Finding good acoustic features for parrot vocalizations: The feature generation approach

Nicolas Giret[a)]
Laboratoire d'Ethologie et Cognition Comparées, Université Paris Ouest Nanterre La Défense, 200 avenue de la république, 92000 Nanterre, France

Pierre Roy
Sony Computer Science Laboratory, 6, rue Amyot, 75005 Paris, France

Aurélie Albert
Laboratoire d'Ethologie et Cognition Comparées, Université Paris Ouest Nanterre La Défense, 200 avenue de la république, 92000 Nanterre, France

François Pachet
Sony Computer Science Laboratory, 6, rue Amyot, 75005 Paris, France

Michel Kreutzer and Dalila Bovet
Laboratoire d'Ethologie et Cognition Comparées, Université Paris Ouest Nanterre La Défense, 200 avenue de la république, 92000 Nanterre, France

(Received 11 August 2009; revised 24 November 2010; accepted 4 December 2010)

A crucial step in the understanding of vocal behavior of birds is to be able to classify calls in the repertoire into meaningful types. Methods developed to this aim are limited either because of human subjectivity or because of methodological issues. The present study investigated whether a feature generation system could categorize vocalizations of a bird species automatically and effectively. This procedure was applied to vocalizations of African gray parrots, known for their capacity to reproduce almost any sound of their environment. Outcomes of the feature generation approach agreed well with a much more labor-intensive process of a human expert classifying based on spectrographic representation, while clearly out-performing other automated methods. The method brings significant improvements in precision over commonly used bioacoustical analyses. As such, the method enlarges the scope of automated, acoustics-based sound classification.

© *2011 Acoustical Society of America*. [DOI: 10.1121/1.3531953]

PACS number(s): 43.80.Ka, 43.80.Ev [MJO] Pages: 1089–1099

I. INTRODUCTION

Bioacoustic research provides fundamental information for understanding communication systems in the wild. A crucial step in this process is to identify acoustic characteristics of a signal to describe the potential meaning encoded by the sender and/or perceived by the recipient. Birdsong has been at the center of a wide range of animal behavior studies since the mid-20th century. The study of their acoustic features brought a better understanding of what information is encoded in songs (see Marler, 2004). Some songbird species, such as domestic canaries, produce complex songs with specific acoustic features that are preferred by females (Vallet and Kreutzer, 1995). Understanding if and how objective characteristics of bird vocalizations (sources) can be mapped to objective characteristics of the contexts of production (targets) is thus central in the study of bird behavior. However, this task, applied to the case of parrots, turns out to be extremely difficult (Cruickshank *et al.*, 1993). A simplification of this problem is instead to consider as a target a categorization produced by human experts. Such an endeavor can be useful for animal behavior studies if one can then interpret human categorizations in terms of specific contexts or other ethological information. The purpose of the current work was to design a sound classification method based on automated extraction of acoustic features, validating its performance against categorization outcomes provided by a human expert. This method was applied to vocalizations of parrots and was compared to other commonly used sound analysis methods.

A. Analysis methods of animal vocalizations

Categorizations produced by humans are traditionally based on spectrographic representations of the acoustic signal (Thorpe, 1954). Three analysis methods commonly used for this purpose are visual comparison of the spectrograms by a trained experimenter (Janik, 1999; Bloomfield *et al.*, 2004), spectrographic cross-correlation (Khanna *et al.*, 1997; Janik, 1999; Cortopassi and Bradbury, 2000; Baker, 2003), and extraction of acoustic features in the signal (Nowicki and Nelson, 1990; Charrier *et al.*, 2004; Draganoiu *et al.*, 2006).

Indeed, the human eye is a complex integrator system that can identify subtle modifications in the signal. However, judgments done by eye are inherently subjective, so these

[a)]Author to whom correspondence should be addressed. Current address: Institut für Neuroinformatik, Universität Zürich/ETH, Winterthurerstrasse 190, CH-8057 Zurich, Switzerland. Electronic mail: ngiret@u-paris10.fr

Author's complimentary copy

classifications cannot always be replicated, especially with large datasets. Moreover, studies that rely on visual inspection generally do not make explicit which criteria are used to establish the classification. Personal criteria evolve during the classification process (see Jones *et al.*, 2001). One way to limit this subjectivity is to have several observers perform the same classification to determine interobserver reliability (Janik, 1999). However, interobserver reliability is itself influenced by several factors that can lead to overestimation or underestimation of the actual reliability (Jones *et al.*, 2001).

Spectrographic cross-correlation, first proposed by Clark *et al.* (1987), produces a similarity matrix between pairs of signals in a database. Each similarity value is measured as the peak of the correlation between the spectrograms of the two signals. A number of software packages, such as AVISOFT-SASLAB PRO (Avisoft-Bioacoustics, Germany) and RAVEN (Cornell Lab of Ornithology, USA), include this method of analysis. Spectrographic cross-correlation is not biased by the subjectivity of the human observer. However, this method suffers from several limitations, mainly that correlation values are sensitive to the fast Fourier transformation (FFT) frame length used for spectrogram generation (Khanna *et al.*, 1997; Cortopassi and Bradbury, 2000).

A third common approach is to extract features directly from the acoustic signal. These features typically represent standard and mathematically well-defined properties of audio signals, such as frequency bandwidth, number of harmonics, duration, amplitude, or the waveform's zero crossing rate (ZCR). The main advantages are that these methods avoid experimenter bias and can be very accurate (Anderson *et al.*, 1996; Tchernichovski *et al.*, 2000). For example, Tchernichovski *et al.* demonstrated that four simple, mono-dimensional acoustic features were sufficient to model the similarity among songs of zebra finch (*Taeniopygia guttata*). However, these methods are also limited by the scope of the feature sets used, which is not always relevant to the specific categorization problem at hand. In the same vein, Schrader and Hammerschmidt (1997) proposed a number of typical features to characterize animal sounds. These features range from formant-like structures, frequency peaks, distribution of spectral energy, and sound structure. However, as noted by the authors themselves, it is often not possible to determine what are the "right features" to characterize animal sounds *a priori*.

Studies comparing the accuracy of these various acoustical methods have produced inconsistent results (Baker, 2003). Khanna *et al.* (1997) demonstrated limitations of the spectrographic cross-correlation methods but were contradicted by Cortopassi and Bradbury (2000), who demonstrated that this method can be reliable when associated with principal coordinates analysis. Janik (1999) demonstrated that judging by eye can be a more efficient system of analysis than pitch extraction with spaced frames (McCowan, 1995) or cross-correlation associated with cluster analyses (Khanna *et al.*, 1997). Thus, determining which classification method is the most efficient remains an open problem.

B. The feature generation approach

This article proposes a new method of classification that associates classification by eye with automatic feature measurement to overcome the limitations of previous approaches based on predefined feature sets. More precisely, visual judgments are used to build a training database, also referred to as subjective standard, which is then used to train and test a classifier using a supervised classification technique.

This approach is termed "feature generation." Acoustic features are first extracted from the audio signals, and a model of the various signal classes is built according to these features during the training phase. In the evaluation phase, classifier predictions are then checked against human classification of the rest of the database (the subjective standard) to assess the model's performance. Eventually, these classifiers can be applied to new acoustic signals.

This approach has been found useful in several areas of research, such as data mining (Mjolsness and Decoste, 2001) or ecological modeling (Stockwell, 2006). The proposed approach differs from these previous ones in that it consists in exploring a huge feature space (potentially several billion), designed to incorporate typical and well-known features such as those described by Schrader and Hammerschmidt (1997), as well as many others. The feature generator is based on a genetic programming algorithm (Koza, 1992). Features such as the ones proposed by Schrader and Hammerschmidt (1997) are all within the reach of feature generation, meaning that the system can find these features or approximations thereof. In fact, feature generation has been shown to outperform approaches based on standard features in the context of animal vocalization (Molnár *et al.*, 2008).

C. African gray parrot vocal behavior

More precisely, our goal here was to investigate whether an artificial-intelligence feature generation system can identify features that can be used to classify African gray parrot (*Psittacus erithacus*) vocalizations as categorized by an expert observer. African gray parrots show complex vocal behavior (Cruickshank *et al.*, 1993; Pepperberg, 2007), are known for their ability to imitate human language (Pepperberg, 1999; Giret *et al.*, 2010), and exhibit complex cognitive skills (Pepperberg, 2006; Al Aïn *et al.*, 2009; Giret *et al.*, 2009). They produce many types of vocalizations with a diverse range of frequencies, amplitudes, and durations. Parrots in general can also imitate a wide range of sounds from their surroundings (Cruickshank *et al.*, 1993; Farabaugh *et al.*, 1994; Farabaugh and Dooling, 1996).

The description of the vocal repertoire of three African gray parrots was based on the visual comparison of spectrographic representations of the calls by a human expert (N.G.) highly familiar with the calls of these birds. This manual classification process raised two issues. First, it was a long and tedious task, which required the expert to listen to and analyze each and every call. Second, the expert's knowledge and perception of the calls were affected by the classification process itself. That evolution led not only to a number of refinements made underway but also three complete reorganizations of the dataset. Both factors make this process difficult to repeat.

D. Study goals

In this article, supervised classification was used to overcome the limitations of the manual approach. Using

Author's complimentary copy

supervised classification reduced the workload by requiring only a small dataset to be classified by the expert, the rest being taken care of automatically by a classifier. Working by eye on a small dataset prevented or at least reduced the risk of committing errors of reliability and of validity. In addition, supervised classification yielded repeatable results, as it depended less on human expertise.

Two experiments were carried out. In a preliminary experiment, the performance of supervised classification was assessed on the full dataset (52 076 calls in 168 types), based on mel frequency cepstral coefficients (MFCCs; Davis and Mermelstein, 1980). Here, performance was poor. In the main experiment, the study was restricted to a limited set of five call types. Feature generation was used to produce a feature set from a subset of these data, and this set provided better results than either MFCCs or spectrogram cross-correlation. This technique also proved successful with a larger dataset. A final experiment then showed that this approach could be used in practice on the whole dataset by using two sets of features.

II. GENERAL METHODS

A. Subjects

Three captive hand-reared African gray parrots were involved in this experiment: two males, Léo and Shango, and one female, Zoé. Subjects were housed together in an aviary of 340 cm × 330 cm × 300 cm in which several perch structures and toys were provided. Water and parrot pellets were available *ad libitum*. Parrots were fed fresh fruits, vegetables, and parrot formula daily.

B. Apparatus

The parrots were recorded regularly from when first arriving at the laboratory, including a period of 4 yr for Léo and Zoé (2003 to 2007) and 2 yr for Shango (2005 to 2007). Recordings were made with a Microphone EMU4534 (LEM Industries, Argenteuil, France) and a MD S301 (Sharp Electronics, Villepinte, France) (BL) minidisc from September 2003 to September 2005 and then converted to 22 kHz (16 bits) *wav* format. Further recordings were made with a Microphone MD 21 U (Sennheiser electronic GmbH & Co, Wedemark, Germany) and a DAT Marantz PMD670 recorder in 22 kHz (16 bits) *wav* format. Although Mini-Disc recorders that use data compression algorithms may yield a loss of harmonic contents at the upper margin of human hearing (Budney and Grotke, 1997), no difference was observed between Mini-Disc and DAT recordings after close inspection of similar calls.

Spectrographic analyses were carried out with AVISOFT SAS-LAB PRO version 4.40. MATLAB version 14 service pack two was used to perform pre-processing on each call. Statistical analyses were conducted with SIGMASTAT version 3.5.

C. Procedures

The parrots were recorded while spontaneously vocalizing (95% of recorded calls) and during elicited recording sessions (5% of calls) in which parrots were either being carried, encouraged to take off, or left alone in the aviary or in which food, toy, or fake predators were shown to the birds. Birds received nine sessions in each situation, except for the fake predator condition in which only four sessions occurred (see Giret *et al.*, 2010). Recordings of spontaneous vocalizations were made at different times of the day and in various areas of the laboratory, such as the aviary and office. Continuous recordings were segmented with AVISOFT-SASLAB PRO by a human expert (N.G.) to extract the individual vocalizations (52 076 calls) used in the two experiments. This expert then classified calls by type based on visual inspection of spectrographic representation. A new call type was defined when at least five instances of that type could be identified. These were instances that exhibited similar acoustic features, such as fundamental frequency, pitch, and minimum and maximum frequencies. Overall, 168 call types were identified.

Further pre-processing was required for automated classification. DC-offset was removed from all calls, which were then amplitude-normalized to 95% of the available 16-bit range. In this study, the identity of the three parrots was not taken into account in order to have a larger sample.

The performance of the classifiers used in this study was typically represented by the confusion matrix between the call types identified by the expert and the call types computed by the classifier. This matrix was in turn used to compute an F-measure for each call type

$$F = \frac{2 \times \text{precision} \times \text{recall}}{\text{precision} + \text{recall}} \qquad (1)$$

Precision was calculated through the following formula:

$$\text{precision} = \frac{N_{\text{tp}}}{N_{\text{tp}} + N_{\text{fp}}} \qquad (2)$$

Recall was determined with the following formula:

$$\text{recall} = \frac{N_{\text{tp}}}{N_{\text{tp}} + N_{\text{fn}}} \qquad (3)$$

For a given vocal category, "tp" (true positive) was the number of elements shared by both classifications (diagonal of the confusion matrix), "fp" (false positive) was the number of elements of another vocal category classified in the given category by the algorithm, and "fn" (false negative) was the number of elements of the given vocal category classified in another vocal category by the algorithm. An F-measure value of 1 (or 100%) implied that the computed classification was identical to the expert classification. The F-measure decreased as a function of increasing fp and fn classification results, with values below 0.3 indicating very poor classification.

III. PRELIMINARY EXPERIMENT: MFCC ANALYSIS ON THE WHOLE DATASET

In this experiment, the performance of the supervised classification approach using standard spectral features was

Author's complimentary copy

evaluated. Spectral envelopes were estimated using MFCC. The cepstrum is the inverse Fourier transform of the log-spectrum of the signal. The mel-cepstrum is computed after a non-linear frequency warping onto the perceptual mel frequency scale (Foote, 1997). The coefficients are called MFCCs. Cepstrum coefficients provide a low-dimensional, smoothed version of the log-spectrum and thus are a good and compact representation of the spectral shape. They are considered as an efficient representation of timbre information, thereby subsuming individual features such as spectral centroid or spectral skewness. MFCCs are widely used as acoustic features for speech recognition (Davis and Mermelstein, 1980) and have also proved useful in music classification (Foote, 1997), musical instrument recognition (Eronen and Klapuri, 2000), as well as for the classification of elephant vocalizations (Clemins *et al*., 2005; Clemins and Johnson, 2006).

For the current work, MFCCs were of interest as a reasonable and generic representation of the spectral content of the signal and were thus used as a reference point for classification performance. Each vocal event was represented by an audio signal. Acoustic features were extracted on each audio signal, and these features were then used to classify the calls using a C4.5 machine-learning algorithm (Quinlan, 1993). The C4.5 algorithm uses the fact that each feature of the data can be used to make a decision by splitting the data into smaller subsets. C4.5 creates a tree-like graph from a set of pre-categorized dataset (the training dataset). This graph is called a decision tree. Each node of the decision tree represents a decision based on the value of a feature of the data and each leaf of the decision tree corresponds to a category in the training dataset, i.e., in our case, each leaf of the decision tree is a call type. The decision tree can be used to classify new data according to the values of the features of that data. The decision tree classifier was trained and tested on two separate parts of the sound database.

A. Method

Ten MFCC values were extracted for each call (52 076 vocalizations, 168 call types). The performance of these features was then assessed with a decision tree algorithm (J48 in the WEKA software, an implementation of Quinlan's C4.5; Quinlan, 1993). Classification performance was evaluated by training the algorithm on a set of 10% of the vocalizations (5207 total) and by testing on the remaining 90% (46 869 total).

B. Results

Overall, 24.9% (11 686) of the calls were correctly classified, while 75.1% (35 183) were not. As shown in Fig. 1, *F*-measures for individual call types were not good. For example, *F*-measures for 155 call types were less than 0.3 (very bad performance), while 13 call types had *F*-measures ranging between 0.3 and 0.7. No call type had an *F*-measure superior to 0.7. In summary, no call type was well modeled by this classifier.

C. Discussion

Figure 1 clearly shows that the classification obtained differed from that of the human expert. Some of this difference may reflect that the expert misclassified some of the training samples. Indeed, the expert estimated that about 30% of the calls were very challenging to classify. This difficulty reflects that some of the 168 call types were hard to distinguish, even for a highly trained observer. In addition, gray parrots produce a large variety of vocalizations, even within a call type. They can modify the length, intensity, and/or fundamental frequency of the calls (Fernández-Juricic *et al*., 1998). Those variations may not be captured by MFCCs, which represent spectral properties of sound. More surprisingly, the accuracy was low even for the call types that had clear perceptual characteristics for the expert. This failure can be partially explained by the aforementioned reason that parrots can modify many different acoustic parameters even within a call type. The limitation of MFCCs in this context is discussed more in depth in the next experiment. The poor performance obtained in this preliminary experiment may also be explained by gaps in the training data and compromised sound quality among the samples. First, 78 call types had less than 200 training samples, and the size of the training dataset was central in supervised classification. Second, recording conditions were often compromised by uncontrollable background noise or by birds flying away from the microphone. Many samples were thus of poor sound quality, making them difficult to classify correctly.

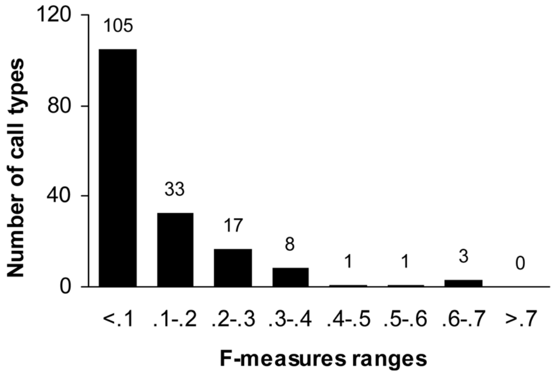

FIG. 1. Numbers of call types are represented according to bins of *F*-measures for the preliminary experiment. Overall, the *F*-measures were low, meaning that the classifier (MFCC) was not efficient on the whole dataset.

In the next experiment, the dataset was reduced to five call types in order to keep the human error-rate to a minimum. The performance of supervised classification of the five call-type problem was assessed and was shown to be much better than on the 168 call-type problem. Moreover, this experiment contrasted two different feature sets, namely MFCCs versus problem-specific features (Pachet and Roy, 2009).

IV. MAIN EXPERIMENT: A FEATURE GENERATION APPROACH

In this experiment, five call types were selected from the parrot repertoire based on being specific to particular contexts (Giret *et al*., 2010) and represented by more than 200 samples of good quality. It was necessary to use a new set of acoustic features to improve the performance of the supervised classification. To do so, one approach would be to look for features that are commonly used in signal-processing

Author's complimentary copy

studies. As noted in the Introduction, however, results of this approach have been disappointing in previous applications. Another possibility, which usually works better, is to compute spectrographic cross-correlation (Cortopassi and Bradbury, 2000) or to design specific features using expertise in signal processing and knowledge of the problem at hand.

Pachet and Roy (2009) introduced a method to automatically create efficient acoustic features for specific audio-classification problems. This approach has been shown successful when applied to dog vocalizations (Molnár *et al.*, 2008). Therefore, the approach by Pachet and Roy (2009) was used to generate features for the current five call-type problem. Results were compared to those obtained with the MFCCs feature set and to those obtained through the classical spectrographic cross-correlation approach.

A. Methods

1. Call types

The call sample consisted of five call types selected to be specific to particular contexts (see Figs. 2 and 3). Type C1 calls were mainly produced as a "protest" either against a conspecific or in response to a human handling the bird. Type C57 calls were mainly produced in situations of fear and were elicited by placing eagle or snake predator models in the aviary. Type C77 calls were produced by a bird when it was alone, just after the experimenter left the aviary. Type C106 calls were mainly produced when the subject wanted "something," such as food or toys. Type C113 calls were produced when the subject was excited and were associated with the behavior of turning and then flying off suddenly. The resulting database included a total of 2971 vocalizations: 691 C1, 1277 C57, 492 C77, 288 C106, and 223 C113 calls.

2. Spectrographic cross-correlation

Spectrographic cross-correlations were carried out with AVISOFT CORRELATOR v2.2. One exemplar of each call type was randomly selected as a reference to which each other call was cross-correlated through the "aligning mode" of AVISOFT CORRELATOR. Each cross-correlation computation provided a peak correlation value between -1 and 1. Values of 0, 1, and -1 indicate that the two calls compared are not correlated, identical, and inverse, respectively. In order to

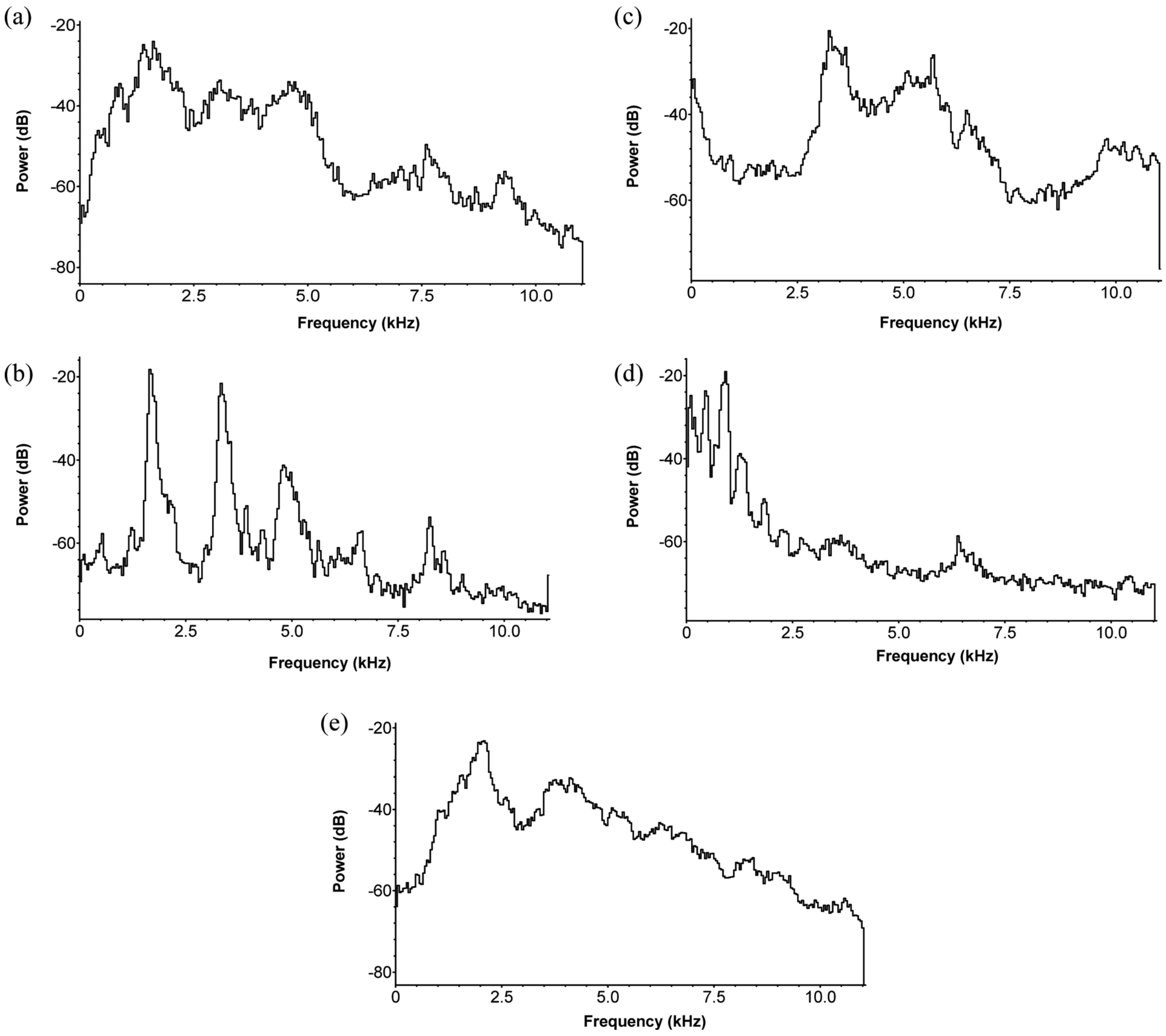

FIG. 2. Mean power spectra are shown for (a) C1, (b) C57, (c) C77, (d) C106, and (e) C113.

Author's complimentary copy

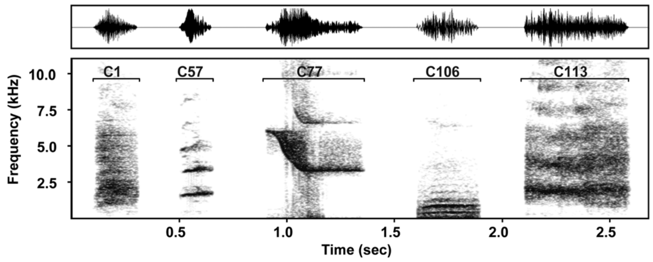

FIG. 3. Oscillograms (top) and spectrographic representations (bottom) for C1, C57, C77, C106, and C113. Spectrograms were calculated using 512-point Hamming windows, a 22 kHz sampling frequency, and 16-bit amplitude sampling.

compare these results with those obtained with feature generation and with MFCC-based classifiers, the highest correlation value was selected as an indicator of the call type in which the call should have been classified. Since there were five call types, five different correlation values were obtained for each call. For example, a call belonging to the call type C1 was cross-correlated with the five reference calls. If the highest correlation value was obtained when cross-correlating this call with the reference of the call type C57, that means that, based on the spectrographic cross-correlation, this call would be classified in the call type C57. F-measures were computed as before.

3. Feature generation

Following identification of the smaller, five call-type dataset, it was divided in two. A total of 500 samples were designated as a feature generation database (FGD), which was used to generate features with extractor discovery system (EDS). The remaining 2471 samples made up a feature evaluation database (FED), which was used to test the efficacy of those features.

The FGD was used to generate problem-specific features using the EDS system (Pachet and Roy, 2009). EDS exploits a library of basic signal-processing operators such as root-mean squared (rms, which measures the power of the waveform), ZCR, and FFT. These basic operators are then assembled using an evolutionary algorithm to produce arbitrary complex mathematical functions of the signal. The function space explored by EDS is huge, so EDS exploits a number of heuristics to prune the search tree in order to quickly identify "interesting" features (Pachet and Roy, 2009). The fitness used to assess feature performance is based on the so-called wrapper approach to feature selection (Guyon and Elisseeff, 2003), meaning that the fitness of a feature is the performance of the associated classifier trained on-the-fly with this feature only. The search process iterates until efficient features have been found. Eventually, the best features are selected using an *ad hoc* feature selection algorithm (Pachet and Roy, 2009).

4. Feature evaluation

The performance of the features generated by EDS was compared to the performance of MFCCs. Evaluations were carried out on FED, to make sure no evaluation sample was used for feature generation, as FED and FGD did not overlap. Note that, as in the preliminary experiment, the classifiers used for this experiment were decision trees trained using a J48 algorithm.

The first step (feature selection) was to select the features to evaluate. This feature selection (see Table I and Appendix) was performed using the correlation-based "Greedy Stepwise" (Hall, 2000) subset search algorithm provided by the WEKA software. The second step (feature extraction) consisted of computing the value of each selected feature for every sample in FED. Those values were then used in the two remaining steps to train and evaluate the classifier performance. More precisely, during the third step (classifier training), a decision tree algorithm was trained on the feature values computed on the 10% samples randomly chosen from FED. In the final, fourth step, the performance of the classifier was evaluated on the FED samples, making up 90% of the entire, five call-type database. Note that classifier training and performance (steps three and four) had to be repeated several times in order to reduce the variability in classifier performance, as the training and testing subsets of the FED created were randomly selected. Fifty repetitions of this training-testing process was selected as an optimal number after a comparison of the F-measures obtained when repeating the evaluation 1, 2, 5, 10, 20, 50, 100, 1000, and 10 000 times.

Based on the evaluation, either 1, 2, 4, 5, 10-EDS features or 20-EDS features were selected and compared to 10- or 20-MFCC features. EDS versus MFCC feature performance was evaluated using Kruskal-Wallis one-way analysis of variance on ranks to compare the number of correctly classified calls for each feature set, followed by *post-hoc* analyses (Student-Newman-Keuls, SNK).

5. Two-stage classification

The feature generation approach produced a classifier designed to be efficient on the five selected call types only. To solve the problem of classifying arbitrary calls into the five reference categories, a two-stage classification was performed. Two abstract categories were defined. Category "A" consisted of calls of one of the five call types C1, C57, C77, C106, and C113. Category "B" consisted of the 163 remaining call types. The first stage consisted of categorizing calls into categories A and B. This first-stage classifier, called CL1, was trained on a database consisting of the 500 samples from FGD, labeled A, and 500 samples labeled B randomly selected from the rest of the corpus. In the second stage, a decision tree classifier was used to classify samples categorized A by the first stage classifier into C1, C57, C77, C106, or C113. For the second-stage classifier, the classifier trained on 20-EDS, called CL2-EDS, was used and its performance was compared to that of the classifier trained on 20-MFCC, called CL2-MFCC.

B. Results

In the preliminary experiment, MFCCs failed to capture the variability of calls within a given type. Specifically, the preliminary experiment yielded F-measures of 0.07, 0.61, 0.36, 0.13, and 0.19 for C1, C57, C77, C106, and C113, respectively.

Author's complimentary copy

TABLE I. Features composing each EDS feature set. In the feature expressions, "*x*" represents the input acoustic signal. The mathematical composition operation is implicitly represented by the parentheses. Each label denotes an operator. Most of them are explained in detail in Pachet and Roy (2009). See Appendix for descriptions of operators not included in Pachet and Roy (2009).

Feature set	Features
1-EDS	1: Abs(Norm(PitchPraat(Blackman(Integration(VMeanNormalization(Arcsin(x)))))))
2-EDS	1 + 2: Sqrt(Abs(Iqr(Centroid(FilterBank(Abs(x)_27.0)))))
3-EDS	1 + 2 + 3: Variance(PitchPraat(Abs(Bartlett(x))))
4-EDS	1 + 2 + 3 + 4: ZCR(Skewness(SplitOverlap(x_32.0_0.5)))
5-EDS	1 + 2 + 3 + 4 + 5: Sqrt(Centroid(Blackman(SpectralKurtosis(MelFilterBank(x_10.0)))))
10-EDS	1 + 2 + 3 + 4 + 5 + 6: Bandwidth(Pitch(Blackman(MelFilterBank(x_10.0)))_50.0) 7: Power(Abs(Mean(LSTER(SplitOverlap(Abs(Hann(x))_32.0_0.3)_0.15_50.0_0.5)))_2.89) 8: Sqrt(Abs(Range(Centroid(MelFilterBank(Normalize(x)_27.0))))) 9: MaxPos(rms(FilterBank(x_10.0))) 10: Crrm(HarmonicSpectralCentroid(HMeanNormalization(FilterBank(x_10.0)))_27.0_3.0)
20-EDS	1 + 2 + 3 + 4 + 5 + 6 + 7 + 8 + 9 + 10 + 11: Skewness(Skewness(SplitOverlap(x_32.0_0.5))) 12: Max(SpectralSpread(MelFilterBank(Triangle(Normalize(x))_10.0))) 13: Sqrt(Max(PitchPraat(Hann(Square(Triangle(PointProcessPraat(x))))))) 14: rms(HarmonicSpectralCentroid(FilterBank(x_10.0))) 15: Sqrt(Abs(Max(Centroid(FilterBank(Arcsin(Hamming(x))_27.0))))) 16: Centroid(SpectralDecrease(Triangle(FilterBank(x_10.0)))) 17: Bandwidth(HNorm(FilterBank(x_10.0))_50.0) 18: SpectralDecrease(x) 19: Sqrt(HarmonicSpectralCentroid(Autocorrelation(x))) 20: rms(SpectralCentroid(Split(x_32.0)))

1. *Spectrographic cross-correlation*

Each call of the database was compared to a reference call of each call type (C1, C57, C77, C106, and C113). *F*-measures were computed to identify correctly or incorrectly classified calls and provided the following results: 0.59, 0.59, 0.90, 0.93, and 0.19 for C1, C57, C77, C106, and C113, respectively.

2. *Feature generation*

EDS generated explicitly about 11 000 features, 33 of which were selected automatically using a feature selection algorithm based on feature fitness and syntactic feature similarity (Pachet and Roy, 2009). Those 33 features formed a feature set labeled 33-EDS.

3. *Feature evaluation*

a. Number of repetitions needed. The minimum number of repetitions of the training-testing process needed to obtain stable *F*-measures was estimated first. Different numbers of repetition were used: 1, 2, 5, 10, 20, 50, 100, 1000, and 10 000. *F*-measures became stable after 50 repetitions (see Table II), which was therefore the value used for further statistical performance comparisons (see below).

TABLE II. Mean percentage *F*-measures obtained for each feature set. The training-testing process was repeated several times in order to reduce the variability in classifier performance: 50 repetitions (or more) ensure almost constant mean *F*-measures.

Number of repetitions	1-EDS (%)	2-EDS (%)	3-EDS (%)	4-EDS (%)	5-EDS (%)	10-EDS (%)	20-EDS (%)	10-MFCC (%)	20-MFCC (%)
1	54.30	71.01	73.54	83.78	84.14	83.48	89.14	74.97	77.74
2	55.30	71.29	73.26	81.96	80.05	83.38	84.78	77.86	76.14
5	55.63	70.08	74.44	79.13	83.11	82.75	85.53	75.01	75.59
10	54.42	70.54	74.07	79.24	81.48	83.26	84.83	76.89	76.82
20	52.33	69.84	73.98	80.35	81.89	83.56	84.61	77.40	76.31
50	**53.70**	**70.29**	**74.09**	**80.89**	**81.30**	**83.95**	**85.29**	**76.74**	**76.41**
100	53.54	70.14	74.10	79.93	82.03	83.95	85.17	76.82	76.77
1000	53.57	70.17	73.80	80.01	82.08	84.13	84.95	77.09	76.55
10 000	53.56	70.12	73.97	79.97	81.95	84.13	84.99	77.06	76.62

Author's complimentary copy

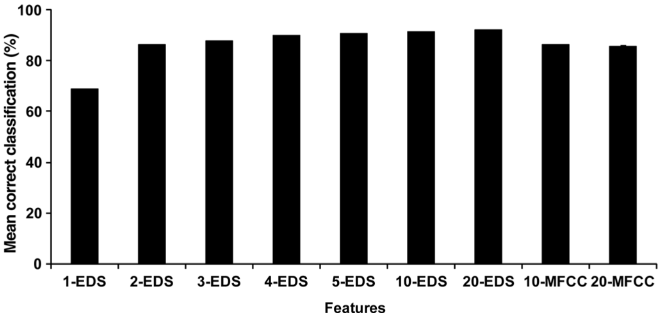

FIG. 4. Mean percentage correct classification for the five calls selected is represented according to the feature sets. Using just four EDS features provided better results than with 10 or even 20 MFCC values, and the results were even better with more EDS features.

b. Evaluation of the features on FED. The overall efficiency of each feature set was evaluated by comparing the number of correctly classified calls for each feature set (with Kruskal-Wallis analyses of variance, see Fig. 4). There was an effect of the feature set used on the number of correctly classified calls ($H_8 = 380.69$, $p < 0.001$). Subsequent SNK *post-hoc* analyses revealed that classification performance using the 2-EDS feature set was not significantly different from that of the 10-MFCC ($q = 3.30$ ns) or 20-MFCC ($q = 0.23$ ns) sets, and that performance with the 10-MFCC set was not different from that of the 20-MFCC set ($q = 4.71$ ns). All the other comparisons were significantly different ($p < 0.05$). Thus, with only four EDS features, the results were better than those obtained with 10 and even 20 MFCC coefficients. Results were even better with 5, 10, and 20 EDS features. The next step was to investigate which particular call types were better-classified due to using EDS features. The number of correctly classified calls obtained independently was compared for each call type and for each feature set (with Kruskal-Wallis analyses of variance, see Table III). Results showed differences for all five call types, including C1 ($H_8 = 329.22$, $p < 0.001$), C57 ($H_8 = 98.01$, $p < 0.001$), C77 ($H_8 = 137.14$, $p < 0.001$), C106 ($H_8 = 167.58$, $p < 0.001$), and C113 ($H_8 = 325.96$, $p < 0.001$). Overall, 10 or 20 EDS features gave the best results for each call type, except in the case of C106 calls (SNK *post-hoc* analyses). Particularly, EDS features provided the greatest improvement compared to the MFCC values for the call types C1 and C113.

4. Two-stage classification

The first stage of the two-stage classification process consisted of sorting calls into categories A and B. Various feature sets were tried to train CL1, and the best results were obtained with 20-EDS (see Appendix). The confusion matrix of CL1 trained with 20-EDS showed that more calls were correctly classified (2188 calls A classified in A and 1956 calls B classified in B) than incorrectly (515 calls A in B and 283 calls B in A). For the second-stage classification, the classification of each call, according to its type, was assessed either with a classifier trained on 20-EDS (CL2-EDS) or with 20-MFCC (CL2-MFCC). Table IV presents confusion matrices obtained with both classifiers. In the two-stage classification, some calls of type B were misclassified by CL1, i.e., they were categorized into class A. In order to compute confusion matrices for the combined classifiers, the classification distributions of calls of type B by the two second-stage classifiers were needed. As some calls of type B were misclassified by CL1 into class A they could then be further deliberately misclassified using CL2 (CL2-EDS or CL2-MFCC) into classes C1, C57, C77, C106, or C113. With those data, one can compute the confusion matrix of the two combined classifiers, using the simple Bayesian inference. Results are presented in Table V. The *F*-measures of the classifier trained with 20-EDS were systematically higher than that of the classifier trained with 20-MFCC. As for the single classifiers C2-EDS and C2MFCC, call types C1 and C113 were substantially better-classified with EDS features than with MFCCs.

The two-stage classifier trained with 20-EDS also outperformed the one-stage classifier trained with 20-MFCC. In other words, the performance improvement brought by EDS features largely compensated for the loss of performance due to errors in the identification of calls of type A versus type B.

C. Discussion

In this experiment, African gray parrot vocalizations were used to study the influence of features on the performance of supervised classification methods. The subset consisted of five call types drawn from various contexts. In this approach, the performance of standard features (MFCCs) and standard approach (spectrographic cross-correlation) was compared to that of feature generation.

TABLE III. Mean percentage correctly classified calls (±SE) for each feature set and call type.

	Call types				
Features	C1 as C1	C57 as C57	C77 as C77	C106 as C106	C113 as C113
1-EDS	15.63 (1.31)	90.51 (0.67)	97.64 (0.09)	83.43 (1.90)	1.53 (0.50)
2-EDS	84.44 (0.5)	94.76 (0.31)	94.95 (0.54)	73.68 (1.56)	7.53 (0.93)
3-EDS	88.68 (0.47)	94.09 (0.36)	94.78 (0.50)	81.95 (1.44)	11.89 (1.29)
4-EDS	90.02 (0.42)	93.32 (0.29)	97.16 (0.23)	88.46 (0.80)	38.25 (1.87)
5-EDS	90.83 (0.62)	94.33 (0.25)	97.19 (0.26)	87.83 (1.13)	34.89 (1.99)
10-EDS	92.12 (0.38)	94.22 (0.22)	97.26 (0.20)	91.82 (1.31)	43.62 (1.53)
20-EDS	91.47 (0.40)	95.01 (0.24)	97.44 (0.15)	93.89 (0.82)	47.53 (1.56)
10-MFCC	80.42 (0.74)	92.17 (0.37)	93.64 (0.48)	93.63 (0.60)	22.49 (1.38)
20-MFCC	78.31 (0.72)	91.70 (0.33)	94.14 (0.44)	94.64 (0.50)	23.64 (1.15)

Author's complimentary copy

TABLE IV. Confusion matrices providing the distribution of the calls according to their original call type after the classification process with CL2-EDS and CL2-MFCC. More calls were classified correctly with CL2EDS than with CL2MFCC. In both cases, calls of type B were misclassified (in C1, C57, C77, C106 or C113).

	CL2-EDS						CL2-MFCC					
Classified as ↓	C1	C57	C77	C106	C113	B	C1	C57	C77	C106	C113	B
C1	523	4	2	0	62	576	385	17	0	0	189	511
C57	26	1043	27	0	81	345	43	935	25	1	173	435
C77	3	6	382	1	0	572	0	7	367	3	15	447
C106	4	5	0	182	0	738	2	0	0	186	0	845
C113	18	2	1	0	99	240	32	7	0	0	84	233

Features obtained by feature generation proved more efficient than spectrographic cross-correlations or using MFCCs, regardless of the feature-set dimensions chosen. The gain in performance was due to two call types, C1 and C113, which were poorly distinguished by both other approaches. Problem-specific analytical features performed much better on this classification subtask.

C1 calls were mainly produced in protest situations, whereas C113 calls occurred when the birds were excited, mainly just before or during a quick flight. These calls had similar bandwidth frequency and overall shape but differed in terms of timbre and temporal sequencing. Specifically, C1 calls consisted of short repeated vocalizations, while C113 calls were longer, more raucous, non-repeated calls.

MFCCs captured timbre information from acoustic signals. However, they failed to distinguish C1 and C113. Automatically generated features better distinguished between C1 and C113 calls. Further study of the analytical features was not carried out as it was not relevant from an ethological standpoint.

V. GENERAL DISCUSSION

The present study described a supervised learning approach to classifying vocalizations of African gray parrots. A preliminary experiment was carried out on a large dataset, with 168 call types, using standard spectral features (MFCCs). The resulting performance was poor due to the inherent difficulty of the classification problem. The study then focused on a smaller and simpler problem with only five call types. This sub-problem was addressed with two feature-set techniques (MFCCs and automatically generated features), as well as a more common approach (spectrographic cross-correlations). As for the general problem, MFCCs lead to poor performance. Spectrographic cross-correlations provided almost perfect classification for two call types (C77 and C106) but very poor performance for others. The approach using automatically generated features generally outperformed other approaches. Importantly, this approach can also be applied when the nature of the sounds is not known.

A wealth of research has investigated the use of acoustic feature extraction for automated recognition of bird calls or songs, whether for species identification (Kogan and Margoliash, 1998; Härmä, 2003; Härmä and Somervuo, 2004; Chang-Hsing *et al.*, 2006; Tyagi *et al.*, 2006) or for syllable classification (Tchernichovski *et al.*, 2000). However, the performance of those approaches depends strongly on the feature set used. In other words, for a given classification problem, specific features must be previously selected from a standard set or devised from scratch. Therefore, the resulting classifiers do not necessarily generalize well to other classification problems.

In the method described here, no assumptions need to be made concerning the nature of the features before the classification. Instead, automatically generated features are produced that are well-adapted to the problem at hand before the training phase of the supervised classification begins. This more general approach can be used to address virtually any audio-classification problem for which a training set of samples for each category is available.

TABLE V. Confusion matrices providing the distribution of the calls according to their original call type and mean *F*-measures obtained for each call type after the classification process with CL1, either followed by CL2-EDS or by CL2-MFCC. More calls were classified correctly with CL2EDS (diagonal of the matrix: 3930 calls) than with CL2MFCC (diagonal of the matrix: 3689 calls). Overall, *F*-measures were greater with CL2EDS than with CL2MFCC.

	CL2-EDS						CL2-MFCC					
Classified as ↓	B	C1	C57	C77	C106	C113	B	C1	C57	C77	C106	C113
B	1956	120	72	119	154	50	1956	107	91	93	176	49
C1	68	463	4	2	0	55	68	341	15	0	0	167
C57	135	23	924	24	0	72	135	38	828	22	1	153
C77	45	3	5	338	1	0	45	0	6	325	3	13
C106	22	4	4	0	161	0	22	2	0	0	165	0
C113	14	16	2	1	0	88	14	28	6	0	0	74
***F*-measures**	*0.83*	*0.76*	*0.84*	*0.77*	*0.64*	*0.46*	*0.83*	*0.62*	*0.78*	*0.78*	*0.62*	*0.26*

Author's complimentary copy

VI. CONCLUSIONS

Classification is a crucial step in any study of behavior, and all classification problems require decisions to be made by a human observer. Avoiding observer bias during the classification process is thus a central issue (Milinski, 1997). On the one hand, although subjective judgments such as classifying spectrogram features by eye can be highly valuable for this task, the error-rate during classification increases with the difficulty of this process. Specifically, the more different call types there are, the more errors are committed. On the other hand, a machine-learning approach cannot identify non-acoustic specificities. Thus, the best approach is arguably one that can combine the advantages of both methods—the complex integrating capability of the human eye and objective specificity of the automatic process. This combination can likely produce the best possible classification performance with the least possible bias. The current work may significantly contribute to creating such a methodology for bioacoustical research, as it provides a means of at least partially automating categorization of vocalizations, with classification being performed in a systematic, repeatable, and fast way.

ACKNOWLEDGMENTS

We thank Alexandre Charrier, Marie Lechardoy, Aurélie Rucart, Erik Gustaffson, Marion Grand, Syrina Al Aïn, Virginie Mazurier, and Amaury La Burthe for their help in conducting the experiments. We also thank Colette Désaleux and Philippe Groué for taking care of the birds. Funding for this research was provided by the 6th PCRD Nest Pathfinder “What it means to be human.” The experiments comply with the French laws concerning animal care.

APPENDIX: OPERATORS APPEARING IN THE FEATURES USED

Norm: reduces a matrix to the vector of the Euclidian norm of the rows.

HNorm: reduces a matrix to the vector of the Euclidian norm of the columns.

PitchPraat: uses Praat to compute the pitch of a signal with the default parameter values (Boersma and Weenink, 1996).

VMeanNormalization: for a given matrix, computes a matrix where each element is the corresponding element of the input matrix minus the average value of the corresponding column.

HMeanNormalization: for a given matrix, computes a matrix where each element is the corresponding element of the input matrix minus the average value of the corresponding line.

Skewness: computes the third statistical moment.

MelFilterBank: for a given matrix, returns a matrix which columns contain copies of the input matrix filtered through different mel-frequency bands. The argument is the number of mel bands.

LSTER: computes a line matrix which elements are the ratios of low short-time energy frames of the columns of the input matrix. The arguments are the energy threshold (percentage), the size of the energy frames (number of elements) and their overlap (percentage).

CRRM: computes a matrix which contains the Cepstrum Resynthesis Residual Magnitude (CRRM) of the input matrix. The first argument is the number of mel band filters and the second one is the order of the smoothing filter.

PointProcessPraat: computes Praat’s PointProcess (to interpret an acoustic periodicity contour as the frequency of an underlying point process, such as the sequence of glottal closures in vocal-fold vibration. Then, it uses Praat’s “To Sound (pulse train)” to generate a pulse at every point in the point process. This pulse is filtered at the Nyquist frequency of the resulting Sound. We use Praat’s default parameter values.

Al Aïn, S., Giret, N., Grand, M., Kreutzer, M., and Bovet, D. (**2009**). “The discrimination of discrete and continuous quantities in African grey parrots (*Psittacus erithacus*),” Anim. Cogn. **12**, 145–154.

Anderson, S. E., Dave, A. S., and Margoliash, D. (**1996**). “Template-based automatic recognition of birdsong syllables from continuous recordings,” J. Acoust. Soc. Am. **100**, 1209–1219.

Baker, M. C. (**2003**). “Local similarity and geographic differences in a contact call of the Galah (*Cacatua roseicapilla assimilis*) in Western Australia,” EMU **103**, 233–237.

Bloomfield, L. L., Charrier, I., and Sturdy, C. B. (**2004**). “Note types and coding in parid vocalizations. II: The chick-a-dee call of the mountain chickadee (*Poecile gambeli*),” Can. J. Zool. **82**, 780–793.

Boerma, P., and Weenink, D. (**1996**). “PRAAT: a system for doing phonetics by computer,” Glot Int. **5**, 341–345.

Budney, G. F., and Grotke, R. W. (**1997**). “Techniques for audio recordings vocalizations of tropical birds,” Ornithol. Monogr. **48**, 147–163.

Charrier, I., Bloomfield, L. L., and Sturdy, C. B. (**2004**). “Note types and coding in parid vocalizations. I: The chick-a-dee call of the black-capped chickadee (*Poecile atricapillus*),” Can. J. Zool. **82**, 769–779.

Chang-Hsing, L., Chih-Hsun, C., Chin-Chuan, H., and Ren-Zhuang, H. (**2006**). “Automatic recognition of animal vocalizations using averaged MFCC and linear discriminant analysis,” Pattern Recogn. Lett. **27**, 93–101.

Clark, C. W., Marler, P., and Beaman, K. (**1987**). “Quantitative analysis of animal vocal phonology: An application to swamp sparrow song,” Ethology **76**, 101–115.

Clemins, P. J., and Johnson, M. T. (**2006**). “Generalized perceptual linear prediction features for animal vocalization analysis,” J. Acoust. Soc. Am. **120**, 527–534.

Clemins, P. J., Johnson, M. T., Leong, K. M., and Savage, A. (**2005**). “Automatic classification and speaker identification of African elephant (*Loxodonta africana*) vocalizations,” J. Acoust. Soc. Am. **117**, 956–963.

Cortopassi, K. A., and Bradbury, J. W. (**2000**). “The comparison of harmonically rich sounds using spectrographic cross-correlation and principal coordinates analysis,” Bioacoustics **11**, 89–127.

Cruickshank, A. J., Gautier, J. P., and Chappuis, C. (**1993**). “Vocal mimicry in wild African grey parrots (*Psittacus erithacus*),” Ibis **135**, 293–299.

Davis, S. B., and Mermelstein, P. (**1980**). “Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences,” IEEE Trans. Acoust., Speech, Signal Process. **28**, 357–366.

Draganoiu, T. I., Nagle, L., Musseau, R., and Kreutzer, M. (**2006**). “In a songbird, the black redstart, parents use acoustic cues to discriminate between their different fledglings,” Anim. Behav. **71**, 1039–1046.

Eronen, A., and Klapuri, A. (**2000**) “Musical instrument recognition using cepstral coefficients and temporal features,” in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, **2**, pp. II753–II756.

Farabaugh, S. M., and Dooling, R. J. (**1996**). “Acoustic communication in parrots: Laboratory and field studies of budgerigars (*Melopsittacus undulatus*),” in *Ecology and Evolution of Acoustic Communication in Birds*, edited by D. E. Kroodsma and E. H. Miller (Cornell University Press, Ithaca, NY), pp. 97–117.

Farabaugh, S. M., Linzenbold, A., and Dooling, R. J. (**1994**). “Vocal plasticity in budgerigars (*Melopsittacus undulatus*): Evidence for social factors in the learning of contact calls,” J. Comp. Psychol. **108**, 81–92.

Fernández-Juricic, E., Martella, M. B., and Alvarez, E. V. (**1998**). “Vocalizations of the blue-fronted Amazon (*Amazona aestiva*) in the Chancaní Reserve,” Wilson Bull. **110**, 352–361.

Author's complimentary copy

Foote, J. T. (**1997**) "Content-based retrieval of music and audio," Proc. SPIE, **3229**, 138–147.

Giret, N., Miklósi, Á., Kreutzer, M., and Bovet, D. (**2009**). "Use of experimenter given cues by African gray parrots (*Psittacus erithacus*)," Anim. Cogn. **12**, 1–10.

Giret, N., Péron, F., Lindová, J., Tichotová, L., Nagle, L., Kreutzer, M., Tymr, F., and Bovet, D. (**2010**). "Referential learning of French and Czech labels in African grey parrots (*Psittacus erithacus*): Different methods yield contrasting results," Behav. Processes **85**, 90–98.

Guyon, I., and Elisseeff, A. (**2003**). "An introduction to variable and feature selection," J. Mach. Learn. Res. **3**, 1157–1182.

Hall, M. A. (**2000**). "Correlation-based feature selection for discrete and numeric class machine learning" (Working paper 00/08), (Department of Computer Science, University of Waikato, Hamilton, New Zealand).

Härmä, A. (**2003**). "Automatic identification of bird species based on sinusoidal modeling of syllables," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Hong Kong, pp. 545–548.

Härmä, A., and Somervuo, P. (**2004**). "Classification of the harmonic structure in bird vocalization," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, Montreal, Canada, pp. 701–704.

Janik, V. M. (**1999**). "Pitfalls in the categorization of behavior: A comparison of dolphin whistle classification methods," Anim. Behav. **57**, 133–143.

Jones, A. E., Ten Cate, C., and Bijleveld, C. C. J. H. (**2001**). "The interobserver reliability of scoring sonagrams by eye: A study on methods, illustrated on zebra finch songs," Anim. Behav. **62**, 791–801.

Khanna, H., Gaunt, S. L. L., and Mccallum, D. A. (**1997**). "Digital spectrographic cross-correlation, tests of sensitivity," Bioacoustics **7**, 209–234.

Kogan, J. A., and Margoliash, D. (**1998**). "Automated recognition of bird song elements from continuous recordings using dynamic time warping and hidden Markov models: A comparative study," J. Acoust. Soc. Am. **103**, 2185–2196.

Koza, J. R. (**1992**). *Genetic Programming: On the Programming of Computers by Means of Natural Selection* (The MIT Press, Cambridge, MA), p. 840.

Marler, P. (**2004**). "Science and birdsong: The good and old days," in *Nature's Music*, edited by P. Marler and H. Slabbekoorn (Elsevier Academic Press, San Diego, CA), pp. 1–38.

Mccowan, B. (**1995**). "A new quantitative technique for categorizing whistles using simulated signals and whistles from captive bottlenose dolphins (*Delphinidae, Tursiops truncatus*)," Ethology **100**, 177–193.

Milinski, M. (**1997**). "How to avoid seven deadly sins in the study of behavior," Adv. Study Behav. **26**, 160–180.

Mjolsness, E., and Decoste, D. (**2001**). "Machine learning for science: State of the art and future prospects," Science **293**, 2051–2055.

Molnár, C., Kaplan, F., Roy, P., Pachet, F., Pongrácz, P., Dóka, A., and Miklósi, Á. (**2008**). "Classification of dog barks: A machine learning approach," Anim. Cogn. **11**, 389–400.

Nowicki, S., and Nelson, D. A. (**1990**). "Defining natural categories in acoustic signals: Comparison of three methods applied to "chick-a-dee" call notes," Ethology **86**, 89–101.

Pachet, F., and Roy, P. (**2009**). "Analytical features: A knowledge-based approach to audio feature generation," EURASIP J. Audio Speech Music Process, pp. 1–39.

Pepperberg, I. M. (**1999**). *The Alex Studies: Cognitive and Communicative Abilities of Grey Parrots* (Harvard University Press, Cambridge, MA), p. 448.

Pepperberg, I. M. (**2006**). "Cognitive and communicative abilities of grey parrots," Appl. Anim. Behav. Sci. **100**, 77–86.

Pepperberg, I. M. (**2007**). "Grey parrots do not always 'parrot': The roles of imitation and phonological awareness in the creation of new labels from existing vocalizations," Lang. Sci. **29**, 1–13.

Quinlan, J. R. (**1993**). *C4.5: Programs for Machine Learning* (Morgan Kaufmann Publishers, San Francisco, CA), p. 302.

Schrader, L., and Hammerschmidt, K. (**1997**) "Computer-aided analysis of acoustic parameters in animal vocalizations: A multi-parametric approach," Bioacoustics **7**, 247–265.

Stockwell, D. R. B. (**2006**). "Improving ecological niche models by data mining large environmental datasets for surrogate models," Ecol. Modell. **192**, 188–196.

Tchernichovski, O., Nottebohm, F., Ho, C. E., Pesaran, B., and Mitra, P. P. (**2000**). "A procedure for an automated measurement of song similarity," Anim. Behav. **59**, 1167–1176.

Thorpe, W. H. (**1954**). "The process of song-learning in the chaffinch as studied by means of the sound spectrograph," Nature **173**, 465–469.

Tyagi, H., Hegde, R. M., Murthy, H. A., and Prabhakar, A. (**2006**). "Automatic identification of bird calls using spectral ensemble average voice prints," in *Proceedings of the European Signal Processing Conference* (EUSIPCO), Italy, p. 5.

Vallet, E., and Kreutzer, M. (**1995**). "Female canaries are sexually responsive to special song phrases," Anim. Behav. **49**, 1603–1610.

Author's complimentary copy