Automatic classification and speaker identification of African elephant (*Loxodonta africana*) vocalizations

Patrick J. Clemins[a] and Michael T. Johnson
*Speech and Signal Processing Laboratory, Marquette University,
P.O. Box 1881, Milwaukee, Wisconsin 53233-1881*

Kirsten M. Leong[b] and Anne Savage
Disney's Animal Kingdom, Lake Buena Vista, Florida 32830

(Received 27 February 2004; revised 26 October 2004; accepted 13 November 2004)

A hidden Markov model (HMM) system is presented for automatically classifying African elephant vocalizations. The development of the system is motivated by successful models from human speech analysis and recognition. Classification features include frequency-shifted Mel-frequency cepstral coefficients (MFCCs) and log energy, spectrally motivated features which are commonly used in human speech processing. Experiments, including vocalization type classification and speaker identification, are performed on vocalizations collected from captive elephants in a naturalistic environment. The system classified vocalizations with accuracies of 94.3% and 82.5% for type classification and speaker identification classification experiments, respectively. Classification accuracy, statistical significance tests on the model parameters, and qualitative analysis support the effectiveness and robustness of this approach for vocalization analysis in nonhuman species. © *2005 Acoustical Society of America.* [DOI: 10.1121/1.1847850]

PACS numbers: 43.80.Lb, 43.80.Ka [JAS] Pages: 956–963

I. INTRODUCTION

One major task in bioacoustic research is determining repertoires for various species and then correlating the different vocalizations with behavior (Berg, 1983; Cleveland and Snowdon, 1982; Poole *et al.*, 1988; Sjare and Smith, 1986a, b). Currently, many features used to determine the vocalization type are extracted by hand from spectrogram plots, introducing bias into the feature values. Improved feature extraction and automatic classification would drastically decrease the time spent analyzing, classifying, and segmenting vocalizations, as well as provide a method for unbiased feature extraction. In addition, automatic classification systems can sometimes identify acoustic patterns correlating to the psychological or physiological state of an animal that are not obvious from simple spectrogram measures.

In the field of bioacoustics, traditionally, animal vocalization analysis is done using statistical methods such as multivariate analysis of variance (MANOVA), discriminant function analysis, or principal components analysis (PCA) (Fristrup and Watkins, 1992; Leong *et al.*, 2002; Owren *et al.*, 1997; Riede and Zuberbühler, 2003; Sjare and Smith, 1986a). By incorporating these traditional methods with a classification system such as that presented here, it is possible to go beyond traditional hypothesis testing and build systems that will classify unknown vocalizations, find new types of vocalizations, and measure how the vocalizations vary within and across classes. One key benefit of this approach is that automatic classification methods can model and compensate for temporal variation of vocalization patterns, making better use of timing information than traditional whole-spectrogram measures.

Previous studies in automatic classification of animal vocalizations include substantial work in feature identification as well as a number of papers on complete classification systems. Fristrup and Watkins (1992) have created an analysis package, Acoustat, capable of automatically extracting 26 different features including median center frequency, bandwidth, and duration. Spectrograms are used to extract the majority of the features. Chesmore (2001) has implemented a complete automatic classification system that uses time-domain-based features and an artificial neural network (ANN) to classify the vocalizations of various species of insects. Also using an ANN-based classifier, Campbell *et al.* (2002) were able to identify individual sea lions by their calls with 71% accuracy using spectral value inputs. Other studies have compared various classification systems' ability to detect biological oceanic signals. These classification systems include ANNs, hidden Markov models (HMMs), and linear spectrogram correlator filters (Potter *et al.*, 1994; Mellinger and Clark, 1993). Finally, Anderson (1999) compared a HMM-based classification system against a dynamic time warping (DTW)-based system using a dataset consisting of two different species of bird song. His conclusion was that while the DTW system worked better with a small amount of training data, the HMM system was more robust to noise and more variable vocalizations.

Since the tasks of speaker identification and vocalization classification, common in bioacoustic analysis, correspond directly to common human speech processing tasks, existing speech processing algorithms can be adapted for application to animal vocalizations. Speech processing methods are attractive because of the large research effort that has been devoted to this field over the past 50 years, and because

[a] Electronic mail: patrick.clemins@marquette.edu
[b] Current affiliation: Department of Natural Resources, Cornell University, Ithaca, New York 14853.

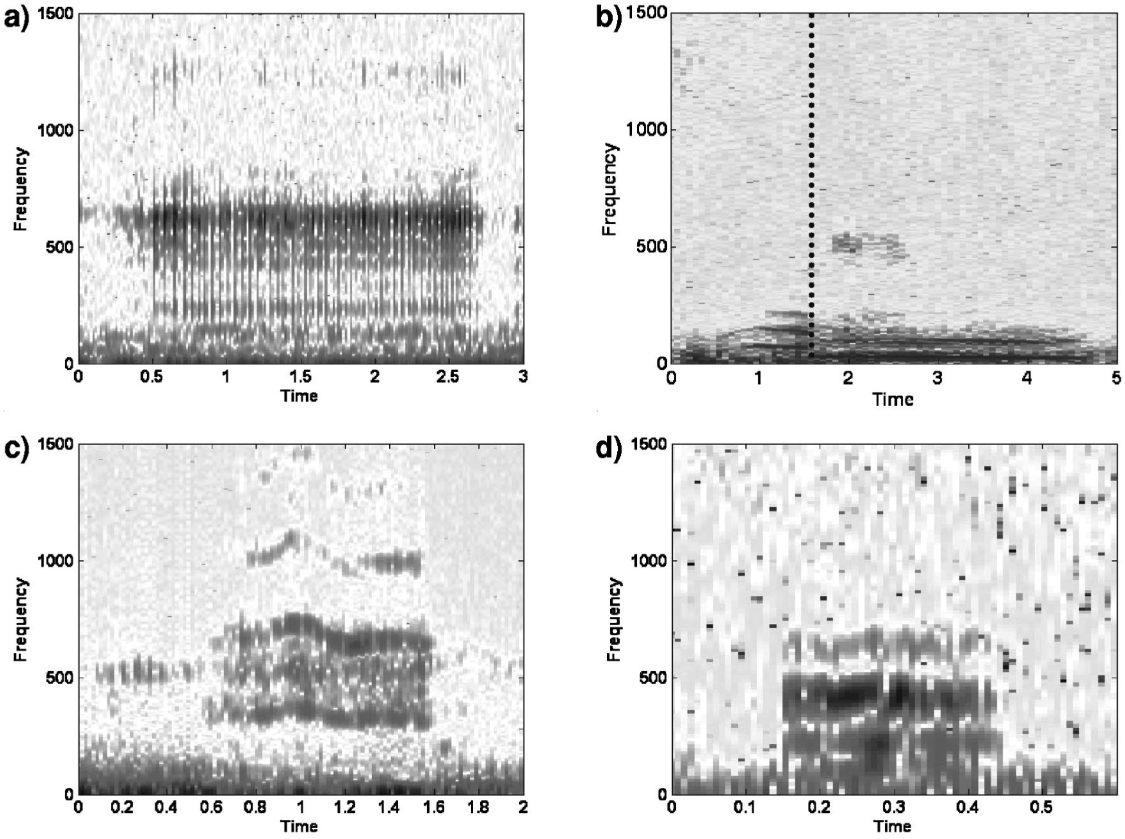

FIG. 1. Spectrograms of various types of elephant vocalizations. (a) Croak. (b) Rev from 1 to 1.5 s, then rumble. (c) Trumpet. (d) Snort.

current speech systems include robust feature extraction techniques coupled with optimal statistical classification models. Justification for the application of speech processing techniques to bioacoustics is supported by studies that suggest that most mammalian vocal production and reception systems are extremely similar (Bradbury and Vehrencamp, 1998; Titze, 1994). Therefore, it is reasonable to envision that the structure of human speech algorithms could be adapted to most mammalian species. Current state-of-the-art human speech systems can achieve classification accuracies of 92% for speech recognition systems on dictated speech (Padmanabhan and Picheny, 2002) and 85% for speaker identification on conversational telephone speech (Reynolds, 2002), although these accuracies can vary widely due to background noise characteristics or the number of speakers enrolled in the system.

While these systems show promise for use in the field of animal bioacoustics, there are challenges with animal vocalizations that include noise and label validity. Noise due to poor recording environments can greatly decrease classification accuracy, especially if the characteristics of the noise vary across the dataset or within individual recordings. Label validity relates to the accuracy and consistency of expert vocalization annotations. When identifying individual animals, it can sometimes be difficult to tell which member of a group is vocalizing even with visual inspection, and when annotating behavior or intended meaning, the animals' behavioral cues are often ambiguous.

African elephants (*Loxodonta africana*) have been extensively studied by researchers for several decades. There is a wealth of information on social dynamics, reproductive strategies, and modalities of communication that provides us with a detailed understanding of the behavioral ecology of this species in the wild. The vocalizations of the African elephant have been categorized using various schemes (Berg, 1983; Leong *et al.*, 2002; Poole *et al.*, 1988). Based on the assessment of spectrograms coupled with behavioral analysis, these studies have found that there are about ten different basic vocalization classes, including the rumble, rev, croak, snort, and trumpet. Many of these classes likely include subtypes. Example spectrograms of a few of these vocalization classes are shown in Fig. 1. The rumble, with much of its energy concentrated in the infrasound range as low as 12 Hz, is the most common vocalization (Leong *et al.*, 2002). The rumble is used to communicate between groups and within each family group. Playback studies have shown that the lower frequency characteristics of the rumble allow it to be used to communicate over long distances (Poole *et al.*, 1988; Langbauer *et al.*, 1991), and this function is often emphasized in discussions of this type of vocalization. Less common than the rumble are the rev, usually emitted when the elephant is startled and often followed by a rumble, and the croak, usually occurring in groups of two or three and commonly associated with the elephant sucking either water or air into the trunk (Leong *et al.*, 2002). Other vocalization classes include the snort, a higher frequency vocalization most generally used as a low-excitement greeting or request for contact, and the trumpet, produced when the elephant is excited (Berg, 1983; Leong *et al.*, 2002; Poole *et al.*, 1988). In addition, there are a few vocalization types that have been observed, but are not used in the present study, including the cry, growl, roar, and bark (Berg, 1983).

This study documents a system, modeled after human speech recognition algorithms, for automatic feature extraction and classification of elephant vocalizations by type and speaker. This system is effective and robust in classifying elephant vocalizations and has implications as a technique to improve and broaden analysis of bioacoustic data.

II. DATA

A. Subjects

The subjects for this study are one male (18 years of age) and six adult nulliparous female (age range 19–30 years) African elephants housed at Disney's Animal Kingdom™, Lake Buena Vista, FL. These elephants are part of a long-term study of elephant communication that incorporates behavioral, hormonal, and vocal data to provide a detailed investigation on the behavioral and reproductive strategies of African elephants. Detailed information on the results of these studies can be found in Leong *et al.* (2002, 2003).

B. Data collection

For a detailed description of the methods used to record elephant vocalizations, see Leong *et al.* (2002, 2003). In brief, each elephant was fitted with a custom-designed collar that contained a microphone and a RF transmitter. Collars were designed, built, and packaged by Walt Disney World Co. Instrumentation Support Division of Ride and Show Engineering. Each collar transmitted to a separate channel of a TASCAM DA-38 8-channel DAT recorder (TEAC America Inc., Montebello, CA) and recorded on separated tracks of a SONY DARS-60MP Digital Audio Tape. The vocalizations were manually extracted from the DAT, passed through an antialiasing filter, and stored on a computer at a sampling rate of 7518 Hz.

The vocalizations are extracted off the DAT tapes using Real-Time Spectrogram (RTS) software (version 2.0) by Engineering Design, Belmont, MA. All vocalizations visually or acoustically identified were saved as individual files. For these experiments, a number of the clearer vocalizations were selected at random using signal-to-noise ratio and the lack of interference for the duration of the vocalization as the main criteria.

III. METHODS

A. Feature extraction

Features were extracted from the vocalizations using a moving Hamming window in a similar manner as in Clemins and Johnson (2003). Window sizes of 30 ms are typical for human speech, based on tradeoffs between frequency resolution and signal stationarity. Since African elephant vocalizations have a fundamental frequency range of 7 to 200 Hz (Langbauer, 2000), much lower than human speech, the window size was increased to 60 ms for the call classification experiment. A window size of 300 ms was used for the speaker identification experiment because only rumbles, the lowest frequency vocalizations, were used in this experiment. In all experiments, the frame rate was one-third the window size, so that consecutive windows overlap by two-thirds. This overlapping allowed improved temporal resolution for time alignment while still maintaining sufficient frequency resolution for spectral analysis.

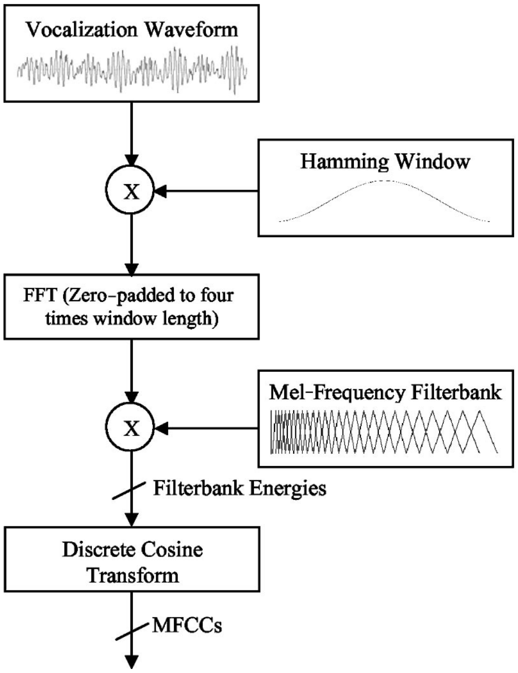

FIG. 2. Feature extraction process. A window of the vocalization waveform is manipulated to generate a number (12 in this set of experiments) of Mel-frequency cepstral coefficients.

Twelve Mel-frequency cepstral coefficients (MFCCs) plus log-energy were extracted from each moving window. Cepstral coefficients (Davis and Mermelstein, 1980) are extremely common spectral features in human speech processing and have a number of beneficial characteristics. These include the ability to capture vocal tract resonances but exclude excitation patterns, the usage of Euclidian distance between sets of coefficients which directly relates to log-spectral distances, and the tendency for coefficients to be largely uncorrelated which makes them well suited for statistical pattern recognition models. The signal processing basis for the cepstral representation is based on the source-filter model of human speech, which also has been used recently to describe the vocal production systems of many different animal species (Fitch, 2003). As shown in the block diagram of Fig. 2, MFCCs were derived by first computing the fast Fourier transform (FFT) of each window. Following this, the frequency axis was warped to the Mel-scale by multiplying the spectrum with a series of Mel-spaced triangular filters. The use of a Mel-spaced frequency scale is supported by evidence that elephants, like humans, perceive frequencies on a logarithmic scale (Heffner and Heffner, 1982; Békésy, 1960). The energy from the frequency band filters was then used as input to a discrete cosine transform, whose output provides cepstral coefficients.

The Mel-frequency filter bank was adjusted to the range 10 to 2000 Hz for the call classification experiment and 10 to 150 Hz for the speaker identification experiment in order to focus on the part of the spectrum used by elephants in the

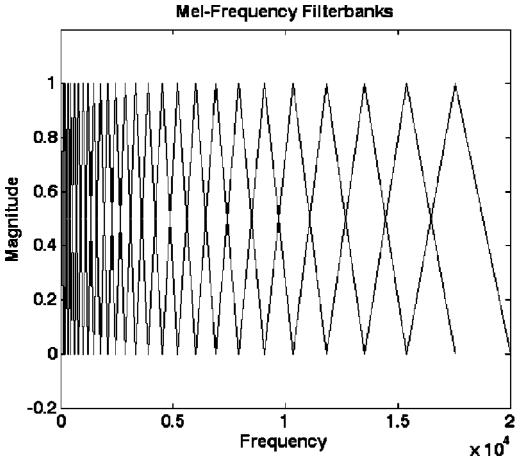

FIG. 3. Mel-frequency filterbanks. This plot shows 26 filter banks spaced between 10 and 2000 Hz using Mel-frequency spacing.

types of calls tested (Langbauer, 2000). A plot of the Mel-frequency filter bank is shown in Fig. 3. Since the signal was recorded at 7518 Hz and the desired filter bank range was only 10 to 150 Hz, the signal was zero padded to four times its original length before calculating the FFT in order to smooth the frequency spectrum.

B. Model parameters

A hidden Markov model (HMM) was used to model each of the different speakers or vocalization types. A HMM is a statistical state machine model, where states represent stationary spectral configurations and transitions between states represent spectral transition (Rabiner and Jaung, 1986). A diagram of a HMM is in Fig. 4. When modeling time series, the states of the HMM are linearly connected with state transitions from left to right, essentially representing time. A HMM is described by its transitions, the probabilities of transitioning from one state to the next, and its state distributions, the probabilities of a particular feature observation occurring while in that state. Each state's observation probability was represented by a multivariate Gaussian distribution. The task of a HMM is essentially to map a sequence of observations, here the sequence of MFCC features throughout a vocalization, onto a sequence of states, and determine the likelihood that the observations could have been generated by that model. To implement a classification task, multiple HMMs are trained, one for each class, and observation examples are classified according to the model yielding the highest likelihood.

HMMs are used in nearly all state-of-the-art speech recognition systems. They were a good choice for this task since they can model both the temporal and spectral differences between similar vocalizations. HMMs have the ability to perform nonlinear temporal alignment during the recognition and training process for vocalizations that may be similar but have different durations. Another reason for using HMMs is that because of their statistical basis, other statistical information, such as grammar models, can be easily incorporated. The programming toolkit used here for model implementation is HTK 3.1.1 from Cambridge University (2002). HTK provides a robust set of tools to implement HMM models, including the Baum-Welch Expectation Maximization algorithm (Baum *et al.*, 1970; Moon, 1996) for training and the Viterbi algorithm (Forney, 1973) for classifying new vocalizations. For these experiments, we used a three-state left-to-right HMM. A silence model was also trained and included before and after each vocalization model to account for varying amounts of leading and trailing silence regions.

IV. RESULTS

A. Vocalization type classification

The vocalization type classification experiment is analogous to an isolated-word speech recognition experiment. Five different African elephant vocalization types were classified in this experiment, using a total of 74 calls from six different animals. The goal of this experiment was to compare how well the HMM system outlined above performs on a classification task that can be easily done by human experts. Using the methodology outlined in the previous section, one HMM was trained for each vocalization type. To maximize training set size given the limited number of examples, leave-one-out cross validation was used for testing, so that each example was tested using models trained on all examples other than itself. The distribution of the data across speakers and vocalization types is shown in Fig. 5.

The confusion matrix for this experiment is shown in Fig. 6. The overall classification accuracy is 79.7%. As can be seen, rumbles were classified most accurately at 90.9% while croaks are classified with the least accuracy at 70.6%.

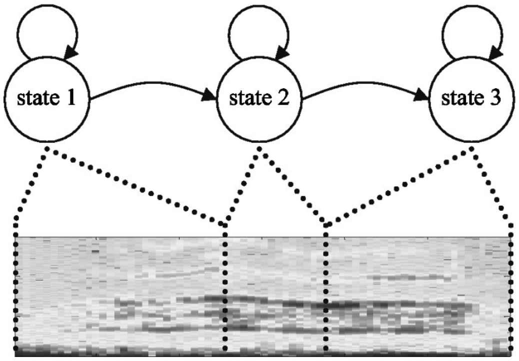

FIG. 4. A hidden Markov model (HMM). Each state of the HMM corresponds to the spectral characteristics of the animal vocalization as they vary in time. These characteristics are modeled by a multivariate Gaussian in each state.

		Vocalization Type				
		Croak	Rumble	Rev	Snort	Trumpet
Speaker	Bala	1 - 1	10 - 6	11 - 7	3 - 2	1 - 1
	Fiki	3 - 2	1 - 1	0 - 0	4 - 0	3 - 3
	Mackie	13 - 2	0 - 0	2 - 0	6 - 0	0 - 0
	Moyo	0 - 0	0 - 0	0 - 0	0 - 0	6 - 5
	Robin	0 - 0	0 - 0	0 - 0	2 - 1	4 - 4
	Thandi	0 - 0	0 - 0	1 - 0	2 - 0	1 - 0

FIG. 5. Distribution of the vocalizations by type and speaker for the vocalization type experiments. The first number in each cell is the number of calls in the complete dataset. The second number in each cell is the number of calls in the clean dataset.

One possible hypothesis for this is that rumbles are the longest vocalization type and therefore have more data windows on which to base a classification decision. Conversely, the snort is one of the shortest vocalizations and, thus, has less data with which to make a decision. It should also be noted that the rumbles are mainly from one speaker while the snorts are more evenly distributed across the speakers. This could also account for the discrepancy in classification accuracy between these two vocalization types.

This experiment was also run taking out vocalizations with poor quantization or poor signal to noise characteristics. All vocalizations were amplitude scaled to normalize their power; however, for this portion of the experiment, those vocalizations with a scale factor larger than 1100 were not included because of the poor quantization of the signal over the full range of possible values. A metric related to the signal to noise ratio, which we call the signal-to-noise characteristic (SNC), was calculated for each vocalization using the following formula:

$$\mathrm{SNC} = \frac{\mathrm{FrameEnergy}_{\mathrm{max}}}{\mathrm{FrameEnergy}_{\mathrm{ave}}}. \tag{1}$$

Those vocalizations with a SNC of less than 5.0 were not used in this portion of the experiment. The distribution of the clean vocalizations across speakers and vocalization type is shown in Fig. 5. Notice that croaks are more evenly distributed in the clean dataset as compared to the entire dataset. The classification matrix of the vocalization type classification experiment with poor quality vocalizations removed is shown in Fig. 6. Note that the classification accuracy of the system improved from 79.7% to 94.3% when only the 35 highest quality vocalizations are used.

In order to visualize the differences captured in each trained HMM, a 15-state HMM was trained for each class using 26 filter bank energies as features. Using these filter bank energies and the 15 temporal states, a spectrogram can be plotted which represents the ‘‘maximum likelihood spectrogram’’ for that vocalization. The maximum likelihood spectrogram for each of the vocalizations is shown in Fig. 7. The blockiness of the plots is a result of relatively low data resolution (15 states horizontally versus 26 filterbanks vertically). The larger low-frequency content of the rumble is evident from the spectrograms as well as the noisy nature of the croak whose spectrogram shows very little structure. The short duration of the snort and rev are also captured in these spectrograms.

While accuracy results demonstrate the ability of the learned models to generalize with respect to unseen test data, they do not provide a statistical measure of the difference between the classes. To test the statistical significance of these differences, a multivariate analysis of variance (MANOVA) test was performed on the 12 MFCC coefficients and log energy measure extracted from each frame of the vocalizations. The HMM state of each 13-parameter data vector was determined by a forced Viterbi alignment using trained HMM models for each vocalization class. In order to show that each state of the trained HMMs is statistically different, both the state and vocalization class were used as independent variables. The result of the MANOVA test using Wilk’s Λ statistic was that the five HMMs represent statisti-

		Classification				
		Croak	Rumble	Rev	Snort	Trumpet
Label	Croak	12	2	2	1	0
	Rumble	0	10	1	0	0
	Rev	0	1	11	2	0
	Snort	0	1	2	13	1
	Trumpet	2	0	0	0	13

		Classification				
		Croak	Rumble	Rev	Snort	Trumpet
Label	Croak	5	0	0	0	0
	Rumble	0	7	0	0	0
	Rev	0	1	6	0	0
	Snort	0	1	0	2	0
	Trumpet	0	0	0	0	13

FIG. 6. Confusion matrices for vocalization type experiments. Left: Confusion matrix over all vocalizations in dataset. Accuracy: 59/74=79.73%. Right: Confusion matrix over clean vocalizations in dataset. Accuracy: 33/35=94.29%.

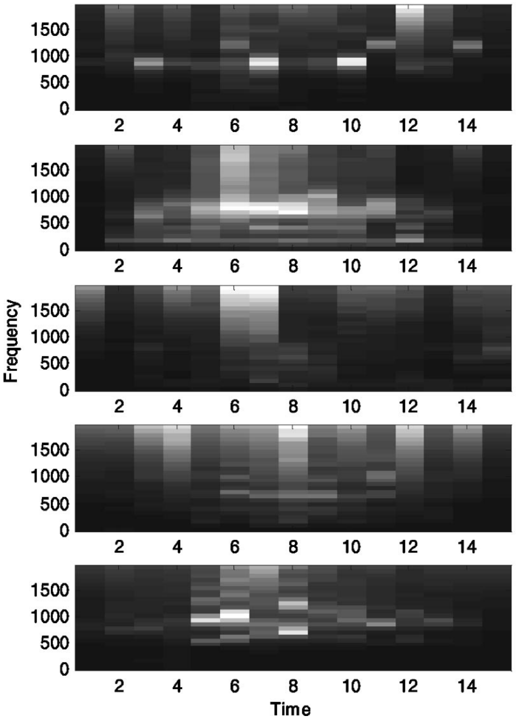

FIG. 7. Maximum likelihood spectrograms for vocalization type experiments. The plots show 26 filterbank energies on the vertical axis across 15 states of a trained HMM on the horizontal axis and graphically represent the HMM for each type of vocalization. From top to bottom: croak, rumble, rev, snort, trumpet.

cally different vocalizations ($F_{104,9483}=150.8, P<0.001$). A second MANOVA analysis was performed disregarding the state information and, therefore, with only one independent variable, vocalization class. This is equivalent to assuming that each class can be represented by a single state HMM. Using Wilk's Λ statistic, each single-state HMM represents statistically different vocalizations ($F_{52,9483}=342.5, P<0.001$).

B. Speaker identification

Speaker identification was performed on data collected in two separate social contexts. The first social context is where the single male was separate from the six females. The second social context is with the male and four of the females grouped together. All vocalizations in the speaker identification data set are rumbles, making it similar to a text-dependent task for human speech. This experiment was proposed to test the hypothesis that, like humans, each elephant has slightly different vocal characteristics that are consistent for certain vocalization types.

The classification matrix for this experiment is shown in Fig. 8. Again, leave-one-out cross validation was used to obtain the confusion matrices. The classification accuracy over the six different elephants was 82.5%. Some individuals were easier to distinguish than others, with accuracies ranging from a low of 75% to a high of 95%, implying that the degree of similarity between the elephants varies somewhat.

This theory is supported by the plot of the maximum likelihood spectrograms for each elephant in Fig. 9. Thandi and Fiki share similar characteristics such as a rather weak fundamental frequency contour and the upper harmonic energy peak coming at the peak of the fundamental frequency contour. The spectrograms for Robin and Bala are also similar. Both spectrograms show a medium strength fundamental frequency contour and the peak in upper harmonic strength comes after the peak of the fundamental frequency contour.

		Classification					
		Bala	Fiki	Mackie	Moyo	Robin	Thandi
Label	Bala	19	0	0	0	1	0
	Fiki	1	23	0	0	0	6
	Mackie	0	0	13	0	0	1
	Moyo	1	0	0	13	2	1
	Robin	5	0	0	0	29	0
	Thandi	0	6	0	1	0	21

FIG. 8. Confusion matrix for speaker identification experiment. Accuracy: 118/143=82.52%.

When the vocalizations are separated by social context, recognition accuracies are comparable: 86.9% for vocalizations made while the male was separate from the females and 79.6% for vocalizations made while the male and four females were together. The similar accuracy numbers across social contexts would support a theory that the elephants do

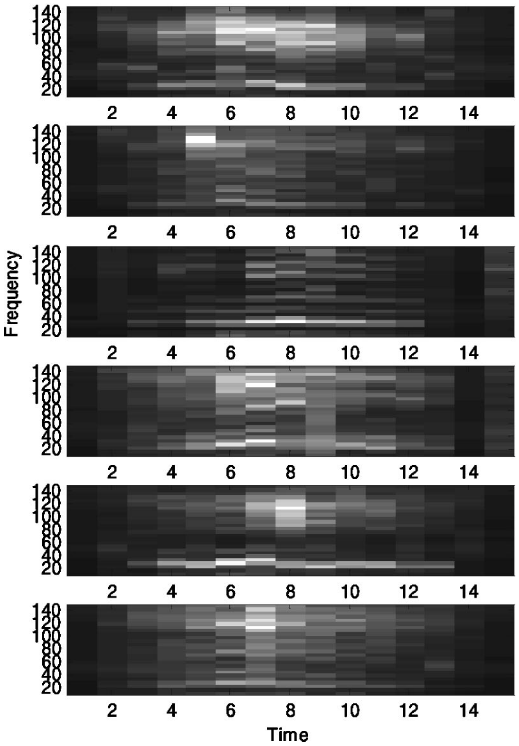

FIG. 9. Maximum likelihood spectrograms for speaker identification experiment. The plots show 26 filterbank energies on the vertical axis across 15 states of a trained HMM on the horizontal axis and graphically represent the HMM for each speaker. From top to bottom: Bala, Fiki, Mackie, Moyo, Robin, Thandi.

not vary their vocalizations significantly while in different social contexts and that the elephants do not use different vocalizations for varying audiences.

The MANOVA test using Wilk's Λ statistic over state and vocalization class showed that each model state represents statistically different vocalizations $(F_{130,13396} = 66.28, P < 0.001)$. A MANOVA test using Wilk's Λ statistic without state information also showed that the HMMs represent statistically different vocalizations $(F_{65,13396} = 115.7, P < 0.001)$.

Our hypothesis is supported by the performance of the classifier as well as by the MANOVA test showing that the models that represent each elephant are different with a high level of statistical significance. In addition, playback experiments demonstrated that elephants can distinguish vocalizations of familiar and nonfamiliar individuals and predicted that the elephants would have to be familiar with the voices of at least 100 adult females to make the observed discriminations (McComb *et al.*, 2000). One reason for the difference between animals could be the vocal tract structure. Although they are functionally identical, individual differences such as length of the vocal tract and shape of the nasal cavity affect the elephant's vocalizations in a consistent way. Another reason for the difference could be a factor that resembles human dialects or accents. Dialects have been found in various species (Dayton, 1990; Santivo, 2000), and, given that the origin of the elephants in this study is varied, the individuals could have developed accents that are present in a certain geographic region or among a specific family group.

V. CONCLUSIONS

This paper has explored the application of speech processing techniques to African elephant vocalizations. Using typical speech processing features and models, African elephant vocalization types were classified with an accuracy of 79.7% (94.3% when poorly quantized and poor SNR examples are removed) and speaker identification resulted in an accuracy of 88.1%. MANOVA tests on both experiments showed that the trained models for each class represent significantly different vocalizations.

There are a number of factors that affect the classification accuracies. The primary factor is the quality of the vocalizations, as is clearly seen in the call-type experiments where removing poor examples reduced error by 73% relative to including all vocalizations. In many bioacoustic studies, the vocalizations are categorized by human experts into groups of varying quality. Then, only the top few categories are used in the analysis. In this study, the lowest quality vocalizations were discarded by experts and a fully automated evaluation mechanism was used to further filter out all but the highest-quality vocalizations.

Another factor that could reduce classification accuracies is the use of suboptimal features to quantify the vocalizations. The features used in these experiments are common to speech processing and are based on human speech production and perception mechanisms. Researchers studying animal communication typically use different features than researchers studying human speech to analyze vocalizations. Features derived from spectrograms such as fundamental frequency and bandwidth are typically combined with time-domain features such as duration to generate a complete feature set. These features are also generally calculated over the entire vocalization instead of on a frame-by-frame basis. The incorporation of more long-term features and more detailed harmonic information is likely to improve the accuracy of the classifier, leading to a continued need to develop automated feature extraction methods for such measures.

Additionally, the validity of the data labels affects classification accuracy. It is well known that elephants use the same general class of vocalization to express different things (Berg, 1983; Poole *et al.*, 1988), as do many other species. For example, rumbles are used to maintain contact with other elephants and to signal that it is time for the herd to move. In addition, numerous other subtypes of rumbles based on behavior have been described (Poole, 2003). Although it is possible that one vocalization is used for all purposes, the variations in spectrogram features indicate that it is likely that the elephants use specific features of the rumbles as well as contextual and visual information sources to discern these different meanings. Thus, labeling rumbles by behavioral context may reveal acoustically distinct subtypes of rumbles, independent of caller identity. The challenge with this approach is determining which behavioral context to assign to which vocalization, as one vocalization often occurs in conjunction with a variety of behavioral events.

These experiments, particularly the call type classification experiment, show that this classification system is robust to noisy conditions. In the call type classification experiment, the system's robustness to noise was shown through a reasonable degradation of classification accuracy when noisy vocalizations were included in the dataset. If the system was not noise robust, the classification accuracy would have dropped off much more significantly when the nosier vocalizations were included in the dataset. The ability to achieve classification accuracies near 80% in both experiments using relatively noisy vocalizations also shows the robustness of the system. It is important to know that the noise-resilient features along with the statistical-based HMM both contribute to this robustness.

The methods presented here are applicable to a wide variety of species. Each animal has different vocal characteristics that make their vocalizations challenging to analyze; however, many of these different characteristics are similar in nature. Each species' sensitivity to different ranges of the frequency spectrum can be modeled by adjusting the filterbanks used to derive the MFCCs. Differences in structural complexity of the vocalizations can be modeled by varying the HMM topology or adding language models to represent these characteristics.

Speech systems provide an adaptable standard framework for many bioacoustic tasks and applications. Applying these systems in bioacoustics research allows us to effectively analyze animal vocalizations and has the potential to reveal more complex vocalization schemes than previously imagined such as complex interactions between harmonics and grammatical structure between syllables with the addition of a statistical language model. Continuing work in this area includes incorporation of additional features related to fundamental frequency and harmonic measures, integrating features at multiple temporal scales, and developing general perceptual-based features that can be easily adapted for different species.

ACKNOWLEDGMENTS

The authors would like to thank the staff of the Wildlife Tracking Center and the Elephant Team at Disney's Animal Kingdom™ for the collection and organization of the acoustic data used in this research.

Anderson, S. E. (**1999**). ''Speech recognition meets bird song: A comparison of statistics-based and template-based techniques,'' J. Acoust. Soc. Am. **106**, 2130.

Baum, L. E., Petrie, T., Soules, G., and Weiss, N. (**1970**). ''A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains,'' Ann. Math. Stat. **41**, 164–171.

Békésy, G. V. (**1960**). *Experiments in Hearing* (McGraw–Hill, New York).

Berg, J. K. (**1983**). ''Vocalizations and associated behaviors of the African elephant (*Loxodonta africana*) in captivity,'' Z. Tierpsychol. **63**, 63–79.

Bradbury, J. W., and Vehrencamp, S. L. (**1998**). *Animal Communication* (Sinauer, Sunderland, MA).

Campbell, G. S., Gisiner, R. C., Helweg, D. A., and Milette, L. L. (**2002**). ''Acoustic identification of female Steller sea lions,'' J. Acoust. Soc. Am. **111**, 2920–2928.

Chesmore, E. D. (**2001**). ''Application of time domain signal coding and artificial neural networks to passive acoustical identification of animals,'' Appl. Acoust. **62**, 1359–1374.

Clemins, P. J., and Johnson, M. T. (**2003**). ''Application of speech recognition to African elephant (*Loxodonta Africana*) vocalizations,'' Proc. of IEEE ICASSP '03 **1**, 484–487.

Cleveland, J., and Snowdon, C. T. (**1982**). ''The complex vocal repertoire of the adult cotton-top tamarin (*Saguinus oedipus oedipus*),'' Z. Tierpsychol **58**, 231–270.

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**(4), 357–366.

Dayton, L. (**1990**). ''Killer whales communicate in distinct 'dialects,''' New Sci. **125**, 35.

Fitch, W. T. (**2003**). ''Mammalian Vocal Production: Themes and Variation,'' in Proceedings of the 1st International Conference on Acoustic Communication by Animals, 27–30 July, pp. 81–82.

Forney, G. D. (**1973**). ''The Viterbi Algorithm,'' Proc. IEEE **61**, 268–278.

Fristrup, K. M., and Watkins, W. A. (**1992**). ''Characterizing Acoustic Features of Marine Animal Sounds,'' Woods Hole Oceanog. Inst. Tech. Rept., WHOI-92-04.

Heffner, R. S., and Heffner, H. E. (**1982**). ''Hearing in the Elephant (*Elephas maximus*): Absolute Sensitivity, Frequency Discrimination, and Sounds Localization,'' J. Comp. Physiol. Psychol. **96**(6), 926–944.

Hidden Markov Model Toolkit (HTK) Version 3.1.1 User's Guide. (**2002**). Cambridge University Engineering Department.

Langbauer, Jr., W. R. (**2000**). ''Elephant Communication,'' Zoo Biol. **19**, 425–445.

Langbauer, Jr., W. R., Payne, K. B., Charif, R. A., Rapaport, L., and Osborn, F. (**1991**). ''African elephants respond to distant playbacks of low-frequency conspecific calls,'' J. Exp. Biol. **157**, 35–46.

Leong, K. M., Ortolani, A., Burks, K. D., Mellen, J. D., and Savage, A. (**2002**). ''Quantifying acoustic and temporal characteristics of vocalizations for a group of captive African elephants *Loxodonta africana*,'' Bioacoustics **13**(3), 213–231.

Leong, K. M., Ortolani, A., Graham, L. H., and Savage, A. (**2003**). ''The use of low-frequency vocalizations in African elephant (*Loxodonta africana*) reproductive strategies,'' Horm. Behav. **43**, 433–443.

McComb, K., Moss, C., Sayialel, S., and Baker, L. (**2000**). ''Unusually extensive networks of vocal recognition in African elephants,'' Anim. Behav. **59**, 1103–1109.

Mellinger, D. K., and Clark, C. W. (**1993**). ''Bioacoustic transient detection by image convolution,'' J. Acoust. Soc. Am. **93**, 2358.

Moon, T. K. (**1996**). ''The Expectation-Maximization Algorithm,'' IEEE Signal Process. Mag. **13**(6), 47–60.

Owren, M. J., Seyfarth, R. M., and Cheney, D. L. (**1997**). ''The acoustic features of vowel-like *grunt* calls in chacma baboons (*Papio cyncephalus ursinus*): Implications for production processes and functions,'' J. Acoust. Soc. Am. **101**, 2951–1963.

Padmanabhan, M., and Picheny, M. (**2002**). ''Large-vocabulary speech recognition algorithms,'' IEEE Comput. **35**(3), 42–50.

Poole, J. H. (**2003**). Personal communication with K. M. Leong.

Poole, J. H., Payne, K., Langbauer, Jr., W. R., and Moss, C. J. (**1988**). ''The social context of some very low frequency calls of African elephants,'' Behav. Ecol. Sociobiol. **22**, 385–392.

Potter, J. R., Mellinger, D. K., and Clark, C. W. (**1994**). ''Marine mammal call discrimination using artificial neural networks,'' J. Acoust. Soc. Am. **96**, 1255–1262.

Rabiner, L. R., and Jaung, B. H. (**1986**). ''An introduction to hidden Markov models,'' IEEE ASSP Mag. **3**, 4–15.

Reynolds, D. A. (**2002**). ''An overview of automatic speaker recognition technology,'' IEEE ICASSP **4**, 4072–4075.

Riede, T., and Zuberbühler, K. (**2003**). ''The relationship between acoustic structure and semantic information in Diana monkey alarm vocalization,'' J. Acoust. Soc. Am. **114**, 1132–1142.

Santivo, S., and Galimberti, F. (**2000**). ''Bioacoustics of southern elephant seals. II. Individual and geographical variation in male aggressive vocalisations,'' Bioacoustics **10**, 287–307.

Sjare, B. L., and Smith, T. G. (**1986a**). ''The vocal repertoire of white whales, Delphinapterus *leucas*, summering the Cunningham Inlet, Northwest Territories,'' Can. J. Zool. **64**, 407–415.

Sjare, B. L., and Smith, T. G. (**1986b**). ''The relationship between behavioral activity and underwater vocalizations of the white whale, *Delphinapterus leucas*,'' Can. J. Zool. **64**, 2824–2831.

Titze, I. R. (**1994**). *Principles of Voice Communication* (Prentice–Hall, Englewood Cliffs, NJ).