Presented at Animal Behavior 2002, Vrije Universiteit, Amsterdam, The Netherlands

AUTOMATIC TYPE CLASSIFICATION AND SPEAKER IDENTIFICATION OF AFRICAN ELEPHANT VOCALIZATIONS

Patrick J. Clemins and Michael T. Johnson

Speech and Signal Processing Lab
Marquette University, Milwaukee, WI USA
patrick.clemins@mu.edu and mike.johnson@mu.edu

ABSTRACT

This paper presents systems for automatically classifying elephant vocalizations by type and for identifying the speaker of a given vocalization. The method applies techniques from the speech processing field, with modifications, to elephant vocalizations. The features used for classification are 12 Mel-Frequency Cepstral Coefficients computed using a chirp Z-transform to interpolate among the lower frequencies. A Hidden Markov Model is trained for each type of vocalization and vocalizations are classified using leave-one-out verification. Using this system, initial classification accuracies of 77.0% for type classification and 72.2% for speaker identification resulted. These systems represent the initial stages of a universal analysis framework that could be applied to other animal species.

1. INTRODUCTION

The vocalizations of the African elephant have been studied for a number of years. There have been various classifications of the different types of vocalizations produced (Berg 1983; Poole et al. 1988; Leong et al. 2002). The studies agree that there are about 10 different basic sounds that the African elephant can produce. These different types are separated based on analysis of the vocalization's spectrogram.

Although these different vocalization types are distinguishable by human experts, there has been no development of computer software to do this classification automatically. Automatic classification would drastically decrease the time spent analyzing and classifying vocalizations compared to current techniques. Other advantages include automatic, unbiased feature extraction, and given a good model, the ability to adapt to temporal variance between vocalizations.

Since there are algorithms available to perform similar tasks on human speech, it is the purpose of this paper to show that these well-developed techniques can be applied to African elephant vocalizations. Speech recognition systems, using language models, can achieve over 92% accuracy on dictated speech (Padmanabhan and Picheny 2002). However, this accuracy drops off sharply when the environment has unpredictable noise characteristics.

Another task widely studied by speech processing researchers is speaker identification. The goal of the research is to create a system that can identify the speaker based on the speaker's acoustic or language characteristics. Systems with identification accuracies of over 85% accuracy on conversational telephone speech have been developed successfully (Reynolds 2002).

This paper will outline the system used to accomplish both of these tasks. First, the vocalization is converted from a waveform into a set of meaningful spectral features. This process is discussed in section 2. Next, a model is applied to the vocalization to capture the individual characteristics of each type of vocalization. The model used in this research is a Hidden Markov Model (HMM). This is currently the most popular model for speech processing research. The HMM is discussed in section 3. Section 4 presents a summary of the classification accuracy of the current system.

2. FEATURE EXTRACTION

To extract features from the vocalization waveform, the waveform is first subdivided into frames. Frames are used so that the signal being analyzed within each frame can be considered stationary. When analyzing human speech, frames of around 30ms are used. However, since African

elephants vocalize at much lower fundamental frequencies than human, a frame size of about 60ms is used for the type classification and a frame size of about 300ms is used for the speaker identification which is done only on rumbles. These frames are overlapped by half to provide better time resolution.

Once the signal is divided into frames, the first 12 Mel-Frequency Cepstral Coefficients (MFCCs) are calculated for each frame. To accomplish this, the FFT of the frame is calculated. Then, the frequency axis is warped to the Mel-scale using a series of frequency band filters. This axis scaling is done because humans perceive sound in the frequency domain on an approximately logarithmic scale. This phenomenon is caused by the spiral shape of the cochlea. Since many mammals have a similar physiology, this scaling is present in other animals, including elephants (Heffner and Heffner 1982). The output from the frequency band filters is then used as input for a discrete cosine transform, whose output provides cepstral coefficients. The cepstral coefficients are, in essence, capturing the general shape of the frequency spectrum. MFCCs are the most common feature used in speech processing today.

Since most energy in elephant vocalizations is concentrated below 1kHz, and for rumbles, below 100Hz, a modification had to be made to the above feature extraction algorithm. Because the FFT provides a frequency resolution between 0 and about 4000Hz (7518Hz sampling rate for the signal), the cepstral coefficients were capturing a lot of the noise information in the signal (2000Hz – 4000Hz). Therefore, instead of an FFT being used, a chirp Z-transform was used to interpolate values and find a frequency spectrum between 0 and 2kHz instead of 0 and 4kHz. This allowed the cepstral coefficients to capture the information in the range 0-1kHz better and therefore give better results.

MFCCs are not the only possible feature that can be used by this classification scheme, although at the current stage of this research, they are the only feature incorporated into the system besides log energy. Some feature extraction algorithms that are being developed are a fundamental frequency tracking algorithm and a feature such as the bicepstrum that contains phase information. The velocity and acceleration of features are also commonly included in the feature vector.

3. HIDDEN MARKOV MODEL (HMM)

HMMs are the most common model used in speech processing today. A typical HMM used for speech is shown in Figure 1. A HMM is a finite state machine where each state is represents a certain observation. Although other configurations are possible, for speech processing, each state's observation is represented by a probability distribution for each feature's value. This probability distribution is usually represented by a Gaussian Mixture Model (GMM). The mathematical representation of a GMM is:

$$\sum_i w_i N_i(\mu_i, \sigma_i)$$

To use an HMM for speech processing, the HMM is constrained to be a left-right HMM so that each state can represent a portion of the vocalization. Each state can transition to the right or back to itself and each of these transitions has a probability associated with it. Therefore, each HMM has means, variances, and transition probabilities as its parameters.

The HMM is trained using the Baum-Welch algorithm, which is an expectation maximization algorithm. Basically, the training examples are put through the HMM model, and the parameters are updated to maximize the probability that the training sequence fits the HMM. To evaluate a test example, the Viterbi algorithm, is used to find the most likely sequence of states (i.e. how long the vocalization spends in each state). A probability is also calculated which represents the probability with which the test example came from that HMM going through that sequence of states. More detailed information on training and evaluation of HMMs can be found in Rabiner and Jaung (1986).

4. RESULTS

4.1. Vocalization Type Classification

The vocalization type classification dataset consists of 74 vocalizations recorded at Walt Disney's Animal Kingdom in Orlando, FL. Researchers record the elephants each day while they are in one of the yards. The vocalizations were recorded from a collar designed and built by Walt Disney World Co. The collar transmitted the vocalizations back to elephant barn where there were recorded on DAT tape for about an hour each day. See Leong et al. (2002) for more information on the recording setup.

The best results acquired to date for the vocalization classification task are shown in Figure 2. This classification was done using 12 MFCCs calculated using a chirp Z-transform between 0 and 2kHz, frame sizes of 68ms with overlap of 34ms, and

		Classification				
		Croak	Rumble	Revv	Snort	Trumpet
Label	Croak	16	0	1	0	0
	Rumble	0	11	0	0	0
	Revv	1	4	7	1	1
	Snort	0	2	3	12	0
	Trumpet	1	1	0	2	11

Figure 2 – Type Classification Results

		Classification				
		Bala	Mackie	Moyo	Robin	Thandi
Label	Bala	7	0	1	2	1
	Mackie	0	8	0	0	0
	Moyo	0	0	9	0	4
	Robin	2	0	1	14	3
	Thandi	2	2	0	2	14

Figure 3 – Speaker Identification Results

a 5 state HMM. The grid cells on the main diagonal represent correct classifications and the other cells represent misclassifications. Leave-one-out verification was used, so all 74 vocalizations were classified given the other 73 as training data.

4.2. Speaker Identification

The vocalization type classification dataset consists of 72 vocalizations recorded at Walt Disney's Animal Kingdom in Orlando, FL. The vocalizations were recorded in the same fashion as for the vocalization type classification experiment.

The best results acquired to date for the speaker identification task are shown in Figure 3. This classification was done using 12 MFCCs calculated using a chirp Z-transform between 0 and 2kHz, frame sizes of 272.4ms with overlap of 136.2ms, and a 5 state HMM. The larger frame size was used because this dataset consisted of all rumbles which have a very low fundamental frequency. The grid cells on the main diagonal represent correct classifications and the other cells represent misclassifications. Leave-one-out verification was used for this experiment as well.

5. CONCLUSIONS

This paper explores the application of speech processing technology to the animal kingdom. Using typical speech processing features and models, African elephant vocalization type classification was done with an accuracy of 77.0% and speaker identification experiments resulted in an accuracy of 72.2%.

These systems are only applicable to elephant vocalizations. In fact, there are currently efforts underway to acquire beluga whale, tamarin, dolphin, and bird vocalizations. Even though each species has different vocal characteristics that make it challenging to analyze, these systems provide an adaptable standard framework that can be applied to other animals. There are also efforts underway to incorporate more traditional bioacoustic features such as bandwidth, vocalization duration, and fundamental frequency.

6. REFERENCES

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

Heffner, R. S. and Heffner, H. E. (1982). Hearing in the Elephant (Elephas maximus): Absolute Sensitivity, Frequency Discrimination, and Sounds Localization. *Journal of Comparative and Physiological Psychology*, **96(6)**, 926-944.

Leong, K. M., Ortolani, A., Burks, K. D., Mellen, J. D., & Savage, A. (2002). Quantifying acoustic and temporal characteristics of vocalizations for a group of captive African elephants (Loxodonta Africana). *Bioacoustics*, in press.

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

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

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

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