4-1-2003

Application of Speech Recognition to African Elephant (Loxodonta Africana) Vocalizations

Patrick J. Clemins
Marquette University

Michael T. Johnson
Marquette University, michael.johnson@marquette.edu

Accepted version. Published as a part of the proceedings of the conference, *IEEE International Conference on Acoustics, Speech and Signal Processing, 2003: ICASSP; Hong Kong, China, April 6-10, 2003*, I-484 - I-487. DOI.

Application of Speech Recognition to African Elephant (*Loxodonta africana*) Vocalizations

P.J. Clemins
Speech & Signal Process. Laboratory, Marquette University, Milwaukee, WI
M.T. Johnson
Speech & Signal Process. Laboratory, Marquette University, Milwaukee, WI

Abstract: This paper presents a novel application of speech processing research, classification of African elephant vocalizations. Speaker identification and call classification experiments are performed on data collected from captive African elephants in a naturalistic environment. The features used for classification are 12 mel-frequency cepstral coefficients plus log energy computed using a shifted filter bank to emphasize the infrasound range of the frequency spectrum used by African elephants. Initial classification accuracies of 83.8% for call classification and 88.1% for speaker identification were obtained. The long-term goal of this research is to develop a universal analysis framework and robust feature set for animal vocalizations that can be applied to many species.

IEEE International Conference on Acoustics, Speech and Signal Processing, 2003: ICASSP; Hong Kong, China, April 6-10, 2003, (April 2003): pg. i-484 – i-487. DOI.

Section 1.

Introduction

The analysis of animal vocalizations is an important research area in bioacoustics. Current topics in this field include the role of vocalizations in the communication process, automatic species detection from acoustic data, the creation of vocabularies for individual species and censusing using vocalization rates. Some of the practical issues researchers face include the difficulty of acquiring high quality acoustic data in adverse environments, imperfect labeling of data and inadequate knowledge about how animals produce and perceive sound.

Recently, there has been interest in performing speaker identification and vocalization classification on animal vocalizations.[1–2,3] Since these tasks correspond directly to the speech processing tasks of speaker recognition and speech recognition, this paper explores the use of speech processing algorithms on animal vocalizations. Speech processing algorithms are attractive because of the large research effort devoted to this field. The long-term goal of the research presented here is to create an analysis framework and robust feature set for animal vocalizations.

Although some animal vocalizations can be classified by human experts[4–5,6]) few systems have been developed to automatically classify vocalizations. Automatic classification could drastically decrease the time spent analyzing and classifying vocalizations. Another advantage of using automatic classification systems includes unbiased feature extraction. Currently, many features of the vocalizations are extracted by hand from spectrogram plots, so the individual performing the feature extraction introduces bias in the feature measurements.

One animal that bioacoustic researchers have studied extensively over the years is the African elephant. The vocalizations of the African elephant have been classified using various schemes.[4–5,6] The studies agree that there are about 10 different basic sound types that the African elephant can produce. The types of vocalizations are separated by animal behavior experts based on spectrogram analysis of the vocalizations. Some of these different vocalization types are shown in Figure 1. Notice that some vocalizations, especially the rumble, have much of their energy concentrated in the infrasound range.

Some research obstacles when dealing with animal vocalizations are noisy data and label validity. The incorporation of noise models is important when dealing with animal

IEEE International Conference on Acoustics, Speech and Signal Processing, 2003: ICASSP; Hong Kong, China, April 6-10, 2003, (April 2003): pg. i-484 – i-487. DOI.

vocalizations since the recording environment is usually poor with many interfering noise sources present. This noise can greatly decrease classification accuracy, especially if the characteristics of the noise vary across the dataset. Label validity is another issue since researchers can only guess as to what the animal is trying to communicate acoustically.

This paper will outline a system used to perform both speaker identification and vocalization classification. The data collection process is outlined in section 2. The feature extraction methods are discussed in section 3. Section 4 presents the results from the various experiments.

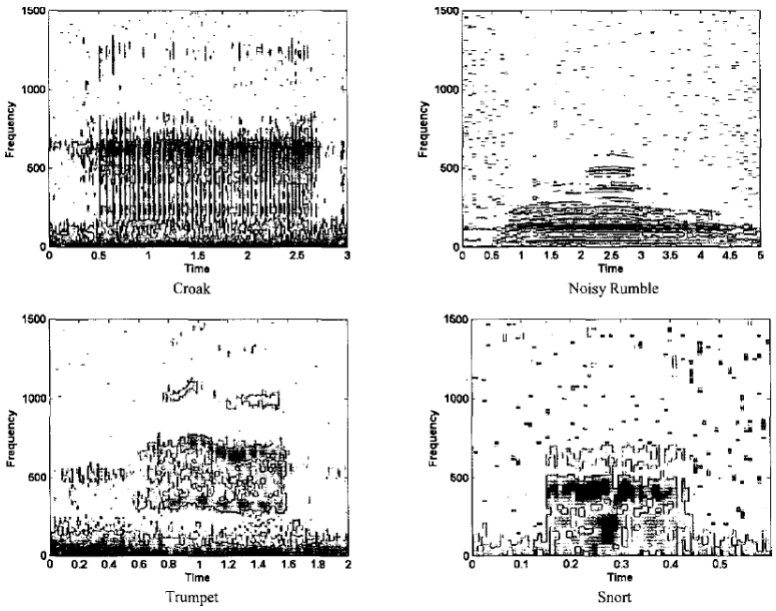

Figure 1 African Elephant Vocalization Types

Section 2.

Data Collection

Animal behavior researchers at Disney's Animal Kingdom™ in Orlando, FL collected the data used in this experiment. Each elephant involved in the data collection project is fitted with a custom designed collar. The collars contain a microphone and an RF radio that broadcast audio to the elephant barn where it is recorded on DAT tapes. The audio is passed through an anti-aliasing filter and stored on computers at a sampling rate of 7518 Hz.

IEEE International Conference on Acoustics, Speech and Signal Processing, 2003: ICASSP; Hong Kong, China, April 6-10, 2003, (April 2003): pg. i-484 – i-487. DOI.

There are 7 elephants involved in the project, one male and 6 females. However, one of the females had very few vocalizations recorded and is not included in these experiments. Based on social dynamics and breeding requirements, the elephants are released into one of three naturalistic yards each day. The two most common configurations in the main yard are all six females together and one male with four females. Along with the audio recordings, time synchronized video is also recorded. In this way, the researchers can label each vocalization with behavior information. More information on the data collection procedure can be found in Leong et. al.[4]

Section 3.

Feature Extraction And Model Parameters

Hidden Markov Models (HMMs) are used to model the different speakers and vocalization types. HMMs are a good choice for this task since they can model the temporal and spectral differences between similar vocalizations. They are also the most popular model used in speech processing.[7]

The programming toolkit used for model implementation is HTK 3.1.1 from Cambridge University.[8] HTK provides a robust set of tools to implement HMM models and is widely used in the speech processing field.

IEEE International Conference on Acoustics, Speech and Signal Processing, 2003: ICASSP; Hong Kong, China, April 6-10, 2003, (April 2003): pg. i-484 – i-487. DOI.

		Classification				
		Croak	Rumble	Rev	Snort	Trumpet
Label	Croak	14	0	2	0	1
	Rumble	0	10	1	0	0
	Rev	0	0	12	2	0
	Snort	0	1	2	13	1
	Trumpet	2	0	0	0	13

Accuracy: 62 / 74 = 83.8%

Figure 2 Type Classification Results

Frame sizes of 30 ms are typical for human speech in order to have several pitch peaks in each frame. However, African elephant vocalizations have a fundamental frequency between 7 Hz and 200 Hz, much lower than human speech.[9] To compensate for this factor, the frame size is increased to 60 ms for the call classification experiment. A frame size of 300 ms is used for the speaker identification experiment because the speaker identification is performed on rumbles which have a fundamental frequency near 10Hz. One-third frame overlap is used in both experiments.

To parameterize the signal, 12 Mel-Frequency Cepstral Coefficients plus log-energy are used. The Mel-Frequency filter bank is adjusted to the range 10Hz to 2000 Hz for the call classification experiment and 10Hz to 150 Hz for the speaker identification experiment in order to filter out noise and focus on the part of the spectrum used by elephants.[9]

The use of a frequency warped scale is supported by evidence that elephants, like humans, perceive frequencies on a logarithmic scale.[10] Since the signal is recorded at 7518 Hz and the desired filter bank range is 10 Hz to 150 Hz, the signal is zero padded in order to smooth the frequency spectrum. An FFT length of four times the frame length, 1200ms, is used for the speaker identification experiment. Smoothing is not required for the call classification experiment since the filter bank bandwidth is larger.

IEEE International Conference on Acoustics, Speech and Signal Processing, 2003: ICASSP; Hong Kong, China, April 6-10, 2003, (April 2003): pg. i-484 – i-487. DOI.

To model the different classes in each experiment, 3-state left-to-right HMMs are used. African elephant vocalizations are largely stationary; therefore, using three states is appropriate given the vocalization length. Because of the small amount of data, single mixture GMMs are used for the observation distributions of each state.

An isolated vocalization setup is used for the experiments. A silence model is included at the beginning and end of each vocalization for the speaker identification experiments. A silence model is not necessary in the vocalization type classification experiment since these vocalizations have been trimmed manually.

		Classification					
		Bala	Fiki	Mackie	Moyo	Robin	Thandi
Label	Bala	9	0	0	1	1	0
	Fiki	0	10	0	0	0	2
	Mackie	0	0	7	0	0	1
	Moyo	0	0	0	12	1	0
	Robin	2	0	0	0	18	0
	Thandi	0	1	0	1	0	18

Accuracy: 74 / 84 = 88.1%

Figure 3 Speaker Identification Results - Dataset I

IEEE International Conference on Acoustics, Speech and Signal Processing, 2003: ICASSP; Hong Kong, China, April 6-10, 2003, (April 2003): pg. i-484 – i-487. DOI.

		Classification					
		Bala	Fiki	Mackie	Moyo	Robin	Thandi
Label	Bala	9	0	0	0	0	0
	Fiki	0	15	0	0	0	3
	Mackie	1	0	5	0	0	0
	Moyo	0	1	0	2	0	1
	Robin	1	0	0	0	13	0
	Thandi	0	4	0	0	1	3

Accuracy: 47 / 59 = 79.7%

Figure 4 Speaker Identification Results - Dataset 2

Section 4.

Results

4.1 Vocalization Type Classification

The vocalization type classification experiment is analogous to a speech recognition experiment on human speech. Five different basic African elephant vocalizations types are classified in this experiment.

The confusion matrix for this experiment is shown in Figure 2. Leave-one-out cross validation has been used to obtain the confusion matrices. The overall classification accuracy from this confusion matrix is 83.8%. As can be seen, rumbles are the easiest to classify while snorts are the most difficult. One possible explanation for this result is that rumbles are the longest vocalization type while the snort one of the shortest.

4.2 Speaker Identification

Speaker identification was performed on two different datasets. The first dataset was obtained while the single male was separate from the six females. The second dataset

IEEE International Conference on Acoustics, Speech and Signal Processing, 2003: ICASSP; Hong Kong, China, April 6-10, 2003, (April 2003): pg. i-484 – i-487. DOI.

was obtained while the male and four of the females were together. Although both datasets could be combined into a single dataset, animal behavior experts suggested that the datasets be treated separately since the females might significantly adjust their vocalizations in the presence of the male. All vocalizations in the speaker identification data set are rumbles. making it essentially a text-dependent task.

Section 5.

Discussion

The classification accuracies for the two datasets are shown in figures 3 and 4. Again, leave-one-out cross validation has been used to obtain the confusion matrices.

The classification accuracy is 88.1% for the first dataset and 79.7% for the second dataset. The lower accuracy for the second dataset may be related to the fact that it is a smaller dataset and therefore has fewer examples to train the models with. Some individuals were easier to distinguish than others, implying that the degree of similarity between the elephants varies somewhat.

This paper explores the application of speech processing to the animal kingdom. Using typical speech processing features and models, African elephant vocalization type classification has been done with an accuracy of 83.8% and speaker identification experiments resulted in an accuracy of 88.1%

Even though these results are promising, there are many factors that result in the deflation of the classification accuracies. The first factor is the quality of the vocalizations. In most bioacoustic studies, the vocalizations are categorized into groups of varying quality. Then, only the top few categories are used in the analysis. In the experiments presented here, all of the vocalizations are used because of the lack of a large number of examples and the desire to create a fully automated system. This results in the use of some poor quality vocalizations which have SNR values below zero decibels.

Another factor is that of feature selection. The features used in these experiments are common to speech processing and are based on human speech production and perception mechanisms. Animal researchers typically use different features than speech researchers to analyze vocalizations. Features derived from spectrograms such as fundamental frequency and bandwidth are typically combined with time-domain features

IEEE International Conference on Acoustics, Speech and Signal Processing, 2003: ICASSP; Hong Kong, China, April 6-10, 2003, (April 2003): pg. i-484 – i-487. DOI.

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 validity of the data labels can also affect classification accuracy. The different types of vocalizations are determined by differences in spectrogram plots, but it is well known that elephants use the same type of vocalization to express different things.[5,6] For example, rumbles are used to maintain contact with other elephants, to call mates and to signal that it is time for the herd to move. Although it is possible that one vocalization is used for all purposes, it is also possible that the elephants are using other features of the rumbles besides spectral magnitude to discern these different meanings. Therefore, the labels may be grouping together vocalizations that are actually dissimilar.

This approach is applicable to other species besides African elephants. We are in the process of acquiring vocalization from other mammalian species. Each species has different vocal characteristics that make their vocalizations challenging to analyze, however, many of the changes are similar in nature. Each species is sensitive to a different part of the frequency spectrum, but this can be easily modeled. Another difference is the complexity of each vocalization. Some animals, such as humans and elephants have relatively simple basic units of speech while other species such as birds and many aquatic mammals have more complex structure in their vocalizations. This difference can be modeled by varying the HMM topology and adding language models to represent these characteristics. Therefore, because of their flexibility, speech systems provide an adaptable standard framework that can be applied to other animals.

References

[1]E. D. Chesmore, "Application of time domain signal coding and artificial neural networks to passive acoustical identification of animals", *Applied Acoustics*, vol. 62, pp. 1359-1374.

[2]K. M. Fristrup, W. A. Watkins, "Marine Animal Sound Classification", *Woods Hole Oceanog Inst. Tech. Rept.*, pp. 94-13.

[3]G. S. Campbell et al., "Acoustic identification of female Steller sea lions", *J. Acoust. Soc: Am*, vol. 111, no. 6, pp. 2910-2928.

[4]K. M. Leong et al., "Quantifying acoustic and temporal characteristics of vocalizations for a group of captive African elephants (*Loxodonta africana*)" in Bioacoustics in press.

[5]J. K. Berg, "Vocalizations and associated behaviors of the African elephant (Loxodonta africana) in captivity", *Z. Tiernsvchol.*, vol. 63, pp. 63-79.

[6]J. H. Poole et al., "The social context of some very low frequency calls of African elephants", *Behav. Ecol. Sociobiol.*, vol. 22, pp. 385-392.

[7]L. R. Rabiner, B. H. Jaung, "An introduction to hidden Markov models", *IEEE ASSP Magazine*, vol. 3, pp. 4-15.

IEEE International Conference on Acoustics, Speech and Signal Processing, 2003: ICASSP; Hong Kong, China, April 6-10, 2003, (April 2003): pg. i-484 – i-487. DOI.

[8]"Hidden Markov Model Toolkit (HTK) Version 3.1.1", 2002.

[9]Jr., W. R Langbauer, "Elephant Communication", *Zoo Biology*, vol. 19, pp. 425-445.

[10]R. S. Heffner, H. E. Heffner, "Hearing in the Elephant (Elephas maximus): Absolute Sensitivity Frequency Discrimination and Sounds Localization", *Journal of Comparative and Physiological Psychology*, vol. 96, no. 6, pp. 926-944.