Chapter 2

LITERATURE STUDY

2.1 CHAPTER OBJECTIVES

In the outline of the study provided in the previous chapter, the research problem was defined as the development of an elephant recording collar, as well as methods of achieving automatic elephant rumble detection by adapting existing speech processing techniques. This chapter aims to give a background and critical discussion of the relevance of elephant vocalizations (Section 2.3) and discuss the history of elephant vocalization research as well as speech processing methods used to process and analyse vocalization recordings (Section 2.4). The wide range of speech processing techniques for VAD and the motivation for the chosen algorithm will be discussed in Section 2.5.

2.2 INTRODUCTION

Despite the fact that the vocalizations of African elephants have been extensively researched, the use of speech processing as an analysis tool for elephant sounds has not been widely explored or documented. The nature of this study necessitates a comprehensive overview of existing knowledge on both elephant vocalization research and certain speech processing techniques, and these aspects will subsequently be expounded in the following sections.

2.3 ELEPHANT VOCALIZATIONS

Elephants are highly social animals and have a complex social structure in which both long distance and short distance communication plays an important role (Langbauer Jr, 2000). They communicate using taste, smell, touch, sound and visual signs. As an example, an elephant bull can go into a state called musth where the testosterone levels rise, causing increased aggression and dominance. The bull will communicate this state to other elephants using specific odours, bodily positions and vocalizations (Poole and Moss, 1981; Rasmussen, 1988). Vocalizations are easier to observe over long distances than the other methods of communication and can be used by researchers to gain information about individual elephants as well as information about the group.

The best known elephant vocalizations are low frequency rumbles or higher frequency trumpets (McComb *et al.*, 2003), but researchers agree that elephants can produce at least 10 different sound types (Clemins and Johnson, 2003; Leong, Ortolani, Burks, Mellen and Savage, 2002; Soltis, Leong and Savage, 2005b). Most vocalizations are produced in the form of infrasonic rumbles which are too low in pitch to be easily perceived by humans. These infrasonic rumbles have a fundamental frequency of between 15 and 25 Hz and harmonics ranging several hundred Hz (Langbauer Jr, 2000). The harmonic frequency at the approximate level of 125 Hz has been shown to be the most important frequency needed for an elephant in the group to correctly establish the identity of the caller (Langbauer Jr, 2000). Figure 2.1 shows a spectrogram of a typical elephant rumble. It can be seen from the figure that almost all the spectral energy of the rumble occurs below 250 Hz. The harmonic nature of elephant rumbles can also be seen in the spectrogram as the yellow stripes that indicate higher spectral energy at multiples of the fundamental frequency. The harmonics of the elephant rumble are indicated by black lines.

Infrasonic elephant rumbles are an effective way of communicating over long distances. Calls can have a sound intensity of 117 dB SPL (Sound Pressure Level) at a distance of one metre (Langbauer Jr, 2000) with a reference sound pressure of 20 uPa. Low frequency rumbles are much less vulnerable to degradation due to the effects of deflection, refraction and absorption. This is due to the fact that low frequency sounds have long wavelengths which allow them to travel past objects relatively smaller than the wavelength itself (Hartmann, 1998). This implicates that these subsonic vocalizations are to a large extent immune to degradation and can travel distances far

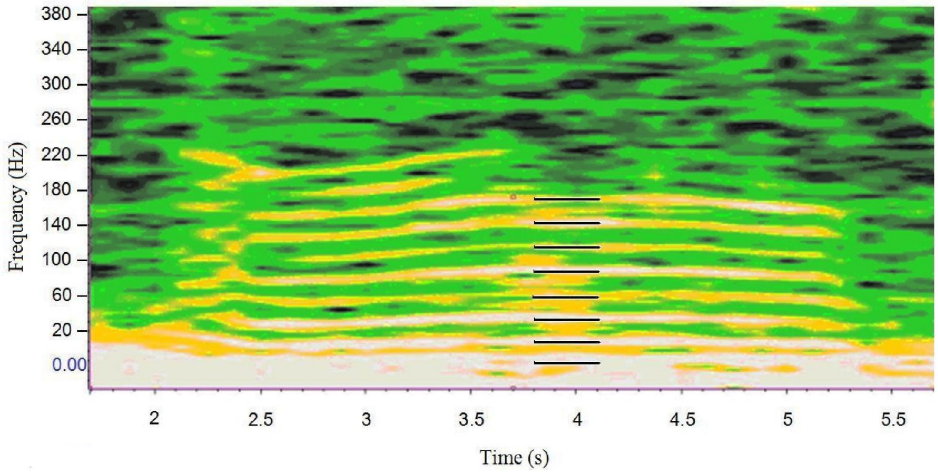

Figure 2.1: A spectrogram of a typical elephant rumble. The harmonics present in the rumble are indicated by black stripes.

exceeding sounds that consist of higher frequencies (like human speech). Experiments have shown that these rumbles can be recognized by elephants separated by a distance of 4 kilometres (Langbauer Jr, Payne, Charif, Rapaport and Osborn, 1991). Other reports have suggested that subsonic communication could even take place over distances of 10 kilometres if optimum conditions exist (Larom, Garstang, Payne, Raspet and Lindeque, 1997). As can be expected the sensitivity of hearing at low frequencies is very well developed in elephants. In fact, elephants have the best hearing at low frequencies of any mammal ever tested (up to 100 times better than humans) (Heffner and Heffner, 1982).

The importance of vocalizations to researchers in the field of elephant behaviour is largely due to the abundance of information that can be retrieved from it (Garstang, 2004; Langbauer Jr *et al.*, 1989; Langbauer Jr, 2000; McComb *et al.*, 2003; O'Connell-Rodwell, Arnason and Hart, 2000; Poole, Tyack, Stoeger-Horwath and Watwood, 2005; Wood *et al.*, 2005). The rate at which vocalizations are observed from an unseen group of elephants can be used to determine the size of a group as well as the number of males, females and calves in the group (Payne, Thompson and Kramer, 2003). Each elephant has specific voice characteristics which means that individuals may be recognized by their vocalizations (Clemins *et al.*, 2005; Clemins and Johnson, 2003; Soltis *et al.*, 2005b). Information about the sexual state of individual elephants can also be determined by analyzing their rumbles (Leong, Burks, Rizkalla and Savage, 2005; Poole, 1999; Soltis, Leong and Savage, 2005a). As is the case with humans,

some parameters of vocalizations can be used to determine the emotional state of an elephant (Clemins *et al.*, 2005; Soltis *et al.*, 2005b).

A large proportion of past research on elephant vocalizations involves the analysis of a collection of recordings. Two methods for recording elephant vocalizations have been identified in the literature. The first recording method entails the use of an RF transmitting collar, as reported by several researchers (Clemins *et al.*, 2005; Leighty, Soltis, Leong and Savage, 2008; Leong, Ortolani, Graham and Savage, 2003; Leong *et al.*, 2002; Soltis *et al.*, 2005a). These collars were all built by Walt Disney World Company Instrumentation Support (Division of Ride and Show Engineering) and were based on an earlier design by William Langbauer, Jr and Steven Powell. The collars consist of a condenser microphone, a radio transmitter and a battery pack. The sound picked up by the microphone gets transmitted to a radio receiver station where the sounds are recorded on a data tape. The recordings were done on captive elephants in Disney's Animal Kingdom. On days that recording sessions were scheduled the collars were mounted in the mornings. At a specific time of the day, the data tapes were activated to initialize a recording session with a duration of one hour. The collars were removed in the evenings so that the batteries could be recharged. This method appears to be effective for recording vocalizations from captive elephants since the elephants cannot move out of range of the receiver station and the batteries can be recharged daily. If a problem develops, for example if the microphone gets clogged with mud, the problem can be solved overnight and no more than a one hour recording session will be lost. The implementation of RF collars for recording wild elephant vocalizations could pose a number of problems. Firstly, wild elephants can travel vast distances in a single day. Elephants have been observed to walk up to 38 km in a 24-hour period (Viljoen and Bothma, 1990). Elephants that wear RF collars could walk out of range of the receiver station resulting in loss of data. To expand the range of an RF collar, a more powerful transmitter would have to be used, resulting in higher power consumption. In addition, wild elephants need to be tranquillized before a collar can be fitted. This is a costly procedure and requires ethical clearance. Therefore, the batteries that power the recording system should have a much greater capacity than those used for captive elephants, since frequent recharging of the batteries, requiring removal and refitting of the collar, is not a viable option in wild elephants.

The second method that has previously been used for elephant recordings is hand-held recordings. This method has been used to obtain vocalizations from wild elephants in

the Kruger National Park as reported by Wood *et al.* (2005). These researchers tracked a single herd of elephants in order to obtain regular recordings of their vocalizations. One of the female elephants in the herd was fitted with an GPS collar that sent the location of the herd to a cellular phone every morning. The researcher responsible for the recordings drove as close to the reported location as the road would permit and then proceeded on foot. A radio locator was used to find the exact position of the elephants, and a sound recorder with an external dynamic microphone was placed as close to the elephants as was safely possible. The movements of the herd had to be followed and recordings were stopped and restarted in the new location every time the herd moved (J.J. Viljoen, personal communication, September 7, 2005). The recorder saved the sound in the wave file format on a CompactFlash memory card. Under good recording conditions, the combined duration of all the recordings made in a day amounted to approximately an hour and a half. On extremely windy or rainy days no recordings could be made.

Besides the intensive field-work required by this method, an additional factor reduces the efficacy of this method. Because of the fact that the recording device needs to be adjusted to a high sensitivity to clearly record elephant vocalizations from a substantial distance, unwanted noises from sources much closer to the microphone can contaminate the recordings. This method of recording has not been successful in obtaining high quality, continuous recordings of elephant vocalizations in the wild.

For both of these recording methods (RF transmitter as well as hand-held recordings), the elephant vocalizations within recordings have to be identified by experts who manually examine spectrogram plots and listen to sped up versions of the recordings (Leong *et al.*, 2002; Poole, Payne, Langbauer Jr and Moss, 1988). When a recording is played at multiple speeds of its normal rate the inaudible low frequency elephant sounds get shifted up into the audible range of frequencies. This is a time consuming procedure, especially if long duration recordings have to be analysed. In addition, the recordings are often contaminated with noise in the infrasonic bandwidth range (Leong *et al.*, 2002), which makes the analysis of these recordings even more complex and time consuming.

In conclusion it can be said that elephants are social animals that communicate with one another in various ways, including vocalizations. Infrasonic vocalization is a very effective way of communicating over both short and long distances. The wealth of

information present in vocalizations as well as the fact that vocalizations are easier to observe than other means of elephant communication render infrasonic vocalization a valuable and useful tool in the study of elephant behaviour. However, previously documented methods used to record these vocalizations have been found to have several limitations, and novel methods should be explored in this regard.

2.4 ELEPHANT VOCALIZATIONS AND SPEECH PROCESSING

In view of the limitations of manual analysis of elephant vocalization recordings, the possibility of using speech processing techniques on these recordings should be considered. Infrasonic elephant rumbles are produced by vocal cords (Garstang, 2004; McComb *et al.*, 2003; Soltis *et al.*, 2005b) so that it may be expected that the resulting sound would have characteristics similar to voiced human speech. In fact, studies have shown that most mammalian vocal production and reception systems are extremely similar (Bradbury and Vehrencamp, 1998; Titze, 1994). This idea is supported by the harmonic nature of the elephant rumbles and justifies the application of speech processing techniques to bioacoustics.

The only known scientific publication where speech processing techniques were used on elephant vocalizations is that of Clemins *et al.* (2005) and Clemins and Johnson (2003). Automatic classification and speaker identification were conducted on a collection of vocalizations and rendered promising results. These vocalizations were recorded from captive elephants by the RF elephant collar discussed previously in this chapter. The classification of elephant vocalizations was similar to speech recognition done on human speech. Five basic elephant vocalizations were classified in the experiment. This was achieved with by using 12 Mel-Frequency Cepstral Coefficients as features and computing log energy by using a shifted filter bank to compensate for the infrasonic range in which elephant vocalizations occur. For the speaker recognition experiments, Hidden Markov Models (HMMs) were used for the modelling of the different speakers. The success rate of the call type classification experiment was 83.8% and that of the speaker recognition was 88.1%. It was noted that in most bioacoustics studies, vocalizations are divided into groups of varying quality, and ultimately only the categories with the best quality recordings are used. In the study reported by Clemins

et al. (2005), however, the number of recorded vocalizations available were too small to allow exclusion of low quality recordings and these were therefore included in the analyses.

The shortage of a sufficient number of vocalizations in the study of Clemins *et al.* (2005) indicates the need for an elephant vocalization recording system that can gather a large amount of continuous acoustic data. A further limitation in the method of these researchers was the need for manual location and isolation of the vocalizations from the recordings. This is a time consuming process, especially for studies where much larger numbers of vocalizations are needed. These limitations underscore the need for the automatic detection of infrasonic elephant vocalizations.

2.5 VAD TECHNIQUES

The challenge of detecting elephant rumbles from recordings appears to be similar to that of identifying voiced human speech from personal audio recordings. Specifically it suggests that the techniques used for Voice Activity Detection (VAD) of human speech may also be suitable for elephant rumble detection. Because of the periodic nature of voiced speech, one way of detecting the presence of speech during a particular interval may be to use a pitch detection (or determination) algorithm. In order to achieve automatic detection of infrasonic elephant rumbles, techniques should be used that capitalize on the specific characteristics of these rumbles that set them apart from background noise, such as the harmonic content of the rumbles.

Most existing applications of pitch detection algorithms used in voice activity detectors as published in literature are limited to clean speech (noiseless speech) in a telecommunications context (Wu *et al.*, 2003). It is, however, more difficult to extract pitch from a recording where numerous other sources of noise are present, as would typically be the case for the elephant call recordings. Elephant recordings typically contain unwanted sounds including bird calls, motor vehicles, wind, walking elephants and other sounds that occur in the wild. A noise robust pitch detection algorithm will be needed to effectively detect periodicity in the noisy recordings.

Pitch detection algorithms are generally classified into three categories namely time-domain (Hung, 2002; Kunieda, Shimamura and Suzuki, 2000; Shimamura and Kobayashi,

2001; Takagi, Seiyama and Miyasaka, 2000), frequency-domain (Davis, Nordholm and Togneri, 2006; Li, Zhang, Cui and Tang, 2005; Woo, Yang, Park and Lee, 2000; Zhang, Zhang, Lin and Quan, 2006) and time-frequency domain algorithms (Wu *et al.*, 2003; Zhao and Ogunfunmi, 1999). Time-domain pitch detection algorithms consider the temporal structure of the waveform. Peak and valley positions, zero-crossings and autocorrelations are used for detecting the pitch period. The simplest and computationally most inexpensive technique for determining pitch would be a simple count of the number of times that the signal crosses the zero reference. This technique is highly inaccurate when the signal contains noise or in the case of a harmonic signal when the fundamental frequency is less energetic than any of the higher harmonics. For this reason, such a technique will not be suitable for pitch detection of elephant vocalizations.

The purpose of autocorrelation routines used as part of time-domain pitch detection is to find the similarity between a signal and a shifted version of the same signal. This is based on the premise that a periodic signal will thus have a periodic autocorrelation function, and a harmonic signal will have an autocorrelation function with peaks at multiples of the fundamental frequency. This technique works well with lower frequencies and is popular in speech processing techniques where the pitch range is limited. The ability of the technique to extract the pitch of a sound with a harmonic structure is attractive, but the addition of noise to a signal degrades the definition of the peaks of the autocorrelation function and diminishes the accuracy of the technique.

Frequency domain pitch detection algorithms typically detect the fundamental frequency by examining the harmonic structure in the short-term spectrum. The fundamental frequency can be determined by computing the greatest common divisor of the frequencies of the higher harmonic components. The greatest common divisor is determined by filling in a frequency histogram for each harmonic frequency and at integer divisions of the harmonic frequency. The greatest frequency peak of the histogram represents the greatest common divisor, and thus the fundamental frequency. Even though this technique is computationally inexpensive, the addition of narrow band noise to the signal or the evaluation of a sound with a changing number of harmonics degrades its performance. The cepstrum, a second order transform of the power spectrum, is also widely used for pitch determination (Ahmadi and Spanias, 1999; Kim and Chung, 2004; Nadeu, Pascual and Hernando, 1991; Noll, 1967; Seiyama, Tohru, Tetsuo and Eiichi, 1992; Zhao and Ogunfunmi, 1999). The term “cepstrum” is formed

by reversing the first four letters of "spectrum". The Fourier transform of a signal is taken to the log-magnitude Fourier spectrum. By implication, if the original spectrum comes from a harmonic signal, the frequency representation will be periodic, so when the FFT (Fast Fourier Transform) is taken again it will result in a peak corresponding to the fundamental frequency. The cepstrum technique can also be seen as a de-convolution procedure. If the original signal is seen as an impulse train that has been convolved with a filter, it results in multiplication in the frequency domain. Applying the log operation translates the multiplication to an addition operation. When the FFT is then applied once more it results in the de-convolution of the original signal which gives us the fundamental frequency. This technique is used with great success for determining the pitch of noiseless speech, but once again, when noise is added to the original signal, the peak indicating the fundamental frequency fades away.

An additional technique used as part of frequency domain pitch detection algorithms is the use of statistical properties of the communication channel and the expected speech signal (Chang, Kim and Mitra, 2006; Chang, Shin and Kim, 2004; Davis *et al.*, 2006; Sohn, 1999). The success of these pitch detection techniques are to a great extent dependant on a priori knowledge of the statistical behaviour of the communication channel as well as the expected speech signal. Often, these systems need to be trained prior to operation. For example, it might be required that a number of both male and female participants repeat certain words in different tones of voice to acquire the statistical models needed for correct operation of the algorithm. The fact that the elephant vocalizations will be recorded in free field and that system training data is not easily obtainable renders this technique unsuitable for use with elephant vocalizations.

Time-frequency domain algorithms first filter the original signal into sub-bands and then perform time-domain analysis on the band-filtered signals (Wu *et al.*, 2003). By filtering sound into sub-bands this technique resembles the way that humans perceive pitch. The pitch of complex harmonic and inharmonic signals can be identified correctly, and is robust in the face of noise and phase changes. Although this technique is computationally expensive, its characteristics makes this type of pitch determination algorithm ideal for use in elephant recordings.

A robust pitch detection algorithm proposed by Wu *et al.* (2003) for use in human voice activity detection in various background noise conditions is an example of a time-frequency pitch detection algorithm. This technique will be discussed in more

detail since it has already been implemented successfully to detect human speech from real life recordings with various background noises (Lee and Ellis, 2006). The reported success of this method despite the presence of background noise makes it an instinctive favourite to use for extracting elephant vocalizations in similar circumstances. The technique features the desired characteristics of previously discussed methods, but outperforms those techniques when a noisy signal is used.

The noise robustness of this algorithm is primarily due to the input signal being divided into a number of sub-bands and only the sub-bands with good signal to noise ratio being used in pitch determination. This perceptual pitch detector combines a cochlear model with a bank of autocorrelators. The algorithm uses the steps shown in Figure 2.2 for determining a pitch track.

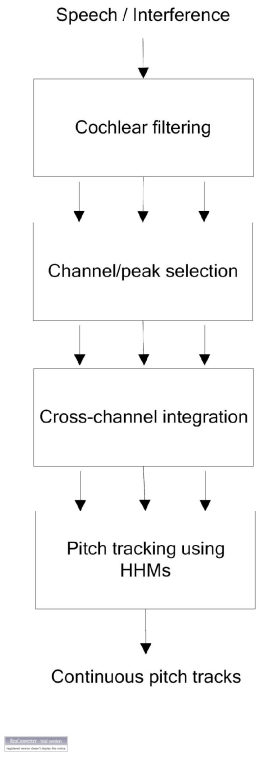

Figure 2.2: The schematic diagram of the model proposed in Wu *et al.* (2003).

The speech signal enters the system where it is firstly filtered into sub-bands by a cochlear type filter. This filter bank comprises an array of fourth order gammatone filters, which is a standard model for cochlear filtering. In fact, the particular sensitivity of the human cochlea to pitch was used as a template for the development of this

filter.

A normalized autocorrelation of each of the resulting channels is then calculated. Noisy channels are discarded and the remaining channels are integrated and used to find an estimation of the pitch present in the original signal. An HMM is subsequently used to form continuous pitch tracks. Some of the same principles used in this algorithm were used as a starting point for the automatic detection of infrasonic elephant rumbles, as will be described in Chapter 3.

2.6 SUMMARY

Chapter 2 presents the literature on which this study is based. The importance of vocalizations and specifically infrasonic rumbles to elephant research were highlighted. Existing methods for recording elephant vocalizations were evaluated. The critical discussion of existing literature demonstrated the potential value of applying speech processing techniques to elephant vocalizations. In addition, the limited application of this technique underscores the need for research that will explore the value thereof in the study of elephant vocalizations. In light of this need, the study at hand will aim to develop an automatic elephant rumble detector based on an existing VAD algorithm that has been successfully used to extract human speech from noisy recordings.