threat posed by the predator.
- **Advertisement:** A demonstration of the mobbing bird's body condition and protective tendencies to potential mates or social allies.

Deterrence may provide either direct or indirect fitness benefits, depending on whether the risk of predation is lowered for the mobbing bird, its kin, potential mates, or social allies. Surveillance and Advertisement provide direct benefits, and Conspecific Communication indirect benefits, though one can easily imagine secondary effects that lead to benefits of another type. For instance, Surveillance could inform Deterrence or Communication behavior that benefits allies or kin, while Communication could produce a population of predator-informed birds who could in turn engage in Deterrence or Communication that benefits the original bird.

The mobbing aggregations of corvids, in particular, are notorious for their size, scale, and duration. Mobs have been observed to contain up to 200 birds and last for over 90 minutes [Frings et al., 1958; Morse, 1971]. I have observed a breeding pair of American crows leave their territory, and their fledged offspring, to fly over a kilometer and join a mob chasing a red-tailed hawk [unpublished data].

Corvid mobbing includes the following five classes of behavior:

- Aggregation: approach to or assembly near the predator;
- Elevation: leaving the ground to perch or fly;
- Flyovers: Repeated low, semi-circular flights above the predator;
- Scolding: the directed production of harsh and irregularly spaced alarm calls; and
- Physical aggression: acts such as dive-bombing or striking the predator with beak or feet.

Interestingly, corvids are also known to exhibit at least four of these behavior classes (aggregation, elevation, flyovers and scolding) in the presence of a dead conspecific, even when no predator is visible. This sort of "cacophonous aggregation" is typically referred to as a "funeral," and has been previously studied in Western scrub-jays (*Aphelocoma californica)* [Iglesias et al., 2012]. Iglesias et al. argued that individual participation in funerals serves the functions of surveillance and communication of predation risk in scrub-jays, implying that funerals and mobbing may share adaptive explanations. (The other categories of explanation for mobbing listed above are unlikely

to apply to funerals; confronting a dead conspecific would offer few deterrence or advertisement benefits.)

In the following experiments, I sought to determine whether aggregation, elevation, flyovers and scolding in Northwestern crows (*Corvus caurinus*) could also be explained by the surveillance or communication hypotheses.

Experiment 1: Collective responses to visual decoys and alarm call exemplars

In the first experiment, I presented groups of wild crows with visual stimuli (a "live" owl decoy and a "dead" crow decoy) and acoustic stimuli (recorded alarm calls) to stimulate mobbing and funeral behavior, and observed the resultant alarm behaviors of aggregation, elevation, flyovers and scolding.

I hypothesized that:

H1. A crow decoy would elicit more intense alarm behaviors than would an owl, because a "live" owl perched near the ground represents a moderate-level threat which has already been detected, while a dead crow suggests the presence of an undetected predator and demands more intensive surveillance and communication of risk.

H2. Alarm playbacks would result in more intense alarm behaviors than silence, since they are signals or cues that other crows perceive an elevated threat level.

H3. An increased initial number of crows in the area would lead to proportionately more scolding, because if the function of scolding is to communicate a threat, the indirect benefits of doing so would increase with the size of the scolding birds' audience.

However, the proportion of elevated birds and the per capita flyover rate would remain the same or decrease, because when the task of surveillance is shared by a larger number of crows, individual birds do not need to invest more time and energy in surveillance .

Methods

Subjects & Field Sites

All trials were conducted on or near the eastern coast of the Puget Sound, between Seattle and Everett, between October of 2014 and January of 2015 [Figure 1]. In each trial, a single crow or group of crows was approached. At least one crow was required to be foraging on the ground, in a park or other grassy area, wherein birds were at least 25 feet from the nearest human. 45 complete trials were conducted, with stimulus combinations randomly chosen from the nine total possibilities.

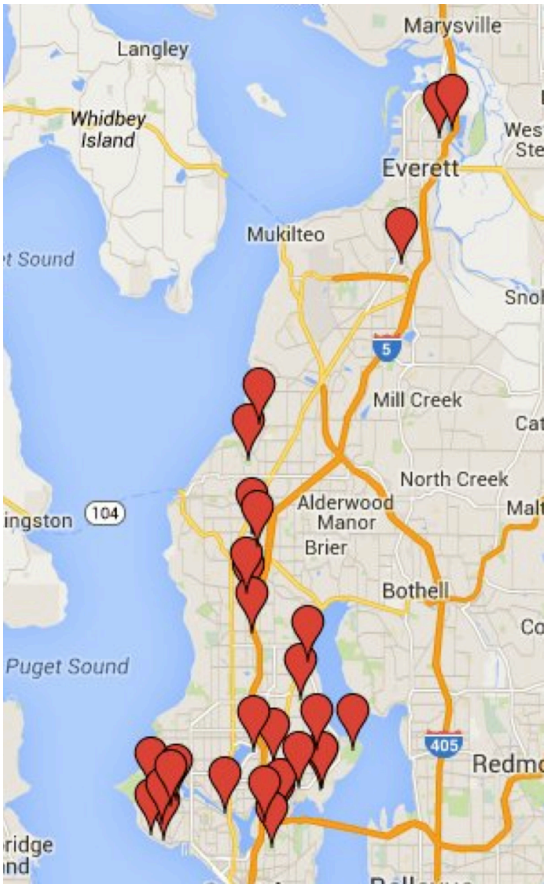


Figure 1: Locations of Experimental Trials

Equipment and Stimuli

The visual stimuli were a plastic owl decoy ("live owl") with movable head, perched on a small stump; a prone, plastic crow decoy ("dead crow") with felted surface; or nothing, in the control condition [Figure 2]. The audio stimulus was a playback of one of two recorded alarm calls, or (in the control condition) silence. Call 1 consisted of an alternating 2-caw and 8-caw bout, both recorded from an American Crow in Ithaca, NY. Call 2 consisted of a 2-caw bout recorded from a Northwestern Crow in Vancouver, BC. [Figure 3] Both calls were repeated with a randomly-varied intercall interval of 2.75 ±

0.3 seconds, and were amplitude-normalized to have identical average sound intensities. Playbacks were made from a portable loudspeaker (ION Block Rocker Bluetooth Portable Speaker System).

Figure 2: Decoys used in experiment. Only one crow decoy was used at a time and it was prone, not mounted as in the image.

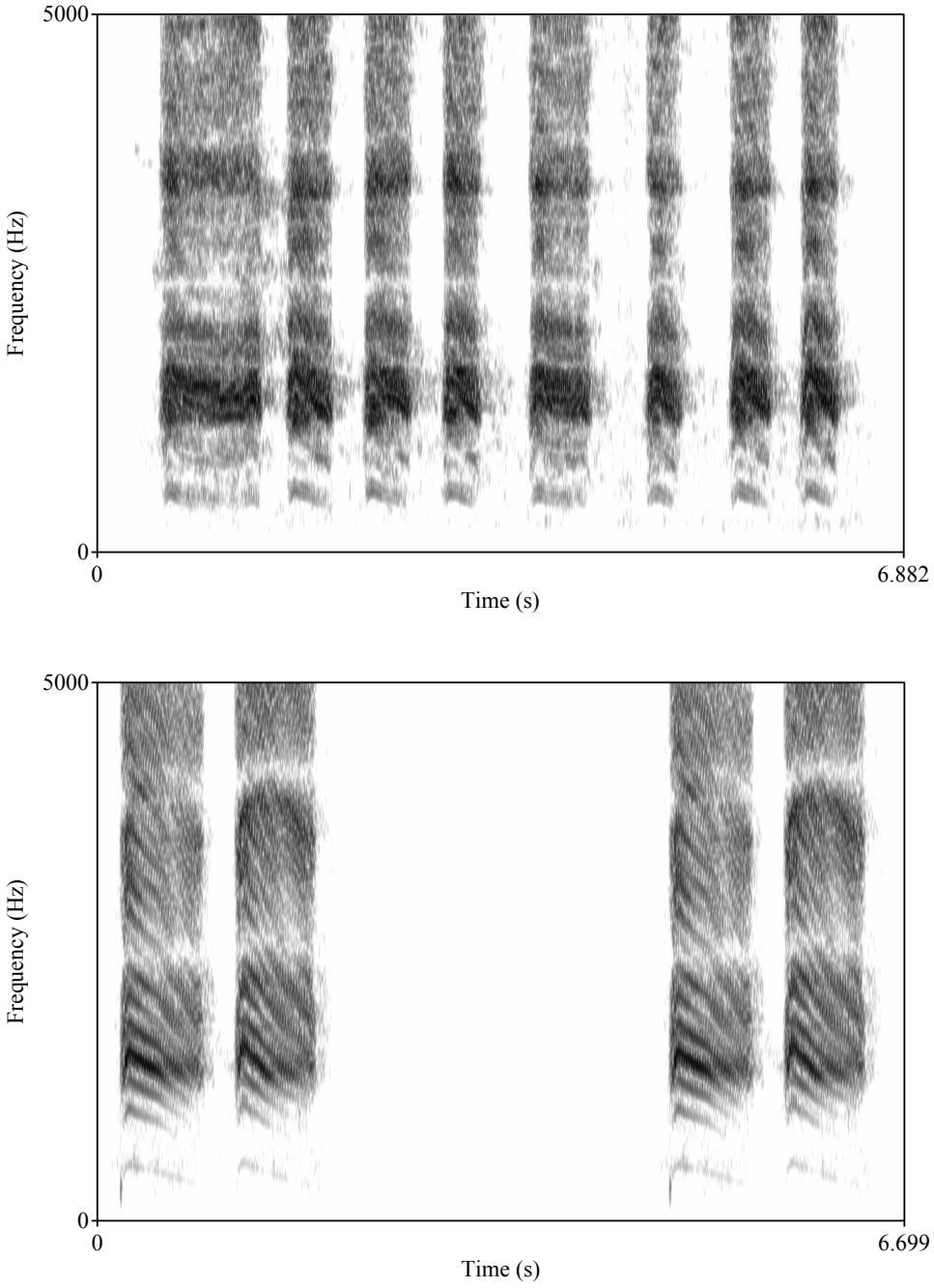


Figure 3. Alarm calls used in playback, recorded from two different individuals.

Procedure

In the experimental trials, the loudspeaker was placed in a shaded location, usually under a bush or tree. A white garbage bag was placed in a conspicuous location (raised if possible) within five meters of the loudspeaker and then removed, revealing a visual stimulus. I walked at least twenty meters away from both the visual stimulus and the loudspeaker, then sat in a secluded location or on a bench to continue observation. After thirty seconds, an audio stimulus began, which lasted for ninety seconds. Observation continued for two more minutes, and then all equipment was removed. If all crows retreated from view, I ended the observation early.

I dictated my observations in real time to a portable audio recorder (TASCAM DR-05). The data recorded included periodic censusing of visible birds (approximately every 20 seconds), and their position as on ground, perched, or airborne. Every flyover over the decoy or speaker was noted using all-occurrences sampling. Flyovers consisted of low altitude (estimated visually as 15-75 ft) flight over the target, usually involving short glides and cocking an eye toward the ground. Flyovers were defined to be unidirectional; if a crow wheeled through a complete circle over the target, this was counted as two flyovers.

Vocal behavior of the observed crows was also noted; both the incidence of calls and the type (e.g. territorial, food recruitment, scolding.) Only scolding calls were used for further analysis. Calls were classified as scolding if they were delivered while facing the direction of the decoy and had the acoustic characteristics (irregularly spaced, long, low, and harsh) associated with alarm calls in Parr (1997) and Mates et al. (2015). For each trial, I recorded the time of production of the first scolding call, and the maximum number of birds visually confirmed to be scolding simultaneously.

Analysis

Observation audio logs were manually transcribed to textgrids of events in Praat 6.0 (Boersma & Weenink, 2016), and then imported into R.

Two statistical analyses were conducted on this data: An analysis of alarm vocalizations (scolding), and an analysis of movements. For the scolding analysis, each trial was treated as a single data point (N=45). The independent variables were the log-transformed number of crows initially visible at the beginning of each trial, plus four contrasts describing the stimuli used:

1. Decoy Presence: decoy = 2, no decoy = 0
2. Decoy Type: crow = 1, owl = -1, no decoy = 0
3. Playback Use: playback = 2, no playback = 0
4. Playback Version: Call 1 = 1, Call 2 = -1, no playback = 0

Because the trial numbers were not completely balanced across conditions, these contrasts were independent but not precisely orthogonal.

The dependent variables were scolding occurrence (at least one scold during the trial = 1, no scolding = 0) and the timing of first scold (before playback or during a trial with no playback = 1, after playback had begun = 0). Peak scold intensity was regressed on the four independent variables with a Gaussian error model. Within the subset of trials where scolding occurred, Fisher's Exact Test was used to determine whether the timing of the first scold depended on the type of visual decoy presented (crow decoy vs. owl decoy or none).

For the movement analysis, the observation time within each trial was broken into 15 second blocks (N=713 including all trials). Each block was evaluated on a series of independent and dependent variables.

The dependent variables consisted of a census of visible birds, the fraction of censused birds that were observed to be elevated (perched or airborne), and the per capita number of flyovers that occurred during that time block. The first two variables were evaluated at the midpoint of each block, via linear interpolation between the nearest census times before and after that point.

Four stimulus contrasts were generated, as used in the scolding analysis, except that the Decoy and Playback Type variables were quadratically transformed to have zero mean and be orthogonal to the Decoy Presence and Playback Use variables.

Four auxiliary variables were associated with each time block:

1. An intercept (constant) variable

2. A linear trend over time, mean-centered within each trial

3. A variable indicating whether playback was occurring, and

4. A variable representing the cumulative effect of playback, which increased linearly while playback occurred and decreased more slowly (with a slope 3/5 as large) after playback ended.

Each auxiliary variable was orthogonalized against the previous ones in the above list.

The two decoy-related contrast variables were then multiplied by each of the first two auxiliary variables, and the two playback-related contrast variables were multiplied by each of the last two auxiliary variables. This resulted in eight independent variables describing each time block:

1. Decoy Presence
2. Decoy Presence, Cumulative (interaction between decoy presence and time elapsed)
3. Decoy Type (crow vs. owl)
4. Decoy Type, Cumulative (interaction between decoy type and time elapsed)
5. Playback Presence
6. Playback Presence, Cumulative
7. Playback Version (Call 1 vs Call 2)
8. Playback Version, Cumulative (interaction between playback version and cumulative effect of playback)

Playback version was included in order to assess the generalizability of any playback effects. Additionally, the mean-centered time and (log-transformed) initial count of

visible crows was included as a fixed effect, and the parent trial of each time block was treated as a random effect with a Gaussian distribution.

Each of the dependent variables was regressed on subsets of these eleven independent variables, as well as the second-level interactions between time, decoy-related and playback-related main effects. The regression models were Gaussian with log transformation for the census count, binomial with logit link for the fraction of perched/airborne birds, and Poisson for the flyover rate.

The 'glmulti' package in R was used for model selection and multimodel inference. An exhaustive search of the model spaces was not possible, given the large number ($\sim 10^7$) of distinct subsets of main effects and interactions. Accordingly, for each dependent variable, a set of three smaller searches was conducted using the options in glmulti: an exhaustive search over main effects, a search over all effects with a model size constraint of at most nine fixed terms; and a search over all effects that minimized AIC using a genetic algorithm. The models generated in these searches were pooled in a consensus set and the terms ordered by model-averaged importance (the combined Akaike weight of the models containing each term), and another exhaustive search was conducted over the fifteen most important terms. Finally, the models in the last search were added to the consensus set.

Results

All pairwise correlations between block-level IVs were below 0.1 in magnitude. Descriptive statistics for block-level DVs are shown in Table 2, broken down into the set

of blocks that coincided with or followed presentation of a auditory or visual stimulus, and the set of blocks that did not

Table 2: Descriptive Statistics for Block-Level DVs

	# of visible birds	fraction of visible birds off ground	absolute # of flyovers in 15 seconds	normalized # of flyovers in 15 seconds
Blocks with No Stimulus (N = 101)				
Maximum	20.23	1	0	0
Mean	3.08	0.49	0	0
Standard Deviation	3.42	0.45	0	0
Blocks with/after Any Stimulus (N = 612)				
Maximum	39.73	1	30	2.45
Mean	6.14	0.67	3.19	0.18
Standard Deviation	5.91	0.41	1.36	0.36

Scolding Analysis

The odds of scolding were significantly increased by decoy presence, and marginally significantly increased by playback presence [Table 5]. The odds of scolding did not

increase with the initial number of visible crows; in fact, there was a non-significant decrease in the odds of scolding as the initial number of crows increased.

Table 5: Most significant model-averaged effects: Odds of scolding occurrence.

**p < .05, .p < 0.1*

Effect	**B**	SE	Model-averaged Importance
Playback Version	0.246	0.644	0.34
Initial Crow Count	-0.330	0.545	0.45
Decoy Type	1.320	1.202	0.71
Playback Presence	2.116.	1.101	0.93
Decoy Presence	2.140*	0.889	0.97

Scolding occurred 36% of the time in trials with a playback and no decoy, 75% of the time in trials with a playback and an owl decoy, and 82% of the time in trials with a playback and a crow decoy; however, no scolding occurred in trials without a playback unless the crow decoy was presented [Table 4]. This suggests the presence of an interaction between decoy type and playback presence. However, such an interaction term could not be added to the existing logistic models for scolding occurrence without a convergence failure, due to complete separation of the response levels.

Table 4: Frequency distribution of scolding. Each cell displays the numbers of [trials with scolding / all trials] for that stimulus combination

	Crow	Owl	None
Call 1	4/4	3/5	3/6
Call 2	5/7	3/3	1/5
None	4/5	0/4	0/6

Therefore, I created a new variable, "stimulus strength," which took the value 1 for trials with a crow decoy or with both an owl decoy and a playback; the value 1/2 for trials with a playback but no decoy; and the value 0 otherwise. In a logistic binomial model, this variable alone predicted the log-odds of scolding more reliably (residual deviance 40.08) than did all four of the original main effects together (residual deviance 42.93). Moreover, the AIC of the model using stimulus strength was significantly lower than that of the best main-effects model ($\Delta AIC = 8.85$). Finally, a glmulti search over all models with terms drawn from the four main effects plus stimulus strength found that only the effect of stimulus strength was significantly different from zero.

The crow decoy was also particularly effective at provoking scolding before playback had begun. Of the 13 trials in which a crow decoy was presented and scolding occurred, the scolding began before playback in seven trials. On the other hand, of the ten trials in

which an owl decoy or no decoy was presented and scolding occurred, the scolding never began before playback. This difference in proportion was significant, $p = .0075$ (Fisher's Exact Test).

Movement Analysis

Four terms were found to significantly affect the number of birds at each census. The number of birds tended to increase over time, and tended to be larger if the initial number of birds was larger; however, the rate of increase was somewhat reduced if the initial number of birds was increased. The number of birds was also increased if a crow decoy was presented, as opposed to an owl [Table 6].

Table 6: Most significant model-averaged effects: (log) census number

**** $p < .001$*

Effect	**B**	SE	Model-averaged Importance
Decoy Type	-0.002	0.021	0.01
Decoy Presence, Cumulative	0.002	0.008	0.04
Initial Crow Count X Time	-0.084***	0.026	0.95
Initial Crow Count	0.734***	0.125	1.00
Time	0.152***	0.013	1.00
Decoy Type, Cumulative	0.135***	0.013	1.00

Four terms, all main effects, were found to significantly affect the proportion of birds that were perched or airborne at each census. This proportion tended to increase over time,

and to do so more rapidly if a decoy or playback was presented. Playback also tended to produce an immediate increase in the proportion of birds off the ground. [Table 7]

Table 7: Most significant model-averaged effects: proportion of birds that were elevated (perched or airborne).

****p < .001*

Effect	**B**	SE	Model-averaged Importance
Decoy Type	0.000	0.019	0.00
Decoy Type, Cumulative	0.003	0.056	0.00
Playback Presence, Cumulative	0.729***	0.179	1.00
Playback Presence	0.627***	0.150	1.00
Decoy Presence, Cumulative	0.543***	0.140	1.00
Time	0.610***	0.143	1.00

Two main effects were found to significantly affect the per capita flyover rate. This rate tended to decrease over time, and to increase during playback [Table 8].

Table 8: Most significant model-averaged effects: per capita flyover rate.

****p < .001*

Effect	**B**	SE	Model-averaged Importance
Decoy Presence	-0.20	0.40	0.22
Playback Presence X Decoy Presence	0.07	0.11	0.33

Playback Type	0.07	0.09	0.44
Decoy Presence, Cumulative	0.08	0.09	0.63
Playback Presence	0.31***	0.07	1.00
Time	-0.36***	0.06	1.00

For the significant terms reported above, raw effect sizes on the various dependent variables are summarized in Table 9.

Table 9: Significant Raw Effects on Block-Level Dependent Variables

		Expected Relative Increase in:		
		Number of birds observed in area	Odds of a given bird being elevated at census	Per capita flyover rate
	2x the initial number of birds	+66.3%		
Time Effects	+15 seconds elapsed	+3.27%	+13.73%	-7.3%
	+15 seconds elapsed, given 2x the initial number of birds	-1.22%		

Decoy Effects	+15 seconds of exposure to decoy		+27.43%	
	+15 seconds of exposure to crow decoy (vs. owl)	+7.49%		
Playback Effects	Playback currently occurring		+376.3%	+115.3%
	+15 seconds of exposure to playback		+53.3%	

Discussion

Hypothesis H1 (that a crow decoy would elicit more intense alarm behaviors than an owl decoy) was partially supported; decoy type had a significant effect on aggregation and scolding occurrence, but not on elevation odds or the flyover rate. The number of observed crows increased particularly rapidly in the presence of a crow decoy, but actually decreased slightly over time when an owl decoy was used. The odds of scolding occurring during a trial did not vary significantly with decoy type, but it was only in the presence of a crow decoy that crows began to scold *without* first hearing a playback.

Hypothesis H2 (that alarm playbacks would elicit more intense alarm behaviors than silence) was supported with regard to elevation, flyover rate, and (marginally) scolding

occurrence, but not with regard to aggregation. Most crows on the ground took off as soon as playback began, and almost all crows were elevated by the end of playback. The per capita rate at which crows overflew the stimulus location more than doubled when a playback was occurring.

Hypothesis H3 (that an increased initial number of nearby crows would elicit more scolding but not more elevation or flyovers) was supported with regard to elevation and flyover rate, but not with regard to scolding. Indeed, the odds of scolding during a trial actually decreased slightly as the initial number of crows in that trial increased, despite there being both a larger audience and a larger number of potential scolders.

The lack of a playback effect on aggregation is consistent with previous studies on American crows by Frings et al. (1958) and Parr (1997). Nonetheless, given the distances over which crows travel to join mobs, and the fact that trials in this study frequently attracted several times as many birds as were initially visible, it would be surprising if alarm calls had no impact whatsoever on aggregation. The explanation may simply be that alarm calls do not induce aggregation over a long distance unless several birds are calling, which was not the case in my playbacks or those in the other referenced studies. In addition, playbacks may not induce aggregation as effectively as live crow calls, since they are not accompanied by visual stimuli such as flyovers and a distinct calling posture. Either of these possibilities would explain why there was an effect of decoy type on aggregation, even though the decoys were placed on the ground where

distant birds could not see them; the crow decoy was more effective at eliciting scolding from several birds, which would constitute a greater stimulus for aggregation than the playback did.

The lack of an effect of decoy type on elevation odds may be explained by the strength of the playback and decoy presence effects. Most foraging crows flew up as soon as they heard an alarm call or registered a decoy of any type; leaving the ground is apparently the minimal response to any potential threat.

Flyover rates were unaffected by decoy type or presence, possibly because flyovers are used to search for threats that are not already visible. In the absence of a playback, crows left the ground, perched in elevated positions and visibly peered at the decoys, but they rarely performed a flyover, even when scolding. Since playback had no impact on the total number of crows observed, it appears playbacks cause more flyovers by increasing nearby birds' tendency to perform them, rather than by attracting more birds.

Trials conducted on larger groups of birds were not more likely to result in scolding, which suggests that communication is not the sole function of that behavior. Scolding is also unlikely to have a deterrent or advertisement function if it is more readily induced by a dead conspecific, or a disembodied playback, than an actual predator such as an owl. My suggestion is that scolding serves a surveillance function. Because crows increase their flyover rate in the presence of alarm calls, and become more likely to scold potential

threats such as owls or human researchers, scolding is an effective means of inducing nearby conspecifics to perform surveillance on the scolder's behalf. If so, it would be adaptive for crows to scold more when in smaller groups, since they face a greater threat from ambush predators at that time. That said, scolding clearly does function as communication as well, since it is also frequently directed at visible predators. I was frequently pursued back to my car by scolding crows after I retrieved the crow decoy, even though I had placed it in an opaque white garbage bag as soon as I picked it up. The crows continued to scold until I had entered my car. Evidently, the crows were capable of changing the target of their scolding from the crow decoy to myself, a potential predator who might have killed it. (When I retrieved the owl decoy, crows generally ceased scolding immediately.)

Why would crows, in general, be more responsive to a decoy of a dead crow than to a decoy of a live predator? One possibility, of course, is that the crow decoy was simply more convincing than the owl decoy. Humans would not find it to be such; the owl decoy had a movable head, and was more realistically shaped and colored than the crow decoy (which was slightly larger than a live crow, and had a felted surface and no legs). Nonetheless, crows may have a more specific search image for predatory owls than for dead conspecifics. The owl's position on or near the ground may also have been unexpected. Swift & Marzluff (2015) also presented wild crows with "dead" crow and "live" predator stimuli, and used a more realistic predator stimulus—a taxidermy-mounted red-tailed hawk perched on a tree branch. They did not find a significant difference in alarm behavior between crows presented with the dead crow and those

presented with the hawk mount. While this suggests that the owl decoy used in my experiment was insufficiently realistic, it also underscores that, even when more realistic predator stimuli are used, crows do not find them more alarming than dead conspecifics.

I suggest that this is because a predator is inherently a more ambiguous threat stimulus than a dead crow. Crows frequently forage near predators, such as humans, bald eagles, and coyotes. Indeed, as kleptoparasites, crows will sometimes deliberately interact with predators (e.g., by tail-pulling) in order to gain access to their kills. Attacking or scolding a predator which does not currently pose a threat can therefore be costly, in terms of lost foraging opportunities. It may be adaptive to wait for additional threat cues from the predator itself, or alarm signals from other crows that have perhaps watched the predator more closely, before beginning scolding and mobbing activities. (In my trials, the playbacks provided such signals.)

On the other hand, there is less opportunity cost of scolding in the presence of a dead conspecific. Furthermore, a dead conspecific may imply the presence of an even greater threat than even a visible, actively hunting predator, since whatever killed it (e.g. disease, poison or a predator that is currently undetected) is not something that the watching crow can easily locate and avoid. The cost-benefit analysis in this case is skewed toward scolding under any circumstance.

The alarm behavior examined in this experiment can plausibly be explained as having the functions of surveillance and communication, which supports the hypothesis that corvid

"funerals" are not distinct from anti-predator mobs. Funeral behavior may simply be the particular type of mobbing which occurs when a dead conspecific is observed without a visible predator in the area, in which case surveillance and communication are the top priorities for other birds. This does not mean that some mobbing behavior–in particular, dive-bombing and other physical attacks–does not serve the function of predator deterrence or advertisement instead. But the "milder" form of mobbing observed in this study is most likely employed by crows to gather and share information on predators with one another. Perhaps little more than this is needed to avoid predation in most cases.

Experiment 2: Individual responses to food and alarm calls with varying production rates and caller numbers

The results of Experiment 1 motivated several new research questions. Given that the observed mobbing behaviors appeared to function in anti-predator surveillance, this raised the question of whether they were used in searching for and monitoring other ecologically relevant targets as well, such as food sources. In addition, having found the unexpected result that alarm-related stimuli did not induce more scolding in larger groups of birds, I sought to confirm this using a more sensitive metric than whether scolding did or did not occur within the entire subject group, over the course of an entire trial. Finally, I wished to further explore how mobbing behaviors in response to visual stimuli were mediated by scolding.

I explored these questions in Experiment 2, in which I presented wild crows with a wider variety of acoustic stimuli. Each 150-second trial was focused on an individual crow so as to control for duration of exposure to the stimuli and record their behavior on a finer timescale. The stimuli used were recorded alarm and feeding calls, produced by one or two individuals, and digitally manipulated to occur at high or low rates. I recorded the elevation pattern, perch changes, flyovers and scolds of the focal crow.

Based on the supposition that the primary function of these alarm behaviors was anti-predator surveillance, I formed the following hypotheses:

H1. Alarm call playbacks would elicit more frequent scolding, elevation, flyovers and perch changes in focal crows than would food call playbacks. This is because alarm calls

indicate the presence of a threat that requires surveillance. Food calls may also motivate crows to search for a target (a food source), but they do not specifically label that target as a threat.

H2. Focal crows would not scold more in the presence of a larger number of conspecifics. This is because the individual need for surveillance remains the same or decreases with more vigilant conspecifics in the area.

In addition, the results of Experiment 1 suggested that the ability of a threatening visual stimulus to induce aggregation might be mediated by the scolding of nearby crows, which would communicate the threat level to conspecifics outside visual range. I therefore hypothesized that:

H3. Alarm call playbacks at higher rates, and/or with more callers, would elicit more intense alarm behaviors in focal crows.

Methods

Subjects & Field Sites

All trials were conducted on or near the eastern coast of the Puget Sound, in Kent and Snohomish Counties, in May through July of 2015. [Figure 3] In each trial, a single crow was approached. It was required to be foraging on the ground at the start of observation, and at least 30 feet from the nearest human. 46 complete trials were conducted, with stimulus combinations randomly chosen from the nine total possibilities. [Table 10]

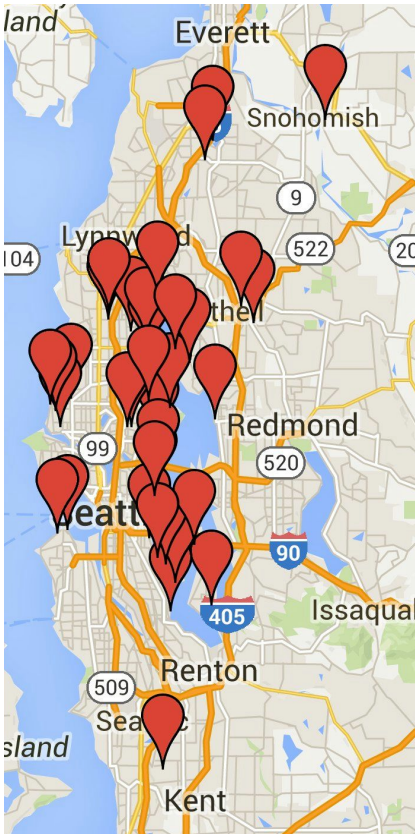


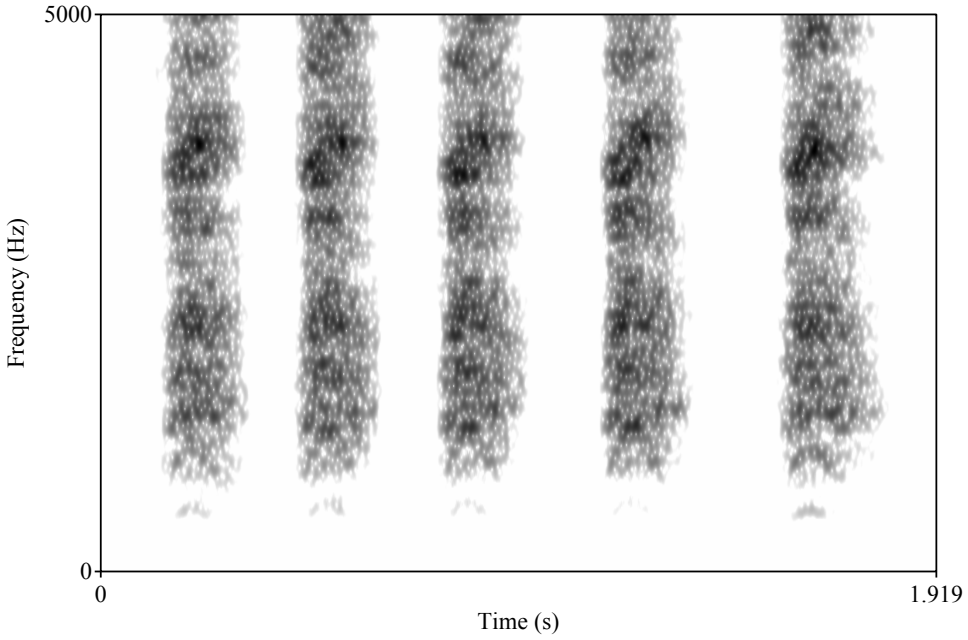
Figure 3: Locations of Experimental Trials

Table 10: Frequency table of stimulus combinations for experimental trials

Trial Stimuli Playback Call Type & Rate	Caller Identity 1	2	Both	Grand Total
Alarm Call	9	7	8	24
High Rate	3	3	4	10
Low Rate	6	4	4	14
Food Call	8	6	8	22
High Rate	2	3	4	9
Low Rate	6	3	4	13
Grand Total	**17**	**13**	**16**	**46**

Equipment and Stimuli

The visual stimulus in every trial was the plastic owl decoy used in Experiment 1 [Figure 2]. The audio stimulus was a playback of one of twelve digitally manipulated call sequences. To create the stimulus call sequences, food recruitment calls were recorded on two occasions from individual wild crows in Seattle, WA after I tossed peanuts beneath their perch locations. Alarm calls were recorded similarly, after I approached foraging wild crows on foot while staring at them and holding up the "dead" crow decoy used in Experiment 1, head downwards [Figure 4]. The caw bouts within each recording were then repeated in random order, at rates approximating their original production.



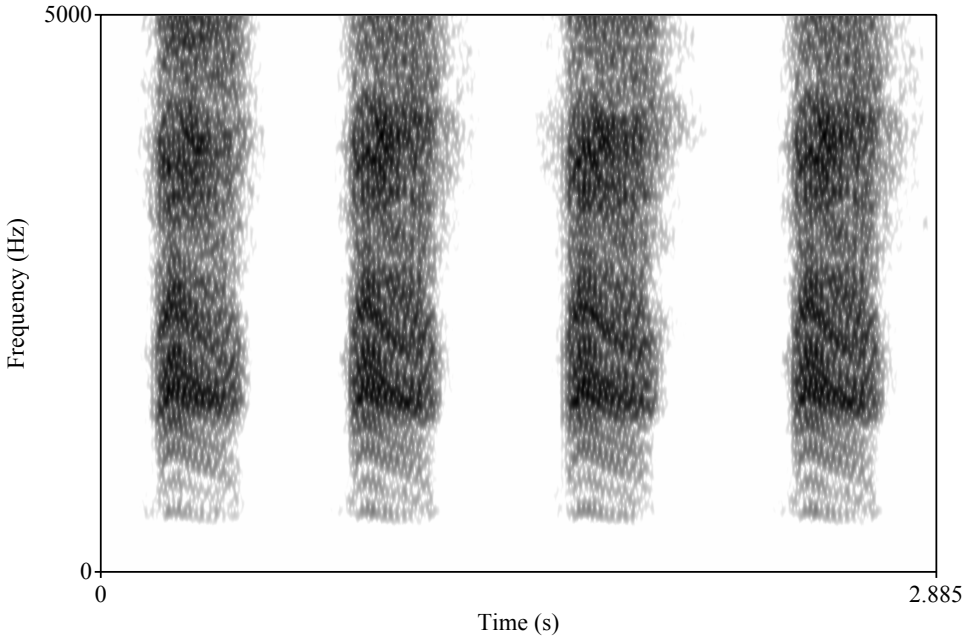


Figure 4. Example of crow caws recorded in the field and used for playback. Above: Food recruitment calls. Below: Alarm calls.

Stimulus recordings differed on the following variables:

Call type: Alarm versus Food (recruitment)

Number of callers: Two (sequences of calls from two individuals were superimposed, with calls occasionally overlapping) vs. one (calls were used from only one of the two recorded individuals for that call type)

Call rate: High (calls made up 30% of playback time) vs. low (calls made up 15% of playback time)

All calls were amplitude-normalized to have identical average sound intensities. Playbacks were made from the same portable loudspeaker (ION Block Rocker Bluetooth Portable Speaker System) as in Experiment 1.

Procedure

Experimental procedure was similar to Experiment 1, although I chose to make the trials slightly shorter because they were focused on individual crows, who tended to return to foraging while other birds in the area were still exhibiting alarm behaviors. The loudspeaker was placed in a shaded location, usually under a bush or tree. A white garbage bag was placed in a conspicuous location (raised if possible) within five meters of the loudspeaker and then removed, revealing the owl decoy. I walked at least 20 meters away from both the decoy and the loudspeaker, then sat in a secluded location or on a bench to continue observation. After 30 seconds, an audio stimulus began, which lasted for 90 seconds. Observation continued for 45 more seconds, and then all equipment was removed. If the focal crow flew too far away to be observed, I ended the observation early; this only occurred in seven trials, and the observation was never ended more than 30 seconds early.

Every scolding call, change in position (between ground, air, and an elevated perch) and flyover of the focal crow was recorded via audio dictation. A census of visible nearby birds was taken at the beginning of the trial, and approximately every 30 seconds thereafter.

Analysis

Observation logs were transcribed into textgrids of events in Praat, and then imported into R.

As in the movement analysis for Experiment 1, observation time within each trial was broken into 15-second blocks. (N=457 including all trials). Each block was evaluated on a series of independent and dependent variables.

The dependent variables consisted of the proportion of time within that block for which the focal bird was observed to be elevated (perched or airborne), and the numbers of scolds, flyovers and perch changes performed by the focal bird during that time block. A perch change was defined as any transition from a ground or airborne position to a perched position; thus, any flyover which ended with the bird perching also counted as a perch change.

Two time-related independent variables were associated with each block, indicating whether it occurred before or after the beginning of playback, and whether it occurred during or after playback. These time variables were also crossed with the three stimulus variables described above (call type, call rate and caller number) to yield six more independent variables, and the (log-transformed) initial count of nearby visible crows was also included. The parent trial of each time block was treated as a random effect with a Gaussian distribution.

Each of the dependent variables was regressed on subsets of these nine independent variables. The regression models were binomial with logit link for the proportion of time spent, and Poisson with log link for the scolding, flyover, and perch change rates.

The 'glmulti' package in R was used for model selection and multimodel inference. An exhaustive search was made of the model space, which contained 512 models for each dependent variable. Regression coefficients were averaged across all models, using Akaike weights; the weighted importance of each effect across the model set was also computed.

Results

All pairwise correlations between block-level IVs were below 0.1 in magnitude. Descriptive statistics for block-level DVs are shown in Table 9, broken down into the set of blocks that coincided with presentation of a playback, and the set of blocks that did not.

Table 9: Descriptive Statistics for Block-Level DVs

	fraction of time spent by focal bird off ground	# of flyovers per 15 seconds	# of perch changes per 15 seconds	# of scolds per 15 seconds
Blocks without Playback (N = 183)				
Maximum	1	2	2	27
Mean	0.41	0.07	0.14	1.85
Std. Dev.	0.49	0.27	0.39	4.85
Blocks with Playback (N = 274)				
Maximum	1	3	4	25

Mean	0.64	0.25	0.39	2.55
Std. Dev.	0.45	0.54	0.67	5.03

Three variables significantly affected the scolding rate of focal crows [Table 10]. Birds scolded at a higher rate after playback had begun, and at a lower rate after it ended. They also scolded at a higher rate during playback of alarm calls rather than food calls. Playback call rate did not significantly affect the scolding rate of focal crows, although models with call rate as a variable constituted 89% of the Akaike weight in the full model set. This apparent paradox reflects the fact that although the models with call rate had better fits, the standard errors of their slope coefficients for call rate were very high. Silva and Tenreyo (2011) note that standard errors in Poisson models can be very large in the case of complete or quasi-complete separation (when the response variable is zero or near-zero for particular combinations of the independent variables). In these observations, almost no focal birds scolded before playback began: there was an average of .06 scolds per 15 seconds beforehand, as opposed to 2.86 scolds per 15 seconds afterwards. Furthermore, in trials where food calls were played back at a low rate, focal birds gave an average of .07 scolds per 15 seconds, while their scolding rates for other combinations of call type and rate were between 1.94 and 4.11 scolds per 15 seconds. When the interaction between call type and call rate was included in a subsequent analysis, the model-averaged effect of call rate did become significant ($B = 2.70$, $SE=0.80$, $p < .001$). However, it is unclear how the significance level should be corrected since this was a post-hoc analysis and the interaction term was not originally included in the regression models.

Table 10: Most significant model-averaged effects: number of scolds given per second

***p < .01, ***p < .001*

Effect	**B**	SE	Model-averaged Importance
Playback begun: Caller number	0.46	1.76	0.10
Playback begun: Call rate	7.32	6.19	0.89
Playback still occurring: Call type	0.36***	0.08	1.00
Playback still occurring	-0.25***	0.04	1.00
Playback begun	6.21**	2.48	1.00

Two variables significantly affected the proportion of time that focal crows spent off the ground. Crows were elevated more often after playback had begun, and less often after it ended [Table 11].

Table 11: Most significant model-averaged effects: proportion of time that bird was elevated (perched or airborne).

*.p < .1, ***p < .001*

Effect	**B**	SE	Model-averaged Importance
Playback begun: Call rate	0.29	0.90	0.33
Playback still occurring: Caller	0.38	0.58	0.36

number			
Playback begun: Caller number	1.19	1.68	0.51
Playback still occurring	-0.90***	0.25	1.00
Playback begun	6.59***	0.83	1.00

Three variables significantly affected the flyover rate [Table 12]. Focal crows performed flyovers more frequently after playback had begun and less frequently after it ended, and more frequently when alarm calls were played back rather than food calls.

Table 12: Most significant model-averaged effects: flyover rate.

**p < .05, **p < .01*

Effect	**B**	SE	Model-averaged Importance
Playback begun: Caller number	0.14	0.74	0.19
Playback begun: Call rate	1.25	1.61	0.65
Playback begun: Call type	3.90*	2.10	0.92
Playback still occurring	0.43**	0.16	0.99
Playback begun	3.88**	1.33	1.00

Two variables significantly affected the perch change rate [Table 13]. Focal crows changed their perch more frequently after playback had begun, and less frequently after the playback had ended.

Table 13: Most significant model-averaged effects: perch change rate.

**p < .05, ***p < .001*

Effect	**B**	SE	Model-averaged Importance
Playback begun: Caller number	0.31	0.72	0.33
Playback begun: Call rate	0.24	0.62	0.36
Playback begun: Call type	-0.49	0.82	0.46
Playback still occurring	0.28*	0.13	0.95
Playback begun	1.91***	0.49	1.00

Discussion

Hypothesis H1 (that alarm call playbacks would elicit more alarm behaviors than would food call playbacks) was confirmed for scolds and flyovers, but not for elevation. This further supports the idea that scolding is used for predator surveillance; if its only function were the communication of a threat, then there would be even more reason to scold a decoy predator when food calls were encouraging conspecifics to forage nearby. The fact that flyovers occurred more frequently in response to alarm calls but perch changes did not, suggests that perch changes are not mobbing behaviors in particular but have a broader surveillance function outside the context of predatory threats. Crows may prefer perch changes within a tree to flyovers when searching for or monitoring food sources, as the former activity is less visible to conspecific competitors. Corvids are

known to pilfer one another's food caches, and to cache in locations that are less likely to be observed by a competitor; common ravens also prioritize food sources that a conspecific has located but cannot yet access (Bugnyar & Heinrich, 2005; Dally, Emery & Clayton, 2005). It would therefore be advantageous for a crow to locate and monitor food sources without advertising its activities to conspecifics.

Hypothesis H2 (that an increased initial number of nearby conspecifics would not elicit more scolding) was confirmed; the number of nearby birds did not significantly affect the rates of any measured alarm behavior. This result runs counter to the communication and advertisement hypotheses, which would predict more scolding in the presence of more conspecifics. Thus, this provides further evidence that scolding serves a surveillance function.

Hypothesis H3 (that alarm call playbacks with higher call rates or caller numbers would elicit more intense alarm behaviors) was not confirmed; call rate and caller number did not significantly affect the rates of any measured alarm behavior. However, as discussed above, this may reflect separation issues in the Poisson models, rather than a true lack of effect. A direct test of the effects of playback characteristics on aggregation was not possible in this experiment, because the numbers of nearby birds could not be censused as frequently as in Experiment 1.

The results of this second experiment largely agree with the first. On an individual as well as a collective level, those mobbing behaviors in Northwestern crows which fall short of actual physical aggression are well explained by the communication and surveillance hypotheses, and less so by the advertisement and deterrence hypotheses.

There is even evidence that scolding functions primarily to recruit conspecifics for surveillance of potential threats. However, the precise relationship between the characteristics of a scolding call and the alarm behaviors it induces remains unclear. Playback experiments with a larger sample size would allow for a wider range of call rate and caller numbers, and a more powerful exploration of the interaction between these properties and the various types of long-ranged calls that Northwestern crows produce.

Bibliography

Boersma, Paul & Weenink, David (2016). Praat: doing phonetics by computer [computer program]. Version 6.0.14, retrieved 11 February 2016 from http://www.praat.org/

Dally, J. M., Emery, N. J., & Clayton, N. S. (2005). Cache protection strategies by western scrub-jays, Aphelocoma californica: Implications for social cognition. *Animal Behaviour*, *70*(6), 1251–1263. http://doi.org/10.1016/j.anbehav.2005.02.009

Dugatkin, L. and Godin, J. (1992). Prey approaching predators: a cost- benefit perspective.

Iglesias, T. L., McElreath, R., and Patricelli, G. L. (2012). Western scrub- jay funerals: Cacophonous aggregations in response to dead conspecifics. *Animal Behaviour,* 84(5):1103–1111.

Mates, E. A., Tarter, R. R., Ha, J. C., Clark, A. B., and McGowan, K. J. (2015). Acoustic profiling in a complexly social species, the American crow: caws encode information on caller sex, identity and behavioural context. *Bioacoustic*s, 24(1):63–80.

Santos Silva, J. M. C., & Tenreyro, S. (2011). Poisson: Some convergence issues. *Stata Journal*, *11*(2), 207–212.

Shields, W. M. (1984). Barn swallow mobbing: Self-defence, collateral kin defence, group defence, or parental care? *Animal Behaviour*, 32(1):132– 148.

Sordahl, T. A. (1990). The Risks of Avian Mobbing and Distraction Behavior: An Anecdotal Review. *The Wilson Bulletin*, 102(2):349–

Swift, K. N., & Marzluff, J. M. (2015). Wild American crows gather around their dead to learn about danger. *Animal Behaviour*, *109*, 187-197.

APPENDIX A

Auxiliary Proofs

1. Proof that any M complex exponentials with distinct signed frequencies bounded by the Nyquist frequency form a linearly independent set on the domain [0,1,...M-1]

The claim is trivially true for M = 1, since any nonzero complex exponential is nonzero everywhere, including at $n = 0$.

Suppose the claim is true for $M = m - 1$, but false for $M = m$.

Then there exist distinct real values $r_1 \dots r_m$ with $-1/2 < r_i \leq 1/2$, and complex values $c_1 \dots c_m$ that are not all zero, such that $f[n] = \sum_{k=1}^{m} c_k e^{2i\pi r_k n} = 0$ for $n \in \{0, 1, ...m-1\}$. Without loss of generality, we can assume that all r_i are nonzero. (If this was not the case, we could simply add a small constant d to all r_i; this would be equivalent to multiplying the above exponential sum by the complex exponential $e^{2i\pi dn}$, which would not affect the number or location of its zeros.)

In that case, if we translate $f[n]$ by -1 along the n-axis, we will leave all but one of its zeros unchanged:

$f[n+1] = \sum_{k=1}^{m} (c_k e^{2i\pi r_k}) e^{2i\pi r_k n} = \sum_{k=1}^{m} c_k e^{2i\pi r_k (n+1)} = 0$ for $n \in \{-1, 0, \cdots m-2\}$.

And by subtracting a constant multiple of $f[n]$ from its translate, we can eliminate the exponential term with the highest frequency, while preserving the $n-1$ zeros which are shared by both $f[n]$ and $f[n+1]$:

$g[n] = f[n+1] - e^{2i\pi r_k} f[n] = \sum_{k=1}^{m-1} c_k (e^{2i\pi r_k} - e^{2i\pi r_m}) e^{2i\pi r_k n} = \sum_{k=1}^{m} c_k (e^{2i\pi r_k} - e^{2i\pi r_m}) e^{2i\pi r_k n} = 0$ for $n \in \{0, \cdots m-2\}$.

The terms of $g[n]$ would constitute $m-1$ complex exponentials that are not linearly independent on the domain [0,1...m-2]. But this contradicts the assumption that the original claim is true for $M = m-1$. Thus, if the claim is true for $M = m-1$, it must also be true for $M = m$.

By induction, the original claim is true for all $M \geq 1$.

2. Error bounds for short-term approximations to the autocorrelation

Let the period-p signal $\vec{s}$ be a real p-element vector, w be the infinite sequence of window values (assumed to be real and nonnegative). Assume further that w is symmetric about n=1/2 ($w[1-n] = w[n]$), that it has support of length L an even integer (w vanishes outside $[1-L, L]$), that w is nondecreasing on [1-L,0] and therefore nonincreasing on [1,L], and lastly that $\ln w$ is concave down. This last condition may seem restrictive, but is satisfied by most common window functions, e.g. rectangular, triangular, Hamming, Hann, Bartlett-Hann, Blackman, and truncated Gaussian. Let $\boldsymbol{P}$ be the permutation matrix which advances the elements of $\vec{s}$ by 1: $(\boldsymbol{P}\vec{s})_k = \vec{s}_{k+1}$. Then

$$\boldsymbol{P} = \begin{pmatrix} 0 & 1 & 0 & \dots & 0 \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \vdots & \vdots & \ddots & \ddots & 0 \\ 0 & 0 & \dots & \ddots & 1 \\ 1 & 0 & \dots & \dots & 0 \end{pmatrix}$$

We can write the true autocorrelation formula for a delay n as $a_s(n) = \frac{\left(\vec{s}^{\,*}\boldsymbol{P}^n\vec{s}\right)}{\vec{s}^{\,*}\,\vec{s}}$.

The short-term approximation introduced in Chapter 3 is

$$R_M(n) = \frac{2 \sum_{t=-\infty}^{\infty} \left[w[t]\overline{s}_{t \bmod p}\right] \left[w[t+n]s_{t+n \bmod p}\right]}{\sum_{t=-\infty}^{\infty} w[t]w[t+n]|s_{t \bmod p}|^2 + \sum_{t=-\infty}^{\infty} w[t]w[t+n]|s_{t+n \bmod p}|^2}$$

.

(Note that the modulo above is defined such that $p \bmod p = p$ rather than 0.)

Let us define $\vec{\omega}(n)$ as the p-element vector with $\omega(n)_k = \sum_{t=-\infty}^{\infty} w[k+n+tp]w[k+tp]$ and $\boldsymbol{W}_n$ as the p x p matrix with $(W_n)_{k,k} = \omega(n)_k$ and all other entries zero. Note that $\omega(-n)_k = \omega(n)_{k-n \bmod p}$ and $\boldsymbol{W}_{-n} = \boldsymbol{P}^{-n}\boldsymbol{W}_n\boldsymbol{P}^n$.

we can write the approximation as

$$R_M(n) = \frac{2\,\vec{s}^{\,*}\boldsymbol{W}_n\boldsymbol{P}^n\vec{s}}{\vec{s}^{\,*}\boldsymbol{W}_n\vec{s} + \left(\boldsymbol{P}^n\vec{s}\right)^*\boldsymbol{W}_n\boldsymbol{P}^n\vec{s}}$$

$$= m(n) = \frac{2\,\vec{s}^{\,*}\boldsymbol{W}_n\boldsymbol{P}^n\vec{s}}{\vec{s}^{\,*}\boldsymbol{W}_n\vec{s} + \vec{s}^{\,*}\boldsymbol{W}_{-n}\vec{s}}$$

Then the error term is $E_M(n) = R_M(n) - a_s(n)$

$$= \frac{2\left(\vec{s}^{\,*}\boldsymbol{W}_n\boldsymbol{P}^n\vec{s}\right)\left(\vec{s}^{\,*}\vec{s}\right) - \left(\vec{s}^{\,*}\boldsymbol{P}^n\vec{s}\right)\left(\vec{s}^{\,*}\boldsymbol{W}_n\vec{s} + \vec{s}^{\,*}\boldsymbol{W}_{-n}\vec{s}\right)}{\left(\vec{s}^{\,*}\boldsymbol{W}_n\vec{s} + \vec{s}^{\,*}\boldsymbol{W}_{-n}\vec{s}\right)\left(\vec{s}^{\,*}\vec{s}\right)} = \frac{N(n)}{D(n)}.$$

$N(n)$ can be rewritten as

$$\left(\vec{s}^{\,*}\boldsymbol{W}_n\boldsymbol{P}^n\vec{s}\right)\left(\vec{s}^{\,*}\vec{s}\right) - \left(\vec{s}^{\,*}\boldsymbol{P}^n\vec{s}\right)\left(\vec{s}^{\,*}\boldsymbol{W}_n\vec{s}\right) \quad + \quad \left(\vec{s}^{\,*}\boldsymbol{W}_n\boldsymbol{P}^n\vec{s}\right)\left(\vec{s}^{\,*}\vec{s}\right) - \left(\vec{s}^{\,*}\boldsymbol{P}^n\vec{s}\right)\left(\vec{s}^{\,*}\boldsymbol{W}_{-n}\vec{s}\right)$$

$$= \left(\vec{s}^{\,*}\boldsymbol{W}_n\boldsymbol{P}^n\vec{s}\right)\left(\vec{s}^{\,*}\vec{s}\right) - \left(\vec{s}^{\,*}\boldsymbol{P}^n\vec{s}\right)\left(\vec{s}^{\,*}\boldsymbol{W}_n\vec{s}\right) \quad + \quad \left(\vec{s}^{\,*}\boldsymbol{W}_n\boldsymbol{P}^n\vec{s}\right)\left(\vec{s}^{\,*}\vec{s}\right) - \left(\vec{s}^{\,*}\boldsymbol{P}^n\vec{s}\right)\left(\vec{s}^{\,*}\boldsymbol{P}^{-n}\boldsymbol{W}_n\boldsymbol{P}^n\vec{s}\right)$$

Define $\tilde{\boldsymbol{W}}_n$ by $(\tilde{W}_n)_{k,k} = \omega(n)_k - \dfrac{a_w(n)}{p}$ and all other entries 0. Then

$$N(n) = \left(\vec{s}^{\,*}\tilde{\boldsymbol{W}}_n\boldsymbol{P}^n\vec{s} + \frac{a_w(n)a_s(n)}{p}\right)\left(\vec{s}^{\,*}\vec{s}\right) - \left(\vec{s}^{\,*}\boldsymbol{P}^n\vec{s}\right)\left(\vec{s}^{\,*}\tilde{\boldsymbol{W}}_n\vec{s} + \frac{a_w(n)\|\vec{s}\|^2}{p}\right) \quad +$$

$$\left(\vec{s}^{\,*}\tilde{\boldsymbol{W}}_n\boldsymbol{P}^n\vec{s} + \frac{a_w(n)a_s(n)}{p}\right)\left(\vec{s}^{\,*}\vec{s}\right) - \left(\vec{s}^{\,*}\boldsymbol{P}^n\vec{s}\right)\left(\vec{s}^{\,*}\boldsymbol{P}^{-n}\tilde{\boldsymbol{W}}_n\boldsymbol{P}^n\vec{s} + \frac{a_w(n)\|\vec{s}\|^2}{p}\right)$$

$$= \left(\vec{s}^{\,*}\tilde{\boldsymbol{W}}_n\boldsymbol{P}^n\vec{s}\right)\left(\vec{s}^{\,*}\vec{s}\right) - \left(\vec{s}^{\,*}\boldsymbol{P}^n\vec{s}\right)\left(\vec{s}^{\,*}\tilde{\boldsymbol{W}}_n\vec{s}\right) \quad + \quad \left(\vec{s}^{\,*}\tilde{\boldsymbol{W}}_n\boldsymbol{P}^n\vec{s}\right)\left(\vec{s}^{\,*}\vec{s}\right) - \left(\vec{s}^{\,*}\boldsymbol{P}^n\vec{s}\right)\left(\vec{s}^{\,*}\boldsymbol{P}^{-n}\tilde{\boldsymbol{W}}_n\boldsymbol{P}^n\vec{s}\right)$$

$$= \vec{s}^{\,*}\tilde{\boldsymbol{W}}_n\left(\boldsymbol{P}^n\vec{s}\vec{s}^{\,*} - \vec{s}\vec{s}^{\,*}\boldsymbol{P}^n\right)\vec{s} \quad + \quad \vec{s}^{\,*}\left(\vec{s}\vec{s}^{\,*} - \boldsymbol{P}^n\vec{s}\vec{s}^{\,*}\boldsymbol{P}^{-n}\right)\tilde{\boldsymbol{W}}_n\boldsymbol{P}^n\vec{s}$$

$$= \vec{s}^{\,*}\tilde{\boldsymbol{W}}_n\left(\|\vec{s}\|^2\boldsymbol{P}^n\vec{s} - a_s(n)\vec{s}\right) \quad + \quad \left(\|\vec{s}\|^2\vec{s}^{\,*} - a_s(n)\vec{s}^{\,*}\boldsymbol{P}^{-n}\right)\tilde{\boldsymbol{W}}_n\boldsymbol{P}^n\vec{s}$$

$$= (\alpha\vec{s}^{\,*} + \beta\vec{s}^{\,*}\boldsymbol{P}^{-n})\boldsymbol{W}_n(\beta\vec{s} + \alpha\boldsymbol{P}^n\vec{s}), \text{ where } \alpha = \sqrt{\|\vec{s}\|^2 + \sqrt{\|\vec{s}\|^4 - a_s(n)^2}}$$

and $\beta = -\operatorname{sgn}(a_s(n))\sqrt{\|\vec{s}\|^2 - \sqrt{\|\vec{s}\|^4 - a_s(n)^2}}$.

So $\|N(n)\| \le \|\alpha\vec{s}^{\,*} + \beta\vec{s}^{\,*}\boldsymbol{P}^{-n}\|\|\tilde{\boldsymbol{W}}_n\|_2\|\beta\vec{s} + \alpha\boldsymbol{P}^n\vec{s}\|$
$= \sqrt{(\alpha^2+\beta^2)\|\vec{s}\|^2 + 2\alpha\beta a_s(n)}\max_k\left[\omega(n)_k - \frac{a_w(n)}{p}\right]\sqrt{[\alpha^2+\beta^2]\|\vec{s}\|^2 + 2\alpha\beta a_s(n)}$
$= 2\left(\|\vec{s}\|^4 - a_s(n)^2\right)\max_k\left[\omega(n)_k - \frac{a_w(n)}{p}\right]$

And $\|D(n)\| = \left\|\left(\vec{s}^{\,*}\boldsymbol{W}_n\vec{s} + \vec{s}^{\,*}\boldsymbol{W}_{-n}\vec{s}\right)\left(\vec{s}^{\,*}\vec{s}\right)\right\|_2 = \|\vec{s}\|^2\left|\vec{s}^{\,*}\boldsymbol{W}_n\vec{s} + \vec{s}^{\,*}\boldsymbol{P}^{-n}\boldsymbol{W}_n\boldsymbol{P}^n\vec{s}\right|$
$\ge 2\|\vec{s}\|^4\min\left[\omega(n)_k\right]$

Finally, $|E_M(n)| \le \dfrac{\left(\|\vec{s}\|^4 - a_s(n)^2\right)\max_k\left[\omega(n)_k - \frac{a_w(n)}{p}\right]}{\|\vec{s}\|^4\min\left[\omega(n)_k\right]}$

To estimate the $\vec{\omega}(n)$-dependent terms in this error bound, I exploit the convexity of $\ln w$ to find the maximum absolute difference between two components of $\vec{\omega}(n)$.

For $0 \le k_1 < k_2 \le p-1$, $\omega(n)_{k_1} - \omega(n)_{k-2} = \sum_{t=-\infty}^{\infty} w[k_1+n+tp]w[k_1+tp] - \sum_{t=-\infty}^{\infty} w[k_2+n+tp]w[k_2+tp]$.

Since $ln(w)$ is concave down, $w[t]w[t+n]$ attains its maximum value (for a given n) at $w[1-\lceil n/2\rceil]w[\lfloor n/2\rfloor]$ and decays monotonically to zero on either side. Let $a_1 = \lceil\frac{-n/2-k_1}{p}\rceil, a_2 = \lceil\frac{-n/2-k_2}{p}\rceil$. Then $\sum_{t=a_1}^{\infty} w[k_1+n+tp]w[k_1+tp]$ and $\sum_{t=a_2}^{\infty} w[k_2+n+tp]w[k_2+tp]$ are nonincreasing nonnegative sequences, while $\sum_{t=-\infty}^{a_1-1} w[k_1+n+tp]w[k_1+tp]$ and $\sum_{t=-\infty}^{a_2-1} w[k_2+n+tp]w[k_2+tp]$ are nondecreasing and nonnegative.

If $a_1 = a_2$, then
$w[k_1+n+tp]w[k_1+tp] \ge w[k_2+n+tp]w[k_2+tp]$ for $t \ge a_1$, and

$w[k_1+n+tp]w[k_1+tp] \le w[k_2+n+tp]w[k_2+tp]$ for $t < a_1$.

Then $0 \le \sum_{t=a_1}^{\infty} w[k_1+n+tp]w[k_1+tp] - \sum_{t=a_2}^{\infty} w[k_2+n+tp]w[k_2+tp] \le w[k_1+n+a_1p]w[k_1+a_1p]$

and $-w[k_2+n+(a_2-1)p]w[k_2+(a_2-1)p] \le \sum_{t=-\infty}^{a_1-1} w[k_1+n+tp]w[k_1+tp] - \sum_{t=-\infty}^{a_2-1} w[k_2+n+tp]w[k_2+tp] \le 0$.

Thus $|\omega(n)_{k_1} - \omega(n)_{k-2}| \le w[1-\lceil n/2 \rceil]w[\lfloor n/2 \rfloor]$.

Similarly, If $a_1 = a_2 + 1$, then $w[k_1+n+tp]w[k_1+tp] \le w[k_2+n+(t-1)p]w[k_2+(t-1)p]$ for $t \ge a_1$, and $w[k_1+n+tp]w[k_1+tp] \ge w[k_2+n+(t-1)p]w[k_2+(t-1)p]$ for $t < a_1$.

Then $-w[k_2+n+a_2p]w[k_2+a_2p] \le \sum_{t=a_1}^{\infty} w[k_1+n+tp]w[k_1+tp] - \sum_{t=a_2}^{\infty} w[k_2+n+tp]w[k_2+tp] \le 0$

and $0 \le \sum_{t=-\infty}^{a_1-1} w[k_1+n+tp]w[k_1+tp] - \sum_{t=-\infty}^{a_2-1} w[k_2+n+tp]w[k_2+tp] \le w[k_1+n+(a_1-1)p]w[k_1+(a_1-1)p]$.

Again, $|\omega(n)_{k_1} - \omega(n)_{k-2}| \le w[1-\lceil n/2 \rceil]w[\lfloor n/2 \rfloor]$.

Set $c_n = w[1-\lceil n/2 \rceil]w[\lfloor n/2 \rfloor] = w[\lceil n/2 \rceil]w[\lfloor n/2 \rfloor]$. Since c_n bounds the absolute difference between components of $\vec{\omega}(n)$, and $a_w(n)$ is the sum of these components, $\frac{a_w(n)-(p-1)}{p}c_n \le \omega(n)_k \le \frac{a_w(n)+(p-1)}{p}c_n$. Then

$$\frac{\max_k \left[\omega(n)_k - \frac{a_w(n)}{p}\right]}{\min[\omega(n)_k]} \le \frac{\min_k[\omega(n)_k] + c_n - \frac{a_w(n)}{p}}{\min[\omega(n)_k]} \le \frac{(p-1)c_n}{a_w(n)-c_n}.$$

To summarize:
$|E_M(n)| \leq \frac{\left(\|\vec{s}\|^4 - a_s(n)^2\right)}{\|\vec{s}\|^4} \frac{(p-1)c_n}{a_w(n) - c_n}$
where $c_n = w[\lceil n/2 \rceil] w[\lfloor n/2 \rfloor]$.

The approximation due to Boersma (1993) gives

$$R_B(n) = \frac{\left(\sum_{t=-\infty}^{\infty} [w[t]\overline{s}_{t \bmod p}] \, [w[t+n] s_{t+n \bmod p}] \right) \left(\sum_{t=-\infty}^{\infty} w[t]^2\right)}{\left(\sum_{t=-\infty}^{\infty} w[t]^2 |s_{t \bmod p}|^2 \right) \left(\sum_{t=-\infty}^{\infty} w[t] w[t+n] \right)}$$

And using the vector notation from above,

$$R_B(n) = \frac{\left(\vec{s}^{\,*} \boldsymbol{W}_n \boldsymbol{P}^n \vec{s}\right) \left(\vec{1}^{\,*} \boldsymbol{W}_0 \vec{1}\right)}{\left(\vec{s}^{\,*} \boldsymbol{W}_0 \vec{s}\right) \left(\vec{1}^{\,*} \boldsymbol{W}_n \vec{1}\right)} = \frac{\left(\vec{s}^{\,*} \boldsymbol{W}_n \boldsymbol{P}^n \vec{s}\right) \left(\vec{1} \cdot \vec{\omega}(0)\right)}{\left(\vec{s}^{\,*} \boldsymbol{W}_0 \vec{s}\right) \left(\vec{1} \cdot \vec{\omega}(n)\right)}$$

Then the error term is $E_B(n) = R_B(n) - a_s(n)$
$$= \frac{\vec{s}^{\,*} \vec{s} \left(\vec{s}^{\,*} \boldsymbol{W}_n \boldsymbol{P}^n \vec{s}\right) \left(\vec{1} \cdot \vec{\omega}(0)\right) - \left(\vec{s}^{\,*} \boldsymbol{P}^n \vec{s}\right) \left(\vec{s}^{\,*} \boldsymbol{W}_0 \vec{s}\right) \left(\vec{1} \cdot \vec{\omega}(n)\right)}{\left(\vec{s}^{\,*} \boldsymbol{W}_0 \vec{s}\right) \left(\vec{1} \cdot \vec{\omega}(n)\right) \left(\vec{s}^{\,*} \vec{s}\right)}$$

Again, define $\tilde{\boldsymbol{W}}_n$ by $(\tilde{W}_n)_{k,k} = \omega(n)_k - \frac{a_w(n)}{p}$ and all other entries 0. Then

$$E_B(n) = \frac{\vec{s}^{\,*} \vec{s} \left(\vec{s}^{\,*} \tilde{\boldsymbol{W}}_n \boldsymbol{P}^n \vec{s}\right) \|\vec{w}\|^2 - \left(\vec{s}^{\,*} \boldsymbol{P}^n \vec{s}\right) \left(\vec{s}^{\,*} \tilde{\boldsymbol{W}}_0 \vec{s}\right) a_w(n)}{\left(\vec{s}^{\,*} \boldsymbol{W}_0 \vec{s}\right) a_w(n) \left(\vec{s}^{\,*} \vec{s}\right)}$$

The upper error bound for the Mates method could be reduced by exploiting the relationship between $\tilde{\boldsymbol{W}}_{-n}$ and $\tilde{\boldsymbol{W}}_n$, but no similar relationship exists between $\tilde{\boldsymbol{W}}_0$ and $\tilde{\boldsymbol{W}}_n$. Therefore the magnitude of each term in the numerator must be estimated separately:

$$\|E_B(n)\| \leq \frac{\|\vec{s}\|^4 \max_k \left[\omega(n)_k - \frac{a_w(n)}{p}\right] \|\vec{w}\|^2 + \|\vec{s}\|^2 a_s(n) \max_k \left[\omega(0)_k - \frac{\|\vec{w}\|^2}{p}\right] a_w(n)}{\min_k [\omega(0)_k] a_w(n) \|\vec{s}\|^4}$$

$$\leq \frac{\|\vec{s}\|^2 (p-1) c_n \|\vec{w}\|^2 + a_s(n) c_0 a_w(n)}{\left(\|\vec{w}\|^2 - (p-1)c_0\right) a_w(n) \|\vec{s}\|^2} = \frac{(p-1)c_n \frac{\|\vec{w}\|^2}{a_w(n)} + \frac{a_s(n)}{\|\vec{s}\|^2} c_0}{\left(\|\vec{w}\|^2 - (p-1)c_0\right)}$$

Thus, the Boersma method also improves in accuracy for slowly-decaying windows with wide support, as c_n and c_0 are small relative to $\|\vec{w}]\|^2$. For a given signal and window, error is expected to be smallest near zeros of the autocorrelation function, but no signal conditions lead to the bound above vanishing.

APPENDIX B

Tenure in current captive setting and age predict personality changes in adult pigtailed macaques [Sussman et al., 2014]

Tenure in current captive setting and age predict personality changes in adult pigtailed macaques

Adrienne F. Sussman[a, b], Exu A. Mates[a], James C. Ha[a, b], Kathy L. Bentson[a, b], Carolyn M. Crockett[a, b]

Affiliations and Correspondence
Adrienne F. Sussman (corresponding author) - adris@u.washington.edu
Exu A. Mates (corresponding author) – xamates@u.washington.edu
[a]Department of Psychology, University of Washington, Seattle, WA
[b]Washington National Primate Research Center, Seattle, WA

ABSTRACT:

Personality stability in nonhuman primates is a topic that warrants more research attention. Many studies focus on intra-individual repeatability, but few note population-wide trends in personality change. In part, this results from the large sample size that is required to detect such trends. This study measured personality in a large sample (N=274) of adult, mother-reared pigtailed macaques over a period of three years. We looked at four personality components (Sociability towards humans, Cautiousness, Aggressiveness, and Fearfulness) derived from behavioral observations at two to four time points per subject. We found these components to have repeatabilities similar to those reported elsewhere in the literature. We then analyzed both population-wide and individual-level stability using linear-mixed effects models. We found that adult personality changed with life experiences (here, tenure at the facility where tested). Our data did not indicate major changes in personality with age, except in cautiousness, but showed that pigtailed macaques became more sociable towards humans and less cautious as the percentage of their lifetime housed in indoor cages at a primate laboratory increased. Incorporating group-level trends into the model improved the predictive power of single measures of personality over an individual-level model. Other researchers may benefit by applying similar methodology to that described here as they extrapolate about personality measures over time.

KEYWORDS: *Macaca nemestrina*, *pigtailed macaque, personality, temperament, stability, repeatability, change*

Introduction

Animals across a broad spectrum of species show consistent individual differences in their behavioral patterns, which are referred to as temperament or personality traits. These traits are analogous to human personality traits in structure and measurement (Gosling, 2001), especially as observed in our nearest relatives, the nonhuman primates (Brosnan et al., 2009; Freeman & Gosling, 2010). However, theories of personality are better developed in humans than in nonhuman primates, especially in regard to how personalities change and adapt throughout an individual's lifetime. In the human literature, temperament is considered to be a genetically rooted set of behavioral tendencies, which eventually develop into personality as these biological predispositions interact with experiences and environment (Rothbart et al., 2000). In humans, individuals' personality trait scores change throughout adulthood, with more change occurring during adolescence and early adulthood, and traits becoming more stable later in life (Haan et al., 1986; McCrae et al., 2000; Roberts & DelVecchio, 2000; Helson et al, 2002; Terracciano et al., 2005; Roberts et al., 2006; Quoidbach et al., 2013). In the quickly growing pool of animal personality studies, only a few examine similar effects in nonhuman primates.

The few studies that have examined population trends in personality change in nonhuman primates have demonstrated that, as in humans, personality trait scores change with age. These developmental changes in infant and juvenile macaque personality are well documented (e.g., Stevenson-Hinde et al., 1980; Sussman & Ha, 2011), as are changes with age in adult chimpanzees (King et al., 2008; Weiss et al., 2007) and other great apes (Weiss et al., 2012). The latter studies suggest that personality change in

chimpanzees follows a similar pattern as in humans, becoming more stable with age. Other studies have noted that personality can be predicted by age in several species, including vervet monkeys (McGuire et al., 1994), capuchins (Manson & Perry, 2013), and callitrichids (Kendal et al., 2005). For most species, though, stability of personality over the lifetime has not been investigated.

Though stability studies are fairly rare in the primate literature, studies reporting personality trait *repeatability* are more common. Repeatability is a measure of trait reliability, and measures how similar a subject's behavioral responses are at two time points. Conceptually, while stability measures how much an individual's personality changes over time, repeatability measures the proportion of behavioral variation in a population that is due to individual differences (David et al., 2012). A repeatability analysis tells us the mean probability that a subject's personality score will be the same at a later time point, while a stability analysis identifies patterns of change over time. Repeatability analyses are vital for validation of proposed personality traits, but stability analyses can provide additional important information for many animal researchers.

A further point of confusion is that repeatability analyses are often compared across studies, which may use very different measurement intervals, ranging from a month or two (Uher et al., 2008; Higley et al., 1996), to a year (Konečná et al, 2012; Maestripieri, 2000; Martau et al., 1985), to several years (Capitanio, 1999; Koski, 2011; Stevenson-Hinde et al., 1980). The Bell et al. (2009) study found that measurements taken over short time periods were significantly more repeatable than measurements over longer time periods. They did not find a difference in repeatability with age, but their analysis included a large variety of taxa, including insects, which might not be expected

to adhere to the same lifetime repeatability trends as humans or other primates. The average repeatability in the Bell et al. study was 0.37 (2009), while in an analysis of nonhuman primate studies, the average repeatability was 0.58 (Freeman & Gosling, 2010). Both of these values suggest less-than-perfect repeatability, which may be due measuring error, random variation, or developmental changes. Repeatability measures do not distinguish between these sources of variance. Moreover, because repeatability is measured from a larger sample, repeatability measures will not accurately reflect future behaviors of all individuals, who may differ in their consistency (Bell et al., 2009).

One explanation for the lack of stability analyses in the primate literature is that they require more data than repeatability analyses. In order to most powerfully identify non-linear patterns of change, researchers require large sample sizes and at least three measurement points (Rogosa et al., 1982). The most powerful data sets are longitudinal measurements, which, through hierarchical analyses, can provide information about both individual- and population-level changes over time (Rogosa et al., 1982). An especially powerful way to analyze nested data is using model-fitting techniques in place of traditional significance testing (Anderson et al., 2000).

In this project, we used a large sample (N=274) of mother-reared adult (≥ 4 years old) pigtailed macaques (*Macaca nemestrina*) to identify patterns of lifetime stability using model fitting techniques. In a previous analysis, we had identified four personality components – Sociability towards humans, Cautiousness, Aggressiveness, and Fearfulness – including subjects in this sample (Sussman et al., 2013). Individuals were tested up to four times over a three-year study period, and our analyses used the

hierarchical nature of the data to identify both population- and individual-level trends within the data.

Our first model identified population-wide patterns of change over the lifetime, and measured the extent to which they were related to age, sex or general experience independent of age. We predicted that we would find that, as in humans, population-wide personality growth curves of macaques would show a quadratic shape, changing more rapidly in younger animals than older animals. In addition to age, we also tested the population-wide relationship between tenure at the current primate facility and personality. We believed that tenure might affect personality as animals habituated to the primate facility, as habituation is well documented in macaques (Capitanio et al., 2006; Crockett et al., 1993, 1994). Unlike long-tailed or rhesus macaques, pigtailed macaques who were at WaNPRC for at least a year scored high on sociability towards humans and low on cautiousness (Sussman et al., 2013). We were interested to learn whether sociability increased and cautiousness decreased as the animals habituated during their first year at the facility, and whether there were any additional changes with increasing time.

Our second model examined within-individual personality stability, and tested how accurately a single personality measure can predict a later personality score –We predicted that, by including the variables of age, sex, testing interval, or current facility tenure, we could account for more of the variance in personality than is accounted for by extrapolating from a single test score. We also predicted that our personality components would have repeatability scores similar to those reported in the literature.

Methods

Subjects and Housing

Behavioral data were collected between 2003 and 2006 on monkeys housed at the National Primate Research Center, University of Washington (WaNPRC), Seattle. The sample included 274 mother-reared pigtailed macaques (*Macaca nemestrina*) (172 female, 102 male) between the ages of 4.0 and 18.6 years at first test (M = 7.8 years, SD=3.6). The subjects were a subset of those in Sussman et al. (2013). We restricted our analysis to the species with the largest sample size, and excluded nursery-reared subjects to reduce the number of variables in our analysis.

When studying animals with complex social systems, like primates, it can be difficult to disentangle the effects of social rank and individual tendencies on an individual's behavior. We used individually housed monkeys for this study. The subjects were housed in single or grooming-contact (Crockett et al., 1997) indoor cages at the WaNPRC. They were subjects in a wide variety of biomedical research studies. Monkeys had visual contact with conspecifics. To address our research goals of modeling the time course of personality changes over time in our Primate Center setting separately from age, our analyses included the variable "tenure," or proportion of the animal's lifespan spent at WaNPRC, which ranged from 1.0% to 91.5% (M = 26.4%; SD = 20.7%). Tenure fraction was chosen because it was relatively uncorrelated with age, unlike raw tenure duration. All monkeys had been at WaNPRC in these conditions for at least 53 days before their first test (M= 608.8 days; SD = 771.0). Subjects in this study originally came from breeding facilities (primarily Tulane National Primate Research Center and Bogor, Indonesia) where they were housed outdoors in breeding social groups. Although they occasionally experienced caged indoor laboratory housing in the

prior facilities, we considered that the move to the WaNPRC facility represented a significant change in the social and physical environmental conditions for these animals.

Animal rooms were maintained on a 12:12 hour light-dark cycle, with an ambient temperature of 22.2° to 25.6° C. Subjects received commercial monkey biscuits twice daily, and ad libitum water, as well as environmental enrichment such as a portable toy and foraging device, and fresh produce or foraging opportunities seven days a week. The University of Washington Institutional Animal Care and Use Committee approved the observational techniques of this study. Our research methods complied with legal requirements of the U.S.A., the state of Washington, and the AAALACi.

Personality Testing

We used a rapid cage-front behavioral assessment to measure personality (RATR: Rapid Assessment of Temperament and Reactivity; Bentson 2003; Sussman et al., 2013). During a 4-minute observation period, the observer (KB) recorded frequencies of 37 variables of interest using a PDA hand-held device. Some behaviors were measured by instantaneous sampling every minute, while others were measured by whether or not they occurred during each minute of the observation period.

Each monkey received two to four tests over the course of three years. Monkeys that entered the WaNPRC during the three-year period (n = 140) were tested beginning 8 weeks after arrival, retested about 8-10 weeks later, and then tested annually. Monkeys that were already at the primate center at the start of the three-year period were tested annually. Our use of growth curve modeling allowed us to combine data for individuals with different numbers of tests into a single, population-wide model. In our sample, 141 individuals received just two tests, 104 received three tests, and 29 received all four tests.

There was an average length of 164 days between the first and second tests (*SD*=158.6 days), 397 days between the second and third tests (*SD* = 136.0), and 373 days between the third and fourth tests (*SD*=117.7). These intervals reflect the mix of testing schedules for monkeys that arrived during the three-year interval and those that were at WaNPRC for over a year before they were first tested.

Personality Component Identification

Personality components were identified as described in Sussman et al. (2013). Though the methodologies were similar, we here use the term "personality" rather than "temperament" because "personality" is most often used in human stability studies. Our analyses focused on twelve behavioral variables of interest identified by Sussman et al. (2013). These included whether the animal was in the front or back of the cage; whether it reached towards or approached the observer; whether it gave a lipsmack to the observer, showed quiet attention towards the observer, or ignored the observer; whether it performed an open-mouth threat, a lunge, or a full threat display towards the observer; and whether it shrieked or fear grimaced. Using a principal components analysis (PCA), we identified four personality components, which we identified as Sociability towards humans, Cautiousness, Aggressiveness, and Fearfulness (Sussman et al., 2013).

In our first identification analysis, we used only a single observation for each individual (the first test conducted when the subject had been at the facility for at least a year), and did not attempt to assess the repeatability of the measures (Sussman et al., 2013). Given the goals of the present study, we wanted to make sure that the structure of the personality components was the same at each of the testing periods within pigtailed macaques before proceeding. To do this, we performed PCAs, using the twelve variables

previously identified, and specifying four components, for each of the first three behavioral observations. We compared the structure of the orthogonally rotated component matrices for the first, second, and third tests using Tucker's Congruence Coefficient (Lorenzo-seva & ten Borge, 2006). The fourth test was not included in the component congruency analysis because of small sample size. We also compared all three tests to the structure of the full-sample PCA from our previously published study, which included a much larger sample size (n=899), and subjects from three species of macaque, including some that had been nursery-reared.

We found that all comparisons between observation periods were congruent, exceeding the minimum congruence level of 0.85 suggested by Lorenzo-seva and ten Borge (2006; Table 1). In other words, this analysis shows that the same personality components existed for the first, second, and third test. To maintain maximum consistency with our past work, we chose to calculate individual component scores for each test using the same methods as in our prior publication. We used the regression method to calculate individual scores, specifying the same equation values as in our previous analysis. We used SPSS 18.0 (IBM, 2008) for these analyses.

Repeatability analyses

In addition to our model fitting, we calculated repeatabilities for each of our temperament components to compare to those published elsewhere in the literature. In the interest of maximizing comparability, we calculated repeatability using both Intra-class correlation (ICC) (Bell et al., 2009) and Pearson's correlations between tests (Freeman & Gosling, 2010). These analyses were conducted in R 2.15.2 using version 2.1 of the "ICC" package.

Stability analyses

Analysis 1: Population-wide stability

To investigate population-wide patterns of change, we fit linear mixed-effect models to each component score with the random effect of subject ID, and a set of predictors. Predictors included sex, age at testing, and tenure fraction (the proportion of the animal's lifespan spent in its current housing location). Both linear and quadratic effects were examined for age and tenure fraction, as well as the interactions of each with sex. All distinct subsets of effects were examined, under the usual constraint that no subset may contain a product or quadratic term unless it also contains the main or linear terms as well. This resulted in a total of 34 models. All models were fit with R 2.15.2, using version 3.1-108 of the "nlme" package (Pinheiro et al., 2010).

The explanatory power of models was judged by the Akaike Information Criterion with small-sample correction (AICc, Hurvich & Tsai, 1989). As p-values can be inaccurate for nested data, this approach is preferable (Pinheiro & Bates, 2000). Because of the large number of models compared and the significant possibility that there would be multiple models with near-minimal AICc values, we followed a model-averaging approach as described by Burnham and Anderson (2002). Each model is assigned an Akaike weight proportional to the inverse of the exponential of its AICc value, and predictions are averaged across all models using these weights. This approach avoids overly privileging the single "best" model if it has very close competitors in AICc values, and has repeatedly shown superior predictive performance to the best-model approach (Burnham & Anderson, 2004). Model averaging was performed in R 2.15.2 with version 1.27 of the "AICcmodavg" package.

For heuristic purposes, we assessed the goodness of fit of our model-averaged predictions using classical R^2. To assess the significance of each individual parameter, we again followed Burnham and Anderson (2002). We examined the subset of models containing that parameter but not containing any associated higher-order parameter. (E.g., for the linear age parameter, models containing a quadratic age parameter or sex-by-age interaction parameter were excluded.) The value and confidence intervals for this parameter were then found by averaging across this subset, using renormalized Akaike weights. If a value differs significantly from zero, this indicates that the parameter has a significant effect on the response in the most informative models that include it.

Analysis 2: Individual Stability

To investigate whether these variables could be used in conjunction with previous test scores to predict later scores, we divided our set of tests into a set of 274 first tests on all animals, and 436 later tests (2^{nd}, 3^{rd}, and 4^{th} tests). For each component score, we fit mixed-effect models to the set of first tests, according to the procedure for Analysis 1, and then computed model-averaged population-level predictions (PLPs) for the scores on the set of later tests. We then fit mixed-effect models to the later test scores, using the predictors: first test score, PLP, and time elapsed from first test to current test. Interactions between time elapsed and first test score, and between time elapsed and PLP, were also considered. Parameter significance and model predictive accuracy were assessed as in Analysis 1.

Results:

Repeatability

Measuring repeatability using ICC, we found that Sociability and Cautiousness had fairly high repeatability, while Fearfulness was slightly less repeatable, and Aggressiveness was not very repeatable (Table 2). We obtained very similar values when using mean Pearson's correlations.

These repeatability findings were consistent with the results of our mixed-effect model stability analysis, though the models better explained the within-individual variation than did the repeatability analyses. We identify the best fitting models (identified by lowest Akaike weight) for each personality component in Table 3 (population-wide model) and Table 4 (within-individual model), along with the significant predictors within that model. We briefly discuss trends in significant predictors for each set of models below.

Population Wide Model

The best-fitting models for population-wide trends explained much of the variance in Sociability and Cautiousness, but not in Aggressiveness or Fearfulness. Specifically, the best-fit model for the population trends explained 35% of the variance in Sociability towards humans and 16% of the variance in Cautiousness, but only 7% of the variance in Aggressiveness and 3% of the variance in Fearfulness (Table 3). These R^2 values increased dramatically when individual ID was added as a predictor variable. When ID was included, all models predicted >50% of the variance in the personality components.

Age was a significant predictor for Cautiousness only, with Cautiousness decreasing as animals aged from 4 years (Figure 1). No components included a quadratic

effect of age as a significant predictor, which had been our prediction based on human and ape studies.

Tenure fraction was a significant predictor for Sociability towards humans, Cautiousness, and Fearfulness. As the fraction of lifespan spent in the facility increased, Sociability increased and Cautiousness decreased (Figure 2). The effect of tenure fraction was quadratic for both Sociability towards humans and Fearfulness, with the predicted values of both components peaking for animals that had spent about 50% of their lives at WaNPRC in Seattle.

Sex was a significant predictor of Sociability towards humans and Cautiousness. Males tended to be more sociable and less cautious than females (Figure 1). There was also a significant sex-by-age interaction for Sociability towards humans, with males decreasing in sociability at older ages and females increasing in sociability (Figure 1). There were significant sex-by-tenure interactions for Aggressiveness and Fearfulness, although the amount of variance they explained was small. In both cases, females had a stronger quadratic component to their score distribution with increasing tenure than did males. The model indicated that females with intermediate tenure fractions were more aggressive and fearful than their male counterparts, but females with very small or large tenure fractions were not.

Individual Stability Model

Similar to the population-wide model, the individual-level model was a better fit for Sociability towards humans and Cautiousness than for the other components, explaining 37% and 30% of the variance in these component scores, respectively. The

model explained only 9% of the variance in Aggressiveness, and 24% of the variance in Fearfulness (Table 4).

First test score was a significant predictor of later test scores for all personality components (Table 4). Using PLPs significantly improved predictive power for Sociability towards humans, Cautiousness, and Aggressiveness. Time elapsed between first and later tests was also a significant predictor for Cautiousness and Fearfulness, with monkeys showing more cautious and fearful behaviors after longer inter-test intervals. Finally, there were significant first test score-by-time elapsed interactions for Sociability and Fearfulness. These interaction effects reflected an *increase* in variance in Fearfulness on later test scores when inter-test intervals were long, and a *decrease* in variance in Sociability under the same conditions.

Discussion

We used repeated assessments of adult pigtailed macaques to evaluate repeatability, individual stability, and population-wide stability of four personality components. Personality components at each assessment were similar to those identified in a previous analysis of a larger dataset including three macaque species (Sussman et al., 2013).

Repeatability

Overall, we found significant component repeatability. Using ICC, the repeatability scores of three of our components (Sociability, Cautiousness, and Fearfulness) exceeded the mean repeatability of 0.37 reported in Bell et al.'s meta-analysis (2009). Aggressiveness, with a repeatability of only 0.26, was our least repeatable component. Our test-retest correlation values also compared favorably with

those reported in the primate literature, as reviewed by Freeman and Gosling (2010); all of our component correlations fell within their reported range of 0.35-0.88, although all were smaller than their weighted average of 0.58. Aggressiveness was, again, the least repeatable component and Sociability towards humans and Cautiousness were the most repeatable. In contrast, aggressiveness tended to be among the most repeatable traits in other studies (Bell et al., 2009). The discrepancy with our results is likely due to limited variance in our measure of Aggressiveness, which reduces ICC. Pigtailed macaques showed lower average scores on both Aggressiveness and Fearfulness, and less variance, than long-tailed and rhesus macaques (Sussman et al., 2013). Our results also differed somewhat from other findings from the primate literature. For example, Uher found that "friendliness to humans" was the least reliable measure in great apes (Uher et al., 2011). This discrepancy might reflect differences in the contexts of the observational measures in each study, or might also reflect true differences across species in repeatabilities of this trait.

Population-wide Stability

The most striking finding of our population-wide model was that, for our animals, the proportion of their life spent at the Seattle WaNPRC facility was at least as good a predictor of personality as was age at testing. While age was included in all of the best-fitting models, it was only a significant independent predictor of Cautiousness. There was also no evidence for a significant quadratic effect of age, unlike what has been documented in the human and chimpanzee literature (e.g., Weiss et al., 2007). Past work demonstrates that pigtailed macaque personality does change significantly during infancy

(Sussman and Ha, 2011), but our present findings suggest that personality components vary only slightly by age in mature adults.

Past studies of nonhuman primates have suggested behavioral habituation in response to changes in captive housing conditions, with behaviors and physiological measures becoming more stable after about 3 months in a given housing situation (Capitanio et al., 2006). In a related study, urinary cortisol levels continued to decline for more than a year after arrival in adult longtailed macaques, *M. fascicularis* (Crockett, et al., 1993, 1994). Our present study found that changes in personality components may reflect another aspect of habituation as it is seen in pigtailed macaques. Our data demonstrate that longer tenures of pigtailed macaques at WaNPRC predict decreased Cautiousness and increased Sociability towards humans. This relationship between tenure and personality might be true of pigtailed macaques, but not other species. A number of studies suggest that pigtailed macaques show unique responses to human interactions. For example, pigtailed macaques are more neophilic than some other macaque species (Montgomery et al., 2005), are likely to direct social behaviors towards human observers (Oettinger et al., 2007), and are more easily trained to perform some tasks than long-tailed macaques (Crockett & Wilson, 1980). Recent work suggests that captivity alters behavior differentially in different species (Mason et al, 2013). Comparative studies are needed to establish whether tenure is an important predictor of personality in other species. Based on the tenure finding, we encourage other investigators to include similar factors in their models of personality stability.

Tenure at WaNPRC was associated with slightly increased Aggressiveness (especially for females), and Fearfulness appeared to peak for individuals who had spent

around 50% of their lives at the Seattle WaNPRC facility (Figure 2). Both of these effects were small, however, and which might be due to the fact that these personality components are lower in pigtailed macaques than in long-tailed and rhesus macaques to begin with (Sussman et al. 2013). In addition, only 12% (n = 85) of our tests were conducted on females with tenure fractions above 50%, so the quadratic effect in Fearfulness should be interpreted with caution. Our data are insufficient to determine whether these component scores really do peak around 50% and then drop thereafter, or whether they simply converge asymptotically to a maximum as tenure increases.

Sex was a significant predictor of all component scores, and the results were consistent with the sex differences we identified previously (Sussman et al., 2013). Males were more sociable towards humans and less cautious than females. The significant sex difference in sociability was found only in younger adult animals (Figure 1). Sex did not appear to be a meaningful predictor of personality stability, however. Females showed a quadratic change in Fearfulness and Aggressiveness with tenure, which was not present for males, with females who had spent about half of their lives at the WaNPRC facility scoring the highest on these traits. However, this quadratic effect should be interpreted carefully for the reasons described above, and these sex-by-tenure^2 interactions do not necessarily represent meaningful effects. On the other hand, these effects do give weak support for the theory that males have more stable personalities than females. Some researchers have suggested that males should be more stable in traits such as aggressiveness, which are strongly linked to male-specific hormones (Wingfield, 1994; Andrew, 1972), while others have suggested that sexually selected traits would be more stable in males than in females (Kokko, 1998; Garamszegi et al., 2006).

Including individual ID in the model greatly increased the amount of variance explained to 50%-75% (Table 3), indicating that individuals had different levels of personality stability. Several other studies have noted similar inter-individual differences in stability, which in some cases is referred to as a distinct personality trait, "consistency" (Dingemanse et al, 2010; Bell et al., 2009).

Individual stability

We found that using the first measure of personality to predict later behavior was a reasonable strategy for Sociability, Cautiousness, and Fearfulness, as the first measure alone explained 20-29% of the variance in later behaviors (Table 4). First test was a poor measure of Aggressiveness, explaining only 6% of the variance in that component. This is consistent with the low repeatability we found for Aggressiveness, and was probably exacerbated by the slightly skewed distribution of this component. When we added information about population-level predictions and time elapsed between tests to our model, they explained considerably larger proportions of variance in trait scores, with incremental f^2 values of 3-13 %. This notable increase makes this approach useful for any investigator trying to connect personality measured at one time with an outcome at a different time. By adding in such information, researchers may be able to make their analyses more informative.

As noted above, individuals differed in their personality consistency. Including information about individual consistency in the individual-level model would most likely improve the predictive power of the model considerably. However, we chose to use first tests only because this is the information that most researchers have access to. Even without adding consistency data, our model predicted an individual's future behavior

significantly better than a repeatability score. For example, using our model with population-level information included, we can explain 37% of the variance in an individual's Sociability towards humans over multiple assessments – compared with just 26% using test-retest correlations. Our model explains more of the variance in all traits except Aggressiveness – again, perhaps due to the non-normal distribution of this component.

Conclusions:

Our analysis offers a significant contribution to the literature, as the first study to model the stability of personality in a large sample of a macaque species over a multi-year time frame. Our repeatability analysis demonstrates that our methods of measuring personality are about as reliable as other methods reported in the literature. Based on our population-level analysis, we recommend that other researchers include information about tenure in current facility into their stability analyses: in our sample, tenure was a better predictor of personality than was age. We also found that an individual-level model that incorporates some information about group-level trends more accurately explains variance in later personality scores. The development of these models for other species and refinement of the model as it is applied to the same species in different settings may be quite useful to all researchers attempting to extrapolate from a single individual personality measure to later outcomes. Overall, the model-fitting techniques used here represent a powerful tool for understanding change and consistency in personality, and we recommend that other primate researchers adopt similar methods in their studies.

References:

Anderson, D.R., Burnham, K.P., Thompson, W.L. (2000). Null hypothesis testing: problems, prevalence, and an alternative. *Journal of Wildlife Management, 64*, 912-923.

Andrew, R.J. (1972). Recognition processes and behavior with special reference to effects of testosterone on persistence. *Advances in the Study of Behaviour, 4,* 175-208.

Bell, A.M., Hankison, S.J., Laskowski, K.L. (2009). The repeatability of behaviour: a meta-analysis. *Animal Behaviour, 77*, 771-783.

Bentson, K.L., Crockett, C.M., Ha, J.C. (2003). A rapid home cage procedure for assessing individual and group differences in behavioral reactivity of monkeys. *American Journal of Primatology, 60*, (Suppl.):77.

Brosnan, S.F., Newton-Fisher, N.E., van Vugt, M. (2009). A melding of the minds: When primatology meets personality and social psychology. *Personality and Social Psychology Review, 13*(2), 129-147.

Burnham, K. P., Anderson, D. R. (2002). *Model selection and multimodel inference: a practical information-theoretic approach.* Second edition. Springer: New York.

Burnham, K. P., Anderson, D. R. (2004). Multimodel inference: understanding AIC and BIC in model selection. *Sociological Methods and Research*, 33, 261–304.

Capitanio, J.P. (1999). Personality dimensions in adult male rhesus macaques: Prediction of behaviors across time and situation. *American Journal of Primatology, 47(*4): 299-320.

Capitanio, J.P., Kyes, R.C., Fairbanks, L.A. (2006). Considerations in the selection and conditioning of old world monkeys for laboratory research: Animals from domestic sources. *ILAR Journa*l, *47* (4), 294-306.

Crockett, C.M., Wilson, W.L. (1980). The ecological separation of Macaca nemestrina and Macaca fascicularis in Sumatra. In: LIndburg, D.G., editor: *The Macaques: Studies in Ecology, Behavior and Evolution*. New York: Van Nostrand Reinhold, pp. 148-181.

Crockett, C. M., Bowers, C. L., Sackett, G. P., Bowden, D. M. (1993). Urinary cortisol responses of longtailed macaques to five cage sizes, tethering, sedation, and room change. *American Journal of Primatology 30*(1), 55-74.

Crockett, C. M., Bowers, C. L., Bowden, D. M., Sackett, G. P. (1994). Sex differences in compatibility of pair-housed adult longtailed macaques. *American Journal of Primatology 32*, 73-94.

Crockett, C. M., Bellanca, R. U., Bowers, C. L., Bowden, D. M. (1997). Grooming-contact bars provide social contact for individually caged laboratory macaques. *Contemporary Topics in Laboratory Animal Science 36*(6), 53-60.

David, M., Auclair, Y., Cezilly, F. (2012). Assessing short- and long-term repeatability and stability of personality in captive zebra finches using longitudinal data. *Ethology 118*, 932-942.

Dingemanse, N.J., Kazem, A.J.N., Reale, D., Wright, J. (2010). Behavioural reaction norms: Animal personality meets individual plasticity. *Trends in Ecology & Evolution, 25*(2), 81-89.

Freeman H.D., Gosling S.D. (2010). Personality in nonhuman primates: a review and evaluation of past research. *American Journal of Primatology, 72*(8): 653-671.

Garamszegi, L.Z., Rosivall, B., Hegyi, G., Szollosi, E., Torok, J. & Eens, M. (2006). Determinants of male territorial behavior in a Hungarian flycatcher population: Plumage traits of residents and challengers. *Behavioral Ecology and Sociobiology, 60*, 663–671.

Gosling, S.D. (2001). From mice to men: What can we learn about personality from animal research? *Psychological Bulletin, 127*(1): 45-86.

Haan, N., Millsap, R., Hartka, E. (1986). As time goes by: Change and stability in personality over fifty years. *Psychology and Aging, 1*(3), 220-232.

Helson, R., Kwan, V.S.Y., Jon, O.P., Jones, C. (2002). The growing evidence for personality change in adulthood: Findings from research with personality inventories. *Journal of Research in Personality, 36*(4), 287-306.

Higley, J.D., King, S.T., Hasert, M.F., Champoux, M., Suomi, S.J., Linnoila, M. (1996). Stability of interindividual differences in serotonin function and its relationship to severe aggression and competent social behavior in rhesus macaque females. *Neuropsychopharmacology*, 1467-1476.

Hurvich, C.M., Tsai, C. (1989). Regression and time series model selection in small samples. *Biometrika, 76*(2), 297-307.

Kendal, R.L., Coe, R.L., Laland, K.N. (2005). Age differences in neophilia, exploration, and innovation in family groups of callitrichid monkeys. *American Journal of Primatology, 66*(2), 167-188.

King, J.E., Weiss, A., Sisco, M.M. (2008). Aping humans: Age and sex effects in chimpanzee (Pan troglodytes) and human (Homo sapiens) Personality. *Journal of Comparative Psychology, 122*(4), 418-427.

Konečná, M., Weiss, A., Lhota, S. & Wallner, B. (2012). Personality in Barbary macaques (Macaca sylvanus): Temporal stability and social rank. *Journal of Research in Personality, 46*, 581-590

Koski, S.E. (2011). Social personality traits in chimpanzees: Temporal stability and structure of behaviourally assessed personality traits in three captive populations. *Behavioral Ecology and Sociobiology, 65*(11), 2161-2174.

Kokko, H. (1998). Should advertising parental care be honest? *Proceedings of the Royal Society of London, Series B, 265*, 1871–1878.

Lorenzo-Seva, U., ten Berge, J.M.F. (2006). Tucker's congruence coefficient as a meaningful index of similarity. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences, 2*(2): 57-64.

Maestripieri, D. (2000). Measuring temperament in rhesus macaques: Consistency and change in emotionality over time. *Behavioural Processes, 49*(3), 167-171.

Manson, J.H., Perry, S. (2013). Personality structure, sex differences, and temporal change and stability in wild white-faced capuchins, Cebus capucinus. *Journal of Comparative Psychology*.

Martau, P.A., Caine, N.G., Candland, D.K. (1985). Reliability of the emotions profile index, primate form, with Papio hamadryas, Macaca fuscata, and two Saimiri species. *Primates, 26*(4), 501-505.

Mason, G., Burn, C.C., Dallaire, J.A., Kroshko, J., Kinkaid, H.M., Jeschke, J.M. (2013). Plastic animals in cages: behavioural flexibility and responses to captivity. *Animal Behaviour, 85*, 1113-1126.

McCrae, R.R., Costa, P.T., Ostendorf, F., Angleitner, A., Hrebickova, M., Avia, M.D, Smith, P.B. (2000). Nature over nurture: Temperament, personality, and life span development. *Journal of Personality and Social Psychology, 78*, 173–186.

McGuire, M.T., Raleigh, M.J., Pollack, D.B. (1994). Personality features in vervet monkeys: The effects of sex, age, social status, and group composition. *American Journal of Primatology, 33*(1), 1-13.

Montgomery, H.B., Bentson, K.L., Crockett, C.M. (2005). Responses to novelty in Macaca nemestrina and Macaca fascicularis varies by species-sex and time in facility. *American Journal of Primatology 66*(1): 146-147.

Oettinger, B.C., Crockett, C.M., Bellanca, R.U. (2007). Communicative contexts of the LEN expression of pigtailed macaques (Macaca nemestrina). *Primates 48*, 293-302.

Pinheiro, J.C., Bates, D. M. (2000). *Mixed Effects Models in S and S-Plus*. New York, NY: Springer.

Pinheiro, J., Bates, D., DebRoy, S., Sarkar, D. (2010). the R Core team (2009) nlme: Linear and Nonlinear Mixed Effects Models. R package version 3.1-96. *R Foundation for Statistical Computing, Vienna.*

Quoidbach, J., Gilbert, D.T., Wilson, T.D. (2013). The end of history illusion. *Science, 339*(6115), 96-98.

Roberts, B.W., DelVecchio,W.F. (2000). The rank-order consistency of personality traits from childhood to old age: A quantitative review of longitudinal studies. *Psychological Bulletin, 126*(1), 3-25.

Roberts, B. W., Walton, K. E., & Viechtbauer, W. (2006). Patterns of mean-level change in personality traits across the life course: A metaanalysis of longitudinal studies. *Psychological Bulletin, 132*, 1–25

Rogosa, D., Brandt, D., & Zimowski, M. (1982). A growth curve approach to measurement of change. *Psychological Bulletin, 92*(3), 726-748.

Rothbart, M.K., Ahadi, S.A., Evans, D.E. (2000). Temperament and personality: Origins and outcomes. *Journal of Personality and Social Psychology, 78*(1), 122-135.

Stevenson-Hinde, J., Stillwell-Barnes, R., Zunz, M. (1980). Subjective assessment of rhesus monkeys over four successive years. *Primates, 21*, 66 – 82.

Sussman, A., Ha, J.C. (2011). Developmental and cross-situational stability in infant pigtailed macaque temperament. *Developmental psychology, 47*(3). 781-791.

Sussman, A.F., Ha, J.C., Bentson, K.L., Crockett, C.M. (2013). Relationships among sex, species, and temperament in rhesus, longtailed, and pigtailed macaques. *American Journal of Primatology 75(4),* 303-313.

Terracciano, A., McCrae, R.R., Brant, L.J., Costa, P.T. (2005). Hierarchical linear modeling analyses of NEO-PI-R scales in the Baltimore Longitudinal Study of Aging. *Psychology and Aging, 20*, 493–506.

Uher, J., Asendorpf, J.B., Call, J. (2008). Personality in the behaviour of great apes: Temporal stability, cross-situational consistency and coherence in response. *Animal Behaviour, 75*(1), 99-112.

Uher, J. (2011). Personality in nonhuman primates: What can we learn from human personality psychology. In A. Weiss, J. King, & L. Murray (Eds.) *Personality and Temperament in Nonhuman Primates.* (pp. 41-76). New York: Springer.

Weiss, A., King, J.E., Hopkins, W.D. (2007). A cross-setting study of chimpanzee (Pan troglodytes) personality structure and development: Zoological parks and Yerkes National Primate Research Center. *American Journal of Primatology, 69*(11), 1264-1277.

Weiss, A., King, J.E., Inoue-Murayama, M., Matsuzawa, T., Oswald, A.J. (2012). Evidence for a midlife crisis in great apes consistent with the U-shape in human well-being. *Proceedings of the National Academy of Sciences of the United States of America, 109*(49), 19949-19952.

Wingfield, J.C. (1994). Control of territorial aggression in a changing environment. *Psychoneuroendocrinology, 19*, 709–721

Table 1: Demographic information for subjects used in this study. Entry age, age at first test, and tenure at 1st test are measured in years, and time elapsed between tests is measured in days.

	Full sample				Males			Females		
	Mean	SD	Range	N	Mean	SD	N	Mean	SD	N
Entry age	5.99	3.38	1.0 – 15.7	293	3.98	1.90	115	7.29	3.49	178
Age at first test	7.85	3.58	4.0 – 17.2	293	6.58	2.79	115	8.68	3.79	178
Tenure at 1st Test	1.86	2.25	0.1 – 12.8	293	2.60	2.39	115	1.38	2.02	178
Time elapsed: 1st – 2nd test	185.24	170.52	20 – 931	293	176.75	164.90	115	190.65	174.24	178
Time elapsed: 1st – 3rd test	580.74	231.84	77 – 1007	150	493.89	171.72	53	628.20	247.01	97
Time elapsed: 1st – 4th test	818.82	109.18	424 - 1000	40	851.81	90.97	21	782.37	118.23	19

Table 3: Repeatability for four temperament traits over 3 years (2-4 testing periods). Measures include ICC (Bell et al., 2009) and mean Pearson's correlation (Freeman & Gosling, 2010). N=293; estimated k for individuals is 2.66

	Sociability	Cautiousness	Aggressiveness	Fearfulness
Single Measures ICC	0.54	0.53	0.28	0.42
95% CI for ICC	0.47-0.60	0.46-0.60	0.19-0.36	0.34-0.50
Within-Test Variance	0.40	0.40	0.30	0.25
Between-Test Variance	0.47	0.45	0.12	0.18
Mean Correlation	0.48	0.45	0.28	0.42
95% CI for Correlation	0.46-0.69	0.46-0.69	0.27-0.50	0.35-0.58

Table 3: Fixed estimates of the best fitting models of population-wide personality change. Equations show the best-fitting model, as determined by highest Akaike weight. Significant predictors are those variables with coefficients that differ significantly from zero in the more informative models. R^2 without subject ID describes the variance in personality explained by the model as shown, and R^2 with ID describes the variance explained by the model with "ID" included as a predictor. Root-mean-square error (RMSE) describes model residuals, and Cohen's f^2 is a measure of effect size.

Outcome:	**Akaike Wt.**	**Significant predictors**	**R^2**	**R^2 with ID**	**RMSE**	**RMSE with ID**	**f^2**
Sociability	= Sex + Entry Age + 1st Test Tenure + Time Elapsed + (Sex * Entry Age) + (Sex * 1st Test Tenure) + (Sex * Time Elapsed)						
	0.83	Sex *** 1st Test Tenure *** Time Elapsed *** Sex-by-Entry Age Interaction*** Sex-by-1st Test Tenure Interaction *** Sex-by-Time Elapsed Interaction *	0.31	0.70	0.83	0.57	0.45
Cautiousness	= Sex + Entry Age + 1st Test Tenure + (Sex * Entry Age)						
	0.25	Sex*** Entry Age*** 1st Test Tenure ***	0.18	0.68	0.91	0.60	0.22
Aggressiveness	= Sex + Entry Age + Time Elapsed + (Sex * Time Elapsed)						
	0.17	Entry Age ***	0.05	0.51	0.97	0.77	0.05
Fearfulness		= Sex + Time Elapsed + (Sex * Time Elapsed)					
	0.36	Time Elapsed *** Sex-by-Time Elapsed Interaction ***	0.03	0.65	0.98	0.64	0.04
*$p < .05$, **$p < .01$, ***$p < .001$							

Table 4: Unstandardized coefficients for predictors in the best fitting models of population-wide personality change.

Component	Predictor	Estimate	Uncond. SE
Sociability	Sex	-0.29	0.05
	Entry age	-0.07	0.02
	1st test tenure	0.06	0.02
	Time elapsed	0.19	0.03
	Sex * Entry age	0.08	0.02
	Sex * 1st test tenure	0.09	0.02
	Sex * Time elapsed	0.08	0.03
Cautiousness	Sex	0.27	0.06
	Entry age	-0.10	0.02
	1st test tenure	-0.15	0.02
	Sex * Entry age	0.04	0.02
Aggressiveness	Sex	0.08	0.05
	Entry age	0.05	0.01
Fearfulness	Sex	0.08	0.05
	Time elapsed	0.12	0.04
	Sex * Time elapsed	0.13	0.04

Entry age, 1st test tenure and time elapsed are measured in years, and mean-centered. Sex is measured with females as 1 and males as -1.

Description of variables included in PCA analysis. Coding is described as either "Point" (behavior recorded once a minute at a specified time; instantaneous sampling) or "One-Zero" (tester recorded whether or not behavior occurred any time in past minute). Point-score variables involved, (1) cage position, (2) degree and type of responsiveness to the observer or other monkeys, (3) posture/locomotion, and (4) facial/vocal expression. These categories were scored on the 15-, 30-, 45-, or 60-second interval, respectively. Related point score variables are grouped together by number in the Coding column. All scores range from 0-4. The behaviors listed in bold significantly contributed to personality components used in this study (see Sussman et al., 2013 for details on inclusion criteria).

Behavior	Description	Coding
Back of cage	Monkey was positioned in back 1/3 of cage.	Point (1)
Front of cage	Monkey was positioned in front 1/3 of cage.	Point (1)
Middle of cage	Monkey was positioned in middle 1/3 of cage.	Point (1)
Attention to others	Attention of monkey, as evaluated by eyes, was focused on conspecifics.	Point (2)
Attention to tester	Attention of monkey, as evaluated by eyes, was focused on observer.	Point (2)
Back to tester	Monkey was turned with back to tester	Point (2)
Ignore	Attention of monkey, as evaluated by eyes was ignoring the observer and conspecifics.	Point (2)
Lean or approach tester	Monkey leans or moves toward the observer, but does not lunge.	Point (2)
Lunge	Monkey lunged at observer.	Point (2)
Move away	Monkey leans or moves away from observer.	Point (2)

Crouch	Monkey was crouched.	Point (3)
Lie	Monkey was lying on floor of cage.	Point (3)
Locomote	Monkey was moving around cage.	Point (3)
Sit	Monkey was sitting on floor of the cage.	Point (3)
Stand on two legs	Monkey was standing on hind legs (without locomoting).	Point (3)
Stand on three or four legs	Monkey was standing on three or four legs (without locomoting)	Point (3)
Grunt	Monkey grunted.	Point (4)
LEN to other monkey	Monkey produced "LEN" facial expression (Lips forward, Ears back, Neck extended), directed toward a conspecific.	Point (4)
LEN to tester	Monkey produced LEN toward observer.	Point (4)
Lipsmack to other monkey	Monkey showed lipsmack toward conspecific.	Point (4)
Lipsmack to tester	Monkey showed lipsmack toward observer.	Point (4)
Open mouth	Monkey showed open mouth.	Point (4)
Quiet face	Monkey showed quiet face (mouth closed, no lip or jaw movement).	Point (4)
Grimace	Whether or not monkey showed fear grimace to observer.	One-Zero
Shriek	Whether or not monkey shrieked.	One-Zero
Grind teeth	Whether or not monkey ground teeth, without eating.	One-Zero
Threat	Whether or not monkey made a threat display (open mouth threat, lunge, and/or stamping foot with eye contact) to observer or conspecifics	One-Zero
Reach	Whether or not monkey reached out of the cage.	One-Zero
Avert gaze	Whether or not monkey averted gaze from observer (direction of gaze moved away from observer as observer directed eyes toward monkey).	One-Zero

Food	Whether or not monkey consumed food.	One-Zero
		One-Zero
Groom	Whether or not monkey self-groomed.	One-Zero
Object	Whether or not monkey manipulated an object on or in the cage	One-Zero
Scratch	Whether or not monkey scratched self.	One-Zero
Shake body	Whether or not monkey shook its body.	One-Zero
Shake cage	Whether or not monkey shook cage.	One-Zero
Urinate	Whether or not monkey urinated.	One-Zero
Yawn	Whether or not monkey yawned.	One-Zero