2018

SPEAKER AND GENDER IDENTIFICATION USING BIOACOUSTIC DATA SETS

Neenu Jose
University of Kentucky, njo239@uky.edu
Digital Object Identifier: https://doi.org/10.13023/ETD.2018.223

STUDENT AGREEMENT:

I represent that my thesis or dissertation and abstract are my original work. Proper attribution has been given to all outside sources. I understand that I am solely responsible for obtaining any needed copyright permissions. I have obtained needed written permission statement(s) from the owner(s) of each third-party copyrighted matter to be included in my work, allowing electronic distribution (if such use is not permitted by the fair use doctrine) which will be submitted to UKnowledge as Additional File.

I hereby grant to The University of Kentucky and its agents the irrevocable, non-exclusive, and royalty-free license to archive and make accessible my work in whole or in part in all forms of media, now or hereafter known. I agree that the document mentioned above may be made available immediately for worldwide access unless an embargo applies.

I retain all other ownership rights to the copyright of my work. I also retain the right to use in future works (such as articles or books) all or part of my work. I understand that I am free to register the copyright to my work.

REVIEW, APPROVAL AND ACCEPTANCE

The document mentioned above has been reviewed and accepted by the student's advisor, on behalf of the advisory committee, and by the Director of Graduate Studies (DGS), on behalf of the program; we verify that this is the final, approved version of the student's thesis including all changes required by the advisory committee. The undersigned agree to abide by the statements above.

Neenu Jose, Student

Dr. Michael T. Johnson, Major Professor

Dr. Caicheng Lu, Director of Graduate Studies

SPEAKER AND GENDER IDENTIFICATION USING BIOACOUSTIC DATA SETS

THESIS

A thesis submitted in partial fulfillment of the requirements for the degree of Master of Science in Electrical Engineering in the College of Engineering at the University of Kentucky

By
Neenu Jose
Lexington, Kentucky
Co-Director: Dr. Michael T. Johnson, Professor of Electrical and Computer Engineering
and Dr. Kevin Donohue, Professor of Electrical and Computer Engineering
Lexington, Kentucky
2018

ABSTRACT OF THESIS

SPEAKER AND GENDER IDENTIFICATION
USING BIOACOUSTIC DATA SETS

Acoustic analysis of animal vocalizations has been widely used to identify the presence of individual species, classify vocalizations, identify individuals, and determine gender. In this work automatic identification of speaker and gender of mice from ultrasonic vocalizations and speaker identification of meerkats from their Close calls is investigated. Feature extraction was implemented using Greenwood Function Cepstral Coefficients (GFCC), designed exclusively for extracting features from animal vocalizations. Mice ultrasonic vocalizations were analyzed using Gaussian Mixture Models (GMM) which yielded an accuracy of 78.3% for speaker identification and 93.2% for gender identification. Meerkat speaker identification with Close calls was implemented using Gaussian Mixture Models (GMM) and Hidden Markov Models (HMM), with an accuracy of 90.8% and 94.4% respectively. The results obtained shows these methods indicate the presence of gender and identity information in vocalizations and support the possibility of robust gender identification and individual identification using bioacoustic data sets.

KEYWORDS: Speaker identification, gender identification, Gaussian Mixture Model (GMM), Hidden Markov Model (HMM), Mice, Meerkat.

Neenu Jose

May 2018

SPEAKER AND GENDER IDENTIFICATION USING BIOACOUSTIC DATA SETS

By

Neenu Jose

Co-Director of Thesis: Dr. Michael T. Johnson

Co-Director of Thesis: Dr. Kevin Donohue

Director of Graduate Studies: Dr. Caicheng Lu

Date: May 2018

To my two wonderful boys...

Acknowledgement

I would like to thank my advisor, Dr. Michael T Johnson, for his immense support and guidance throughout my research at the University of Kentucky. I also would like to thank my graduate committee members, Dr. Kevin Donohue and Dr. J Robert Heath, for their willingness to provide feedback toward my thesis and defense completion.

I would like to thank Dr. Micheal Dent and Kali Burke of University of Buffalo for sharing the Mice data and helping me with all my questions. I would also like to thank Dr. Martha Manser and Ariana Strandburg-Peshkin of University of Zurich for sharing the meerkat data.

I would like to thank my parents, my family, and my friends for their love and support during my University years. Finally, thanks to God for all his blessings, without whom this work would not have been possible.

Table of contents

List of tables

List of figures

Chapter 1: Introduction

1.1. Background and motivation

The research work presented here focuses on analysis of ultrasonic mice vocalizations (Zippelius and Schleidt 1956, Sales 1972) and meerkat vocalizations (Clutton-Brock, Russell et al. 2005), with an emphasis on individual identity and gender classification. Such vocalizations may provide insights for studies of genetic foundations of vocal communication in humans (Fischer and Hammerschmidt 2011) and can be used for understanding animal behavior.

Mice are the most commonly used species in biomedical research, neuroscience and experimental psychology. The facts that they are inexpensive, easy to handle and have 98% genetic overlap to human genes makes them ideal candidates for research about various human conditions.

Meerkats are socially obligated, cooperatively breeding, highly territorial mammal who live in groups (Clutton-Brock, Russell et al. 2005). Meerkats forage as a group and has a highly developed vocal communication system which help them to coordinate group movements, identify predators and maintain group cohesion. In this study analysis of Close calls are used for gender and individual identification. Close calls are low amplitude pulsated calls used for group cohesion and are encoded with gender, individuality and group signature (Townsend, Hollén et al. 2010, Townsend, Allen et al. 2012, Mausbach, Braga Goncalves et al. 2017). The study of meerkat vocalizations might help us understand the social dynamics and social learning in species that live in small groups.

Bioacoustics is a multidisciplinary area of research and requires extensive manual labor to perform basic tasks such as detection, segmentation and manual labeling of voice activity from long recordings of data from the field. The three main tasks involved in the automated analysis of bioacoustic signals are detection, classification, and clustering of vocalizations from noisy recordings. Each of these will be described in more detail in 2.3.1.

Detection is the process of identifying the presence of a particular type of vocalization, including start and end points of each vocalization. In contrast, classification involved dividing vocalizations into categories such as call type, species, speaker, gender, and behavioral patterns Classification and detection algorithms are usually trained using supervised learning approaches which builds models out of expertly labeled data. In contrast, unsupervised clustering groups vocalizations into categories based on similarity with the goal of separating groups and determining the number of such groups present based on some threshold criterion, without any predefined categories. Unsupervised clustering can also be applied to the individual identification of vocally active species.

Individual identifying information within vocalizations occurs when interindividual variation in a vocalization exceeds intraindividual variation in that vocalization, as a result of temporal or spectral variations in vocalizations (Pollard and Blumstein 2011). Individual vocal distinctiveness has specific communicative function, being essential for species living in larger groups where individual interactions are more important for offspring and mate recognition, territorial or coalitional behaviors, signaler reliability assessment, and social hierarchies (Pollard and Blumstein 2011). In addition, individual vocal distinctiveness is related to non-communicative characteristics as well, such as simple physiological differences within the vocal production mechanisms such as

body size and shape. Corresponding to this, speaker specific acoustic features are observed in many species, for example individual identity cues have been found in birds (Adi, Johnson et al. 2010), mammals (Clemins, Johnson et al. 2005, Volodin, Lapshina et al. 2011) and marine mammals (Brown, Smaragdis et al. 2010).

Like individual differences, gender specific differences in animal vocalizations can differ in two ways, by acoustic shape or by sequence or timing of delivery (Green 1981). Green categorized gender differences in vocalizations as vocalizations which are produced by both sexes, but which differ in acoustic shape due to sexual dimorphism, vocalizations that are present in one gender but entirely absent in the other and vocalizations which are produced by both sexes but have different purposes. Gender identification using vocalizations can be seen in many species, for example, gender specific vocalization patterns have been found in birds, for example black-capped chickadee songs and oriental white stork (Eda-Fujiwara, Yamamoto et al. 2004, Hahn, Krysler et al. 2013) and in mammals, baboons and goitred gazelles (Rendall, Owren et al. 2004, Volodin, Lapshina et al. 2011).

In animal bioacoustics, vocalizations can be analyzed by both qualitative and quantitative methods (Terry, Peake et al. 2005). In the qualitative approach, visual examination of spectrograms or listening in the field is used for identification. A spectrogram is a visual representation of energy present in various frequencies of acoustic waveform over time. Listening in the field needs extensive experience and is limited to a small number of speakers. The most commonly used qualitative method is visual analysis of spectrograms, since humans have good skills at pattern recognition. Visual analysis has only modest accuracy, and thus quantitative analysis will often follow qualitative analysis.

In quantitative analysis, there are several analysis methods, from simple statistical methods to automatic methods. Acoustic features used for simple statistical methods include initial, final, mean, minimum and maximum frequencies and duration (Shapiro 2010), which can be directly measured from spectrograms. The features are extracted from entire vocalizations and are fed into statistical analysis tools like Discriminant function analysis (DFA) (Favaro, Gamba et al. 2015), stepwise discriminant function analysis (SDFA) (Hoffmann, Musolf et al. 2012), Principal coordinates analysis (Charrier and Harcourt 2006), multivariate analysis of variance (MANOVA) or analysis of variance (ANOVA) (Boughman and Wilkinson 1998). Since these methods use statistics on acoustic features taken from individual vocalization frames, the disadvantage of these methods is that they fail to incorporate information about the temporal patterns of the vocalizations.

Another simple method for classification which does incorporate such temporal information is template matching, where a target vocalization is selected as a template and cross-correlated against test vocalizations. There are two common types of template matching techniques, spectrogram cross-correlation (SCC) and matched filtering. Spectrogram cross-correlation operates in the spectral domain and matched filtering operates in temporal domain. The main disadvantage of template matching is that small fluctuations in vocalizations can result in negative correlation.

Using automatically extracted acoustic features like Greenwood Function Cepstral Coefficients (GFCC) (Clemins, Trawicki et al. 2006) and generalized Perceptual Linear Prediction coefficients (gPLP) (Clemins and Johnson 2006, Clemins, Trawicki et al. 2006), more powerful statistical classification methods are possible. GFCC and gPLPs are generalized forms of Mel frequency Cepstral Coefficients (MFCC) (Davis and

Mermelstein 1980) and Perceptual Linear Prediction (PLP) (Hermansky 1990) coefficients respectively, extracted automatically using frame-based processing of vocalizations.

MFCCs and PLPs are currently the most widely used feature extraction methods for human speech. For the speech recognition task in particular, Hidden Markov Models (HMM) using MFCC features, sometimes modeled statistically and sometimes using deep neural networks, are the most common approach (Gales and Young 2007). HMMs model the temporal variation of the vocalizations as states, and each state has a statistical model of acoustic features for the temporal pattern that state represents.

Many studies have proven successful application of HMMs and GFCCs in Bioacoustics (Li, Tao et al. 2007, Ren, Johnson et al. 2009, Adi, Johnson et al. 2010). Another method for classification is dynamic time warping (DTW) (Sakoe and Chiba 1978), a template-based method that uses a dynamic programming algorithm. DTW estimates the lowest distance path by aligning test vocalization against a template vocalization.

Many recent developments in human speech analysis have used Deep Neural Network (DNN) (Hinton, Deng et al. 2012) based HMMs. The underlying models are HMMs for temporal representation just like HMM-GMMs, but state observation probabilities are modeled by DNNs instead of GMMs.

1.2. Contributions and significance

This study focuses on applying speech processing techniques to bioacoustics, specifically individual and gender classification of ultrasonic mice vocalizations and speaker identification with meerkat vocalizations. Features of the vocalizations are

extracted using Greenwood Function Cepstral Coefficients, with a classification model based on statistical Gaussian Mixture Model discrimination and Hidden Markov Models.

Although speaker identification and gender identification in mice and speaker identification in meerkats using Gaussian Mixture Models and Hidden Markov Models is the main contribution of this work, this model can be extended to any species. This work may contribute to understand behavior and communication among mice. Since mice serve as models for biomedical research, this work may help better understand the evolution of vocal communication in humans and other terrestrial mammals. Meerkats are among the most social mammals with a rich vocal repertoire, which makes them a model to understand the evolution of social behavior, animal communication and cognition. Speaker identity and gender identity have been explored in a wide variety of species, and here we extend this to ultrasonic mice vocalizations and meerkat vocalizations through a speech processing-based approach to bioacoustics classification.

1.3. Plan of thesis

The remainder of this thesis is organized as follows: Following this introduction, Chapter two gives a brief background description of the technical areas of speech processing, feature extraction, machine learning, bioacoustics, individual recognition and gender recognition. Chapter three introduces the GMM based method for identifying gender and speaker from ultrasonic vocalizations of mice, the data and experimental methodology to be used, and gives a detailed review of the results of the study. Chapter four introduces the GMM and HMM based method for identifying speaker from Close calls of meerkats, the data and experimental methodology to be used, and gives a detailed review

of the results of the study. Chapter five concludes the thesis and describes contributions and future work.

Chapter 2: Background and related works

2.1. Overview

This chapter provides a broad overview of the fields of study connected to this research work, including an introduction to topics in speech processing and bioacoustics. The specific tasks associated with this work are speaker identification and gender identification of animal vocalizations, so a particular focus will be given to these two topics and to applications of speech technology to animal vocalizations.

The first section of the chapter gives an overview of the source filter model of speech production and introduces the basic concept of frame-based speech processing and feature extraction, with a focus on Mel Cepstral Coefficients. Also, the basic theory of Gaussian Mixture Models, commonly used for both gender and speaker identification, is explained.

The second section gives a brief overview of bioacoustics, with a focus on the bioacoustic tasks and analysis approaches as well as applications of speech processing techniques to problems in animal vocalization analysis and classification. This section covers the Greenwood Function Cepstral Coefficients for the feature extraction from bioacoustic signals and mentions previous studies of classification using GMMs in bioacoustic field.

The third section deals with the production, usage and vocal repertoire of the mice Ultrasonic vocalizations and meerkat vocalizations involved with the present study, with a summary and conclusions in the final section.

2.2. Overview of speech processing and analysis

Human speech processing started long before the advent of the computer. As early as 1791, there were attempts to implement speech synthesis using a mechanical speech synthesizer that could produce vowels and consonants (Benesty, Sondhi et al. 2007). The highly acclaimed VOCODER by Dudley in 1930, which can produce arbitrary sentences, marked the beginning of modern era of speech processing. Today, speech processing has become a part of everyday life, with speech recognition integrated into smartphones and many other devices. Research in speech processing continues in the areas of Automatic Speech Recognition (ASR), automatic speaker identification and verification, gender identification, speech enhancement, speech coding speech synthesis, language modeling and machine translation. The two main areas of speech research covered in this work are speaker identification, the determination of which individual is vocalizing, and gender identification, the determination of the gender of the individual vocalizing.

2.2.1. Speech production

The human speech system consists of phonation organs (lungs and larynx) and articulatory organs (lips, tongue and teeth). Forced air from the lungs vibrates the vocal folds in the larynx to generate the excitation signal. The vocal tract and articulators filter the excitation signal, producing many different types of sounds. Humans produce two basic categories of sounds, voiced and unvoiced, depending on the vibratory status of the excitation signal. Voiced speech has a nearly periodic input excitation signal and the unvoiced signals are produced by a pseudo white noise excitation signal. Production of voiced excitation happens when forced air from lungs build up a pressure beneath closed vocal folds until the pressure forces these to open. When the pressure beneath the vocal

folds returns to normal, the vocal folds close from muscle tension and this cycle repeats. This process generates a quasi-periodic airflow which is the excitation signal for voiced speech. In contrast, unvoiced sound is produced by forcing air through an open vocal fold, as a result the excitation signal generated is a white noise.

The voiced and unvoiced sounds are modified by the movements of vocal tract and articulators, such as the lips and tongue. The basic unit of speech is called a phoneme, which can be considered as a unique set of articulatory gestures within the vocal tract and excitation characteristics, that together create an acoustic signal that differentiates meaning within a language. American English has around 42 phonemes, classified as vowels, semivowels, diphthongs and consonants. Commonly used phonetic alphabets include the International Phonetic Alphabet (IPA) and the ARPAbet, developed by Advanced Research Projects Agency (ARPA) (Deller, Hansen et al. 2000).

From a speech processing perspective, the overall process of speech production can be represented as a source-filter model. A source represented by air flow through vocal folds being filtered by the resonances of vocal tract generates the speech signal.

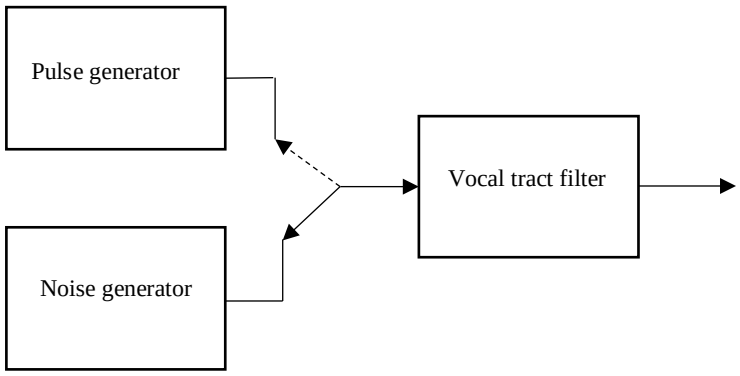

Figure 1: Source filter model of speech production

The mathematical representation of the source filter model for speech production can be given as

$$e[n] \otimes h[n] = s[n], \tag{2.1}$$

where the excitation $e[n]$ and filter $h[n]$ are convolved to produce speech signal $s[n]$.

2.2.2. Spectral analysis and feature extraction

Since speech is produced from a time varying vocal tract excited by time varying excitation signal, speech is nonstationary by nature. Thus, the spectral properties of speech are time varying and short-term processing is used as an analysis tool. In short term processing a sliding window approach is used, where each individual window, or frame, is assumed to be stationary (Deller, Hansen et al. 2000). Each frame is a product of shifted window $w[n]$ with original speech signal $s[n]$,

$$f[n;m] = s[n]w[n-m]. \tag{2.2}$$

Speech processing systems for speaker identification and gender identification consists of feature extraction, acoustic modeling and statistical classification. Frame sizes of 10 – 30 ms are used in human speech processing in order to approximate the stationarity for spectral analysis. Features are extracted from each frame and combined to form a feature vector. This feature vector is then used as an input to the classification system. Although most speech processing systems use spectral domain features, temporal domain features such as short-term energy and zero crossing rate can also be useful for some applications.

According to the source filter model, speech is produced by the convolution of an excitation signal with the impulse response of the vocal tract. Although direct spectral

analysis through methods such as the Fourier Transform provides good information about speech characteristics, homomorphic techniques such as cepstral analysis are often used for various speech processing techniques because of their ability to separate the convolutional mixture of excitation signal and vocal tract response. In the cepstral domain, the excitation signal and the vocal tract response are linearly combined and occupy different regions of the cepstral domain, making them easy to separate (Deller, Hansen et al. 2000).

Mathematically, the cepstrum of a signal is the inverse Fourier transform of the logarithm of the Fourier Transform Magnitude of the signal. This is represented mathematically as

$$c(n) = F^{-1}(log\,|\,F(s[n]\,|), \tag{2.3}$$

$$= F^{-1}(log\,|\,S[m]\,|), \tag{2.4}$$

$$= F^{-1)}(log\,|\,E[m]H[m]\,|), \tag{2.5}$$

$$= F^{-1}(log\,|\,E[m]\,| + log\,|\,H[m]\,|), \tag{2.6}$$

where $s[n]$ is the signal and F represents the Fourier transform operation. The logarithm operation acting on the real, positive Fourier Transform magnitude separates the convolved excitation signal and vocal tract response into summed components in the cepstral domain. Liftering, defined as splitting different regions of the cepstrum, then separates these components. Note that the terms cepstrum and liftering are derived from swapping letters in spectrum and filter respectively.

Feature extraction plays an important part in accurate classification of automatic speech processing systems. Feature extraction is the process of estimating a reduced set of relevant variables that will be effective for further analysis, modeling, and classification.

Although there are many feature extraction methods for speech like Linear Prediction Coefficients (LPCs), Perpetual Linear Prediction (PLP) coefficients, and Mel Frequency Cepstral Coefficients (MFCCs), MFCCs are the most commonly used feature extraction method for speech recognition. MFCCs operate in the cepstral domain, making it easy to separate the linear combination of excitation and vocal tract characteristics. MFCCs use a nonlinear frequency scale which represents the human auditory system (Huang, Acero et al. 2001) which makes them an ideal candidate for speech processing applications.

A block diagram for calculating MFCCs is given in Figure 2. The speech signal is divided into frames generally using a sliding window and the log magnitude spectrum of each frame is warped according to the Mel frequency scale. The Discrete Cosine transform (similar to the inverse Fourier Transform operation, but with real rather than complex coefficients) of the warped frequency log magnitudes yields Mel frequency cepstral coefficients. Each step in calculating MFCCs is explained below in detail.

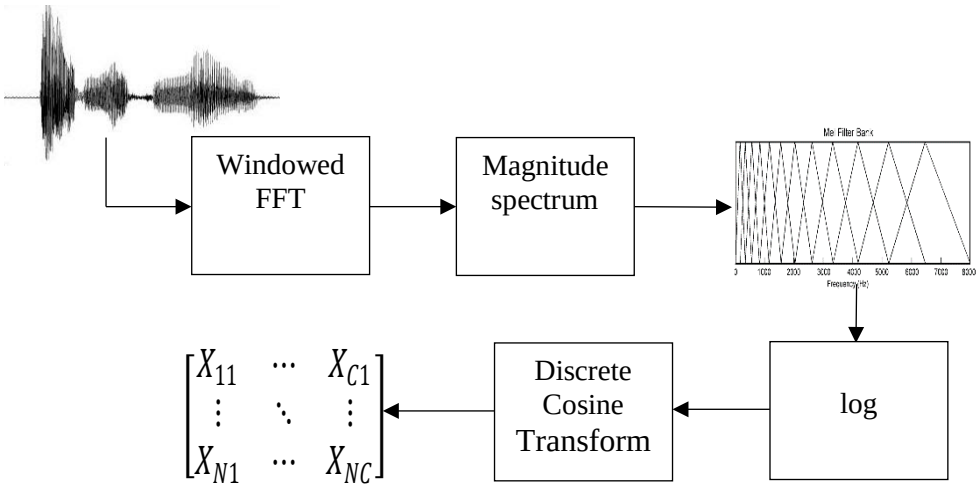

Figure 2: MFCC Block diagram

The speech signal is divided into frames using a windowing function, typically using a hamming window. A window size of 20 – 30ms is used in human speech processing, to trade off the typically syllable rate of human speech and the need for larger frames for accurate frequency resolution. Choosing the frame length is a tradeoff between spectral and temporal resolution, with a lengthy frame yielding better spectral resolution but losing the stationarity of the speech signal. A narrow frame gives better temporal resolution but poor spectral resolution.

Following the calculation of the log-magnitude Discrete Fourier Transform (DFT), the next step in finding MFCCs is to warp the DFT output to the Mel scale (Stevens and Volkmann 1940) using filter bank analysis. The human hearing system is linearly sensitive to frequencies below 1000 Hz and logarithmically sensitive to frequencies above it. The Mel scale successfully models the non-linearity in human speech perception. Mel scale is defined as

$$f_{mel} = 2595 \log(1 + \frac{f}{700}), \tag{2.7}$$

where f is the frequency in Hz and f_{mel} is the Mel frequency. Mel scale is often calculated using a filter bank as shown in Figure 3.

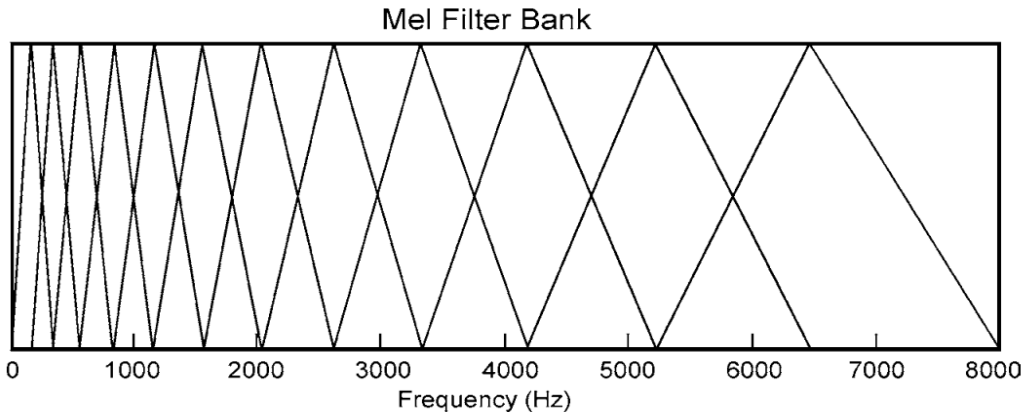

Figure 3: Mel Filter Bank

Generally triangular band pass filters are used with higher number of filters in the lower frequency region and lower number of filters in higher frequency region. Other filter shapes can be used and are more reflective of the non-linear shape of critical-band filtering in the human auditory system, but the simplicity of triangle filters makes them an attractive option which is often used in practice. Log spectral energies in each filter are calculated by taking the logarithm of sum of coefficients after multiplying each filter with the magnitude spectrum of the frame.

$$\Theta[m] = ln[\sum_{k=0}^{N-1} | X[k] |^2 H_m[k]], 0 \le m \le M \tag{2.8}$$

where M is the number of filters and $H_M[k]$ is the filter bank. At lower frequencies the filter is narrow and is more sensitive to spectral energy variations. At higher frequencies, the filters get wider and less sensitive to spectral energy variations. These match the sensitivity of human hearing.

The final step in the MFCC calculation is taking the Discrete Cosine Transform (DCT) of the log spectral energies of each filter. The DCT acts similarly to the inverse FT of the cepstral calculation, but with real coefficients. The DCT decorrelates the filter bank

magnitudes and creates a compact and efficient Mel cepstral representation (Davis and Mermelstein 1980). The DCT is defined as

$$c[n] = \sum_{m=0}^{M-1} \Theta[m] \cos(\pi n(m+0.5)/M), 0 \le n \le M. \quad (2.9)$$

The time derivatives of MFCCs also provide useful information about trajectories of MFCCs over time (Yang, Soong et al. 2007). The dynamic time derivative features taken along with the static features MFCC provide better recognition in automatic speech analysis. The time derivative can be calculated as

$$d_t = \frac{\sum_{k=1}^{N} k(c_{t+k} - c_{t-k})}{2\sum_{k=1}^{N} k^2} \quad (2.10)$$

where d_t is a delta coefficient at time *t* computed in term of the static coefficients c_{t-k} to c_{t+k}. The second derivative, known as a delta-delta coefficient, can be calculated by applying a similar computation to delta coefficients.

Another feature that can be used in speech and speaker recognition problems is short term energy. Short term processing of speech can be used to find the short-term energy of a speech frame. Since speech is a time varying signal, the energy associated with speech is also time varying. Voiced speech will have higher energy compared with unvoiced speech, and different phonemes have different average energies. Thus, the short-term energy can be an important feature to include for analysis and classification. The short-term energy of N length frame ending at time m,

$$E(m) = \sum_{n=m-N+1}^{m} S^2[n] \quad (2.11)$$

2.2.3. Acoustic modeling

Acoustic models are representations constructed from the features extracted from speech signals. An automatic speech processing system uses such models for comparison of feature vectors and pattern recognition. Template, statistical, and machine learning models may all be used for acoustic modeling. Template based models matches selected speech templates with test templates and calculates distance between both. Common template matching models include Spectrogram Cross Correlation (SCC), Matched Filtering and Dynamic Time Warping (DTW). In contrast, statistical methods like Gaussian Mixture Models (GMMs) characterize the statistical properties of speech signal. Newer state-of-the-art techniques for automatic speech recognition replace statistical GMM approaches with deep neural networks (DNNs) that directly estimate posterior probabilities of the feature vectors.

2.2.3.1. Gaussian Mixture Models (GMM)

In this work speaker and gender models of mice are characterized using Gaussian Mixture Models (GMMs). GMMs are widely used in many speech processing applications as a statistical model for acoustic features extracted from speech. Advantages of GMMs include that they are computationally inexpensive and are highly generalizable. GMMs are a linear combination of more than one Gaussian distributions and can model any continuous density accurately with sufficient number of mixtures. A GMM is defined for a n-dimensional feature vector $\overrightarrow{X}$ as

$$p(\overrightarrow{X} \mid \lambda) = \sum_{i=1}^{M} w_i p_i(\overrightarrow{X}), \tag{2.12}$$

where w_i is the weight of i^{th} Gaussian Mixture. The weights must satisfy the condition that $\sum_{i=1}^{M} w_i = 1$. $p_i(\vec{X})$ is a multivariate Gaussian distribution with means μ_i and covariance $\sum_i$ of i^{th} Gaussian Mixture:

$$p_i(\vec{X}) = \frac{1}{((2\pi)^{(n/2)} |\sum_i|^{(1/2)})} e^{(-1/2(\vec{X}-\vec{\mu_i})^T \Sigma_i{}^{(-1)}(\vec{X}-\vec{\mu_i}))}. \quad (2.13)$$

where μ_i is an n-dimensional vector with $\mu_i = E(x)$ and Σ_i is a n by n covariance matrix with $\Sigma_i = E[(x-\mu_i)(x-\mu_i)']$, and $|\Sigma_i|$ is the determinant of covariance matrix. Collectively the parameters of the GMM are denoted as, $\lambda = \{w_i, \mu_i, \Sigma_i\}, 1 \le i \le M$. Although the model supports a full covariance matrix, using a diagonal covariance matrix has the advantage of computational efficiency since repeated matrix inversions are not required. Since cepstral coefficients are already largely uncorrelated by nature, using a diagonal covariance matrix is a reasonable approach. In a study by Reynolds, Quatieri and Dunn, it has been suggested that diagonal covariances outperform a full covariance matrix (Reynolds, Quatieri et al. 2000) for speaker identification.

The likelihood of the GMM parameters are trained using the Expectation Maximization (EM) algorithm. EM is an iterative maximum likelihood algorithm used when there are unknown hidden parameters, in this application the knowledge of which mixture is associated with a particular training feature vector. The expectation step of the algorithm uses the current parameter estimates to identify the mixture likelihoods for each feature vector, and then the maximization step uses a mixture-likelihood weighted combination of features to re-estimate the model parameters which maximizes the likelihood of GMM. With an initial model λ, a new model $\overline{\lambda}$ is estimated such that

$p(X \mid \overline{\lambda}) \geq p(X \mid \lambda)$. The new model then becomes the base model and the process repeats until some convergence is reached. The likelihood of GMM model λ for the training vector $X = \{\overrightarrow{x_1}, \overrightarrow{x_2}, ..., \overrightarrow{x_T}\}$ is calculated by,

$$p(X \mid \lambda) = \prod_{t=1}^{T} p(\overrightarrow{x_t} \mid \lambda). \tag{2.14}$$

2.2.3.2. Hidden Markov Model (HMM)

The state-of-the-art method used for temporal sequencing in speech processing is the HMM. HMMs sequentially model and combine the statistical models of the acoustic parameters of a speech signal, with each state of the HMM modeling the acoustic characteristics of a particular time-region of the speech data. HMMs align the frames of acoustic data against the HMM states and calculate the overall likelihood of the acoustic data generated by the model. Traditionally, GMMs have been used to represent the probability distribution of each state, but most modern techniques have now replaced GMMs with DNNs that directly model state probabilities. The main advantage of HMM is its ability to model the changes in temporal pattern with spectral patterns of vocalizations.

In a Markov process, the probability of a random variable at a given time depends only on the probability at the preceding time (Rabiner 1989).

$$p[q_{t+1} = S_j \mid q_t = S_i, q_{t-1} = S_k .....] = p[q_{t+1} = S_j \mid q_t = S_i] \tag{2.15}$$

where S_i, i=1….N , is the N distinct states of the system at any given time and q_t is the state at time t. A Markov process needs less memory and is called an observable Markov model, since each state corresponds to an observable event.

A Hidden Markov Model, an extension of the Markov process, is a double embedded stochastic process, with an underlying stochastic process that cannot be observed directly. A Hidden Markov Model, ρ can be defined by,

- The number of states in the model N, where individual states are denoted by $S = \{S_1, S_2, ...S_N\}$ and the state at time t as q_t.
- The number of observation symbols M, per state $V = \{v_1, v_2.....v_m\}$.
- The state transition probability matrix, a_{ij}, transition from state *i* to state *j*.

$$a_{ij} = p(q_{t+1} = S_j \mid q_t = S_i), 1 \le i, j \le N. \tag{2.16}$$

- The output probability matrix, $B = b_j(k)$, where $b_j(k)$ is the probability of emitting symbol, v_k, at state *i*.

$$b_j(k) = p(v_k \mid q_t = S_j), 1 \le j \le N; 1 \le k \le M \tag{2.17}$$

- The initial state distribution $\pi_i = p(q_1 = S_j), 1 \le i \le N$.

HMM can be conveniently denoted as $\rho = (A, B, \pi)$. Since $a_{ij}, b_j(k)$ and π_i are probabilities they must satisfy,

$$\begin{aligned} & a_{ij} \ge 0, b_j(k) \ge 0, \pi_i \ge 0 \forall i, j, k \\ & \sum_{j=1}^{N} a_{ij} = 1, \\ & \sum_{k=1}^{M} b_j(k) = 1, \\ & \sum_{i=1}^{N} \pi_i = 1. \end{aligned} \tag{2.18}$$

Basic problems of HMM

1) The evaluation problem - Given the observation sequence $O = \{o_1, o_2 ..... o_T\}$ and model $\rho = (A, B, \pi)$, how to efficiently calculate $p(O \mid \rho)$, the probability of observation given model.
2) The decoding problem - Given the observation sequence $O = \{o_1, o_2 ..... o_T\}$ and model $\rho = (A, B, \pi)$, how to efficiently choose a state sequence, $q = q_1, q_2, ... q_t$, which best explains the observation sequence.
3) The learning problem – how to adjust model parameters, $\rho = (A, B, \pi)$, to maximize $p(O \mid \rho)$.

Solution to Evaluation problem – Forward Backward Procedure

In order to find $p(O \mid \rho)$, the sum probability of all possible state sequence needs to be calculated. Let $Q = q_1, q_2, ... q_T$ be such a state sequence, then,

$$p(O \mid \rho) = \sum_{allQ} p(O \mid Q, \rho) p(Q \mid \rho) \tag{2.19}$$

$$= \sum_{q_1, q_2 .... q_T} \pi_{q_1} b_{q_1}(O_1) a_{q_1 q_2} b_{q_2}(O_2) ..... a_{q_{T-1} q_{T-2}} b_{q_T}(O_T). \tag{2.20}$$

This method of calculation is computationally expensive, since it has a time complexity of $O(N^T)$, where N is the number of states and T is the number of observations. A more efficient way of solving the problem is through a dynamic programming algorithm called the forward backward algorithm.

Let a forward variable $\alpha_t(i)$ be defined as

$$\alpha_t(i) = p(O_1, O_2, ... O_t, q_t = S_i \mid \rho). \tag{2.21}$$

Then $\alpha_t(i)$ can be calculated recursively by,

1) Initialization

$$\alpha_1(i) = \pi_i b_i(O_1), 1 \le i \le N. \tag{2.22}$$

2) Induction

$$\alpha_{t+1}(j) = \left[\sum_{i=1}^{N} \alpha_t(i) a_{ij} \right] b_j(O_{t+1}), 1 \le t \le T-1; 1 \le j \le N. \tag{2.23}$$

3) Termination

$$p(O \mid \rho) = \sum_{i=1}^{N} \alpha_T(i) \tag{2.24}$$

Step 1 initializes the forward calculation, step 2 iteratively calculates all the forward probabilities as illustrated in Figure 4, and step 3 calculates $p(O \mid \rho)$ as the sum of forward variables. The time complexity for this method is $O(N^2 T)$.

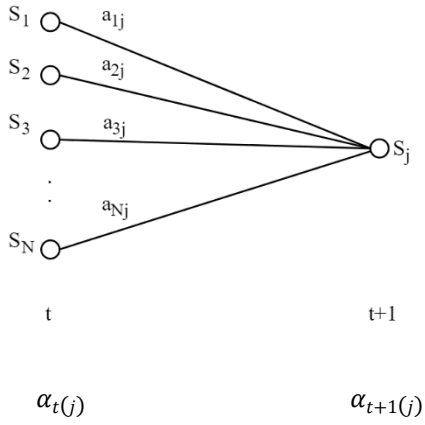

Figure 4: Computation of forward variable

The backward probability is defined similarly to the forward probability as

$$\beta_t(i) = p(O_{t+1}, O_{t+2}, ...O_T \mid q_t = S_i, \rho) \tag{2.25}$$

$\beta_t(i)$ can be calculated iteratively in a time-reverse fashion as

1) Initialization

$$\beta_T = 1, 1 \le i \le N. \tag{2.26}$$

2) Induction

$$\beta_t(i) = \sum_{j=1}^{N} a_{ij} b_j(O_{t+1}) \beta_{t+1}(j), t = T-1, T-2, ...1, 1 \le i \le N \tag{2.27}$$

3) Termination

$$p(O \mid \rho) = \sum_{i=1}^{N} \beta_i(1) \tag{2.28}$$

Solution to decoding problem – The Viterbi Algorithm

The forward algorithm calculates the overall probability that a given HMM generates a particular observation sequence. In speech recognition it is also important to find an optimal state path that generates an observation, which the forward algorithm does not accomplish. An alternative dynamic programming method called the Viterbi Algorithm is used to find the single highest probability state sequence for a specific observation sequence.

The viterbi algorithm is a recursive algorithm which finds the best state sequence as follows:

1) Initialization

$$\delta_1(i) = \pi_i b_i(O_1), 1 \le i \le N \tag{2.29}$$

$$\varphi_1(i) = 0. \tag{2.30}$$

2) Recursion

$$\delta_t(j) = \max_{1\le i\le N}\left[\delta_{t-1}a_{ij}\right]b_j(O_t), 2\le t\le T; 1\le j\le N. \tag{2.31}$$

$$\varphi_t(j) = \arg\max_{1\le i\le N}\delta_{t-1}(i)a_{ij}, 2\le t\le T; 1\le j\le N. \tag{2.32}$$

3) Termination

$$P^* = \max_{1\le i\le N}\delta_t(i), \tag{2.33}$$

$$q_T^* = \max_{1\le i\le N}\delta_T(i). \tag{2.34}$$

4) Back Tracking

$$q_t^* = \varphi_{t+1}(q_{t+1}^*), t = T-1, T-2, ..., 1 \tag{2.35}$$

$Q^* = q_1^*, q_2^*....., q_T^*$ is the desired optimal state sequence.

Solution to learning problem- Baum Welch Algorithm

The solution to the learning problem requires a method to adjust the model parameters (A, B, π) to maximize the probability of observation given the model. The learning problem is solved using the iterative procedure of the Baum Welch algorithm, which similar to the GMM estimation process is also an Expectation Maximization method.

Let $\gamma_t(i)$ be the probability of being in state S_i at time t

$$\gamma_t(i) = p(q_t = S_i \mid O, \rho). \tag{2.36}$$

Equation (2.36) can be written in terms of forward backward variables as

$$\gamma_t(i) = \frac{\alpha_t(i)\beta_t(i)}{p(O\mid\rho)} = \frac{\alpha_t(i)\beta_t(i)}{\sum_{i=1}^{N}\alpha_t(i)\beta_t(i)}. \tag{2.37}$$

Let $\xi_t(i, j)$ be the probability of being in state S_i at time t and S_j at time $t+1$

$$\xi_t(i, j) = p(q_t = S_i, q_{t+1} = S_j \mid O, \rho) \tag{2.38}$$

$$= \frac{\alpha_t(i) a_{ij} b_j(O_{t=1}) \beta_{t+1}(j)}{\sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_t(i) a_{ij} b_j(O_{t=1}) \beta_{t+1}(j)} \tag{2.39}$$

The relationship between $\gamma_t(i)$ and $\xi_t(i, j)$ can be shown as follows,

$$\gamma_t(i) = \sum_{j=1}^{N} \xi_t(i, j). \tag{2.40}$$

The expected number of times that state S_i is visited or the expected number of transitions made from state S_i can be calculated as follows,

$$\sum_{t=1}^{T-1} \gamma_t(i) = \text{expected number of transitions from } S_{i.} \tag{2.41}$$

$$\sum_{t=1}^{T-1} \xi_t(i, j) = \text{expected number of transitions from } S_i \text{ to } S_j. \tag{2.42}$$

Using these formulas, we can define the HMM parameters (A, B, π),

$$\overline{\pi_i} = \text{expected frequency in state } S_i \text{ at time (t=1)} = \gamma_1(i). \tag{2.43}$$

$$\overline{b_j}(k) = \frac{\text{expected number of times in state j and observing symbol } v_k}{\text{expected number of times in state } j}$$

$$= \frac{\sum_{\substack{t=1 \\ O_t = v_k}}^{T} \gamma_t(j)}{\sum_{t=1}^{T} \gamma_t(j)}. \tag{2.44}$$

After the re-estimation process the new model $\overline{\rho} = (\overline{A}, \overline{B}, \overline{\pi})$ will have a higher likelihood than the previous parameters $p(O \mid \overline{\rho}) > p(O \mid \rho)$. The re-estimation is done iteratively by replacing ρ by $\overline{\rho}$ until it converges. The expectation step is the calculation

of $\overline{\rho}$ from ρ and in the maximization step, maximization over $\overline{\rho}$. A more detailed description of HMMs can be found in a tutorial by Rabiner (Rabiner 1989).

2.2.4. Speaker identification

Speaker identification is a subset of speaker recognition, which includes the tasks of both speaker verification and speaker identification. Speaker identification is the problem of determining which specific speaker is speaking from a set of known speakers. In contrast, speaker verification is a binary classification problem that takes a claimed/proposed individual identity and answers the question of whether the identify claim is true or false. Within speaker identification, there are two types of tasks, closed set and open set. In closed set speaker identification, the best match for the speaker is selected from a known group of speakers, there is no rejection strategy. As the number of speakers in the known group increases the difficulty in identifying the speaker increases. In the open set speaker identification, there is an additional identification category of "none of the above", such that if a claimed speaker's verification fails, there won't be any identification result. Speaker recognition is further divided into text-dependent and text independent categories according to whether the input text is specifically prompted (Reynolds 1995). Figure 5 and Figure 6 shows the basic structure for a typical speaker recognition system.

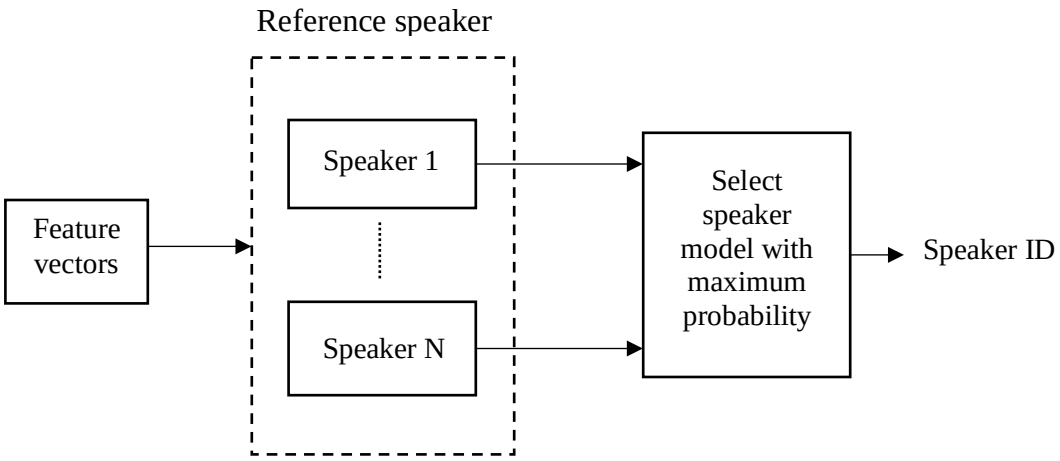

Figure 5: Speaker Identification System (Reynolds 1995)

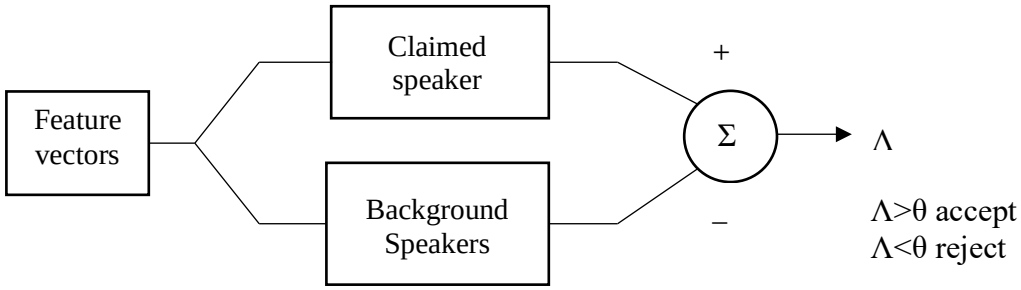

Figure 6: Speaker Verification System (Reynolds 1995)

Human speech contains many speaker specific characteristics, which are due to both physiological and learned differences. Vocal fold characteristics and vocal tract shape are the main physiological factors that contribute to speaker specific features in a person's voice. Air flow through vocal folds during speech production creates resonances in vocal tract that changes the spectral content of the speech wave as indicated by typical speech features such as MFCCs. Another distinguishing feature for speaker identification is the

fundamental frequency or pitch of the speech waveform. Learned aspects of speech like dialect and speaking rate also helps to distinguish between speakers.

Like speech recognition, speaker identification employs short term processing of speech by segmenting the voice activity extracted from recorded speech waveform typically into 20 – 30ms frames. Noise and silence removal is important in speaker identification tasks, since these can represent a false model related to aspects other than individual identity. Feature vectors are extracted from each frame using feature extraction methods, generally Mel frequency cepstral coefficients due to its ability to match frequency sensitivity of human ear. The next step in speaker identification is to model a statistical representation of speaker from the feature vectors, typically using methods such as GMMs and HMMs.

In Gaussian mixture speaker modeling, the techniques explained in section 2.2.3.1. for acoustic modeling are used. This model is then applied to speaker identification and verification. Speaker identification typically uses a Maximum Likelihood classifier. For a reference set with S speaker models $\hat{\lambda} = \{\lambda_1, \lambda_2 ..... \lambda_S\}$, a speaker $\hat{S}$ with maximum likelihood for the test feature vector $X = \{x_1, x_2 ..... x_T\}$ is

$$\hat{S} = \arg\max_{1<s<S} \frac{p(X \mid \lambda_s)}{p(X)} pr(\lambda_s). \tag{2.45}$$

If we assume equal prior probabilities $pr(x_t \mid \lambda_s)$ and a constant $p(X)$ for all speakers and calculate log probabilities, from equation (2.45) ,

$$\hat{S} = \arg\max_{1\le s\le S} \sum_{t=1}^{T} \log p(x_t \mid \lambda_s), \tag{2.46}$$

where T is number of feature vectors and $p(x_t \mid \lambda_s)$ is calculated from Equation (2.12).

In HMM speaker modeling, a HMM model $\rho = (A, B, \pi)$ for each speaker is trained using the methods detailed in section 2.2.3.2. The feature vector for the test speech is given to the system as observation sequence $O = \{o_1, o_2, ..., o_T\}$ and the likelihood of the observation sequence for all speaker models is calculated. The speaker model with maximum likelihood is selected as the predicted speaker.

$$\hat{S} = \arg\max_{1 \le s \le S} p(O \mid \rho_s) \tag{2.47}$$

2.2.5. Gender identification

Gender identification from speech determines whether the speech is uttered by a male or female speaker. The differences in male and female voices arise from physiological, acoustical and perceptual factors (Wu and Childers 1991). Physiological differences are due to the differences in vocal tract length, vocal fold length and vocal fold thickness (Titze 1989). These differences in physiology contribute to acoustic differences in male and female speech. The formant frequencies of females are related to formant frequencies of males by a scaling factor that is inversely proportional to vocal tract length. Formant frequencies are typically 20% higher for female than that of males (Wu and Childers 1991). The fundamental frequency (pitch) is higher for females than males, which is also a distinguishing factor when determining gender from speech. It has been found that fundamental frequency is scaled according to the vocal fold length (Titze 1989). In a study by Singh and Murry, perceptual factors used to characterize female speakers were nasality, pitch and effort, whereas effort, hoarseness and pitch were used for male speakers (Murry and Singh 1980).

Gender models can be implemented using GMMs/HMMs and MFCCs, similarly to the speaker modeling methods discussed in section 2.2.3 speaker identification. Feature vectors are extracted from short frames of speech after removing the noise and silence regions using MFCCs. MFCCs are then used to train the GMMs/HMMs for male and female class. Test speech is then classified into male or female category using the maximum likelihood values obtained by comparing the test speech with both male and female models.

2.3. Bioacoustics

Bioacoustics investigates how sound is produced and received by animals. Animals rely upon their vocalizations for a wide variety of purposes, including communicating with the members of same species and monitor their surroundings. In recent years, the speech processing and machine learning techniques from human speech processing techniques have begun to be used to study animal communication for detection and classification, with applications to censusing, acoustic ecology and understanding the effect of noise on animal communication.

In addition to signaling information, animal vocalizations convey information about species, gender, group and individual identity. The analysis and classification of animal sounds can be a powerful tool for monitoring the diversity of animal communities. This can be a noninvasive and economical way to study vocally active species who live in habitats that are difficult to reach or are sensitive to human intervention and may help biologists for conservation of endangered species.

In this section tasks associated with Bioacoustics, in particular, speaker identification and gender identification and techniques incorporated from speech processing are discussed.

2.3.1. Bioacoustics tasks

The main tasks associated with bioacoustics are classification, detection and localization. These tasks have a broad range of applications across many species. The human speech processing community has addressed similar tasks for many years, although with a different set of constraints and challenges, and the bioacoustic community has the potential to use such human speech processing techniques to address the tasks associated with bioacoustics. There are many challenges associated with this, since data collection is much more difficult and since human speech is better understood in terms of the relationship between acoustics and meaning. However, by using human speech processing tools such as HMMs, GMMs and DNNs, the performance on bioacoustics tasks can be improved.

2.3.1.1. Classification

Classification is the task of classifying vocalizations into one or a set of predefined categories, which may include species, call type, individual identity, gender or behavioral context. Classification is a supervised task that involves training data having annotated labels which is used to train models for classification. The results of classification are usually represented as a confusion matrix, which provides a visual representation of correct classification against misclassified data. Classification methods are often used as an

underlying layer for detection in classifying a specific type of category within a larger data set (Clemins 2017).

The classification task includes various applications such as call classification (Clemins, Johnson et al. 2005, Garland, Castellote et al. 2015), species classification (Roch, Soldevilla et al. 2007, Trifa, Kirschel et al. 2008), individual classification (Brown, Smaragdis et al. 2010, Ji, Johnson et al. 2013, Dvorakova, Ptacek et al. 2017) and gender identification (Volodin, Volodina et al. 2015), to classify vocalizations into predefined categories using a model created from labeled vocalizations from each category.

As discussed in section 1.1, there are several template matching methods like Dynamic Time Warping (DTW), Spectrogram Cross Correlation (SCC) and Matched filtering as well as statistical models like Gaussian Mixture Models (GMM) and Hidden Markov Models (HMM) for classification. In template matching chosen vocalizations are considered as templates and the test vocalizations are matched against the templates to measure similarity. DTW (Kogan and Margoliash 1998, Brown and Miller 2007) is a template matching method that was widely used method for human speech recognition before HMMs and GMMs because of its ability to compare non-linear waveforms. DTW finds optimal match between two sequences of speech feature vectors by finding the path that reduces the total distance between them, using dynamic programming algorithm. In standard DTW algorithm the distance between the test signal and template signal is calculated initially by arranging them on the vertical and horizontal axis respectively using the equation,

$$\text{Dist}[i, j] = |\, \text{test}[i] - \text{template}[j] \,|, \tag{2.48}$$

where i and j are the indices along the vertical and horizontal axis respectively. Then the dissimilarity is calculated as cost using a cost matrix C given by

$$C[i, j] = \text{Dist}[i, j] + \min \begin{Bmatrix} C[i-1, j-1] \\ C[i-1, j] \\ C[j-1, i] \end{Bmatrix} \tag{2.49}$$

DTW performs better in simple less noisy data sets for isolated vocalization classification however when noisy or complex vocalizations are present DTW requires careful selection of data to achieve a higher performance.

Another common method for automatic animal vocalization classification is Spectrogram Cross Correlation (SCC) (Khanna, Gaunt et al. 1997, Mellinger and Clark 2000) which cross correlates the template vocalization with the test vocalization. The resulting series of recognition values represents the similarity between the target and test vocalizations. Given the target set and the spectrogram for the test signal, the similarity is calculated as

$$\alpha(t, f) = \sum_t \sum_f \text{template}(t, f)\text{test}(t, f) \tag{2.50}$$

The disadvantages of SCC include performance dependency on size of FFT, window length and type. Since SCC is a quantitative method a good amount of correlation is required for successful classification and variations in patterns of vocalization can affect SCC performance. Variations in the ambient background and noisy environment can affect correlation between two spectrograms.

Another template matching method is matched filtering (Mellinger and Clark 1997), which is a template matching technique works in the temporal domain unlike SCC, which works in the frequency domain. Matched filtering is used to identify known signal

from a signal corrupted with white noise, accomplished by cross correlation between template and test vocalization in time domain.

GMMs and HMMs, which are the primary modeling and classification approaches used in this work, are explained in section 2.3.2.2 and section 2.3.2.3 respectively.

As discussed in Section 2.2.4, speaker identification is the task of determining the speaker who is vocalizing in a particular speech segment. Speaker identification plays an important role in conservation of animals, as it can be used as an effective tool for population monitoring, helping to understand behavioral traits and generating data for conservation tools (Terry, Peake et al. 2005). In situations where animals are sensitive to human interference, identifying individual from their vocalizations can be used as an alternative, non-invasive marking method.

Several studies have been conducted in the bioacoustics field regarding the presence of individual identity characteristics in animal vocalizations. Studies involving Rhesus monkeys have shown individual and kin recognition in contact calls of female rhesus macaques (Rendall, Rodman et al. 1996). This study used playback calls from close related individuals to show that the subjects can identify individuals from their contact calls. Another study of Bottleneck dolphins by Janik (Janik, Sayigh et al. 2006) found that individual identity is present in dolphin signature whistles. In this study playback experiments using synthetic signature whistles of close relatives induced a favorable reaction in dolphins which indicated the individual identity in signature calls.

In the two species under study in this work, mice and meerkats, there have thus far been few studies related to individuality. For mice, it has been pointed out by Holy and Guo (Holy and Guo 2005) that the songs of male mice have individual characteristics. A

study by Penn (Musolf, Hoffmann et al. 2010, Hoffmann, Musolf et al. 2012) revealed that the ultrasonic vocalizations of mice are embedded with individuality and kinship. In meerkats, playback experiments showed meerkats being vigilant for a long duration when presented with close calls of subordinate meerkat, near the test subject and within seconds, from a physically impossible geographical position away from the test subject (Townsend, Allen et al. 2012).

Some studies have implemented human speech processing techniques for individual identification from animal vocalization. For example, a study by Brown et al. in killer whales used HMMs and GMMs for individual identification (Brown, Smaragdis et al. 2010). Studies using HMM-GMMs include tigers (Ji, Johnson et al. 2013), Norwegian ortolan bunting (Trawicki, Johnson et al. 2005, Tao, Johnson et al. 2008, Adi, Johnson et al. 2010), African elephants (Clemins, Johnson et al. 2005) and Asian elephants, chicken (Ren, Johnson et al. 2009).

Gender identification determines the gender of the vocalizing animal from its vocalizations. Studies in some avian species with no visible sexual dimorphism has shown gender differences in vocalization, which is an effective non-invasive method for gender determination (Carlson and Trost 1992, Volodin, Kaiser et al. 2009, Volodin, Volodina et al. 2015). In study on screams of chimpanzees and bonbons using discriminant function analysis yielded 80% accuracy for gender identification in chimpanzees and 70% in bonbons (Mitani and Gros-Louis 1995). 22KHz alarm cries by rats in potential threat situation than actual threats showed sex differences (Litvin, Blanchard et al. 2007). Another study using vowel like grunt vocalizations in baboons showed gender differences in adults (Rendall, Owren et al. 2004).

Gender is often used in human speech recognition to generate gender-specific phonetic models, which implicitly involves gender classification as part of the automatic speech recognition process.

2.3.1.2. Detection

Detection involves identifying the presence of a specific type or subset of vocalization patterns from a recording of acoustic data. Detection includes both a binary classification problem, the presence/absence of a target vocalization, and an estimation problem, the determination of the start and end points of each vocalization. The task is significantly compounded by the presence of environmental noise and non-target calls both from the same species and other species, as well as the possibility of overlap between multiple target vocalizations. Often detection incorporates classification methods - for example, detection using HMMs can simultaneously detect a vocalization and classify it into a set of trained categories. The most common methods for detection include spectrogram cross-correlation and match filtering, each of which involves using a sliding window approach and an evaluation function to generate a detection output, which is then compared against an established threshold. Assessment of detection results is measured using two methods, detection accuracy and timing accuracy. Detection accuracy includes both miss rates and false alarm rates, and timing accuracy involves the correct detection of start and end points.

One particular application of detection is simple signal detection or “Voice Activity Detection” (VAD), which identifies the start and end times of vocal activity in the long recoding data of audio. This does not involve classification, but often involve separating overlapped vocalizations. For human speech this is a well-studied area, and there are a

number of approaches used. One common approach includes a hypothesis testing approach that sets up statistical models for silence and non-silence using a set of features such as energy and spectral information.

2.3.1.3. Clustering

Clustering is an unsupervised method in which data with no category labels are clustered into subgroups according to some measurement criteria. There are generally two types of approaches, divisive or agglomerative. Divisive clustering is a top-down approach where all data is initially placed into a single cluster and then the criteria is used to determine how to divide the data into groups, whereas agglomerative clustering is a bottom-up approach where each exemplar is initially identified as a cluster of one point, and then the criteria is used to determine how to combine groups together. The evaluation of clustering algorithms is an extremely difficult task due to the lack of ground truth. The consistency of clustering algorithms can be measured by evaluating the results across multiple runs and measuring the stability of the results obtained. However, consistency of clustering algorithms can change due to environmental noise and variations in gender or social group. Environmental noise may create clusters according to the noise rather than the vocalizations. Likewise, gender and social variability in larger groups can create subgroup clusters. Due to this, clustering results often cannot provide concrete conclusions and involves post-hoc analysis methods.

Unsupervised clustering is used for vocal repertoire analysis of single species into call categories and to determine number of those groups. Unsupervised clustering can also be implemented in individual identification, where vocalizations of a single species are grouped according to the individual variability. To avoid the variability in call type affect

the accuracy of individual variability, methods may use only data from a single call type when clustering for individual identification (Adi, Johnson et al. 2010).

2.3.2. Application of speech processing techniques to bioacoustics

Historically, the bioacoustic community has generally analyzed and classified animal vocalizations by visual inspection of spectrograms. This analysis method is time consuming and sometimes less effective than automatic methods, since it relies on only visibly-obvious features like frequency variation and duration from vocalizations. In recent years the speech processing community started incorporating the advanced techniques developed for human speech processing into bioacoustics. For example, feature extraction techniques like MFCCs and PLPs, statistical modeling techniques like GMMs and HMMs, and most recently HMM based DNNs, have been applied to bioacoustics tasks. This enabled the bioacoustics community to significantly improve accuracy on difficult tasks like species, individual and gender identification or vocalization classification of animals. This next section gives an overview of various speech processing methods used for bioacoustic signal analysis, with an emphasis on techniques used in this work.

2.3.2.1. Greenwood Function Cepstral Coefficients (GFCC)

As discussed in Section 2.2.2, MFCCs are most commonly used feature representation method in human speech processing techniques. MFCCs warp the perceived frequency to Mel-scale cochlear frequency map. Since the MFCCs are suitable for various speech processing tasks like speech recognition and speaker recognition, GFCCs (Clemins, Trawicki et al. 2006) were introduced as a generalized form of MFCCs to improve the bioacoustic signal processing of any given species. GFCC features use a generalized form

of the Mel-frequency scale in humans to create a cepstral coefficient feature representation that are theoretically well-founded across nearly all terrestrial mammals and give good vocalization representation across nearly all species. They can be implemented using only very basic knowledge of the minimum and maximum frequency range for a species. The advantage of GFCCs is that they use the information about the perceived frequency of the species under study.

GFCC's are derived from the Greenwood function (Greenwood 1990)

$$f = A(10^{ax} - k) \tag{2.51}$$

where a, A and k are species specific constants and x is the cochlea position. This equation is used to convert perceived frequency to measured frequency and vice versa through following equations

$$F_p(f) = (\frac{1}{a})\log_{10}(\frac{f}{A} + k) \tag{2.52}$$

$$F_p^{-1}(f_p) = A(10^{af_p} - k), \tag{2.53}$$

where f is the real frequency and f_p is the perceived frequency. The constant k can be approximated to 0.88 for a wide range of terrestrial mammals as shown by LePage(LePage 2003). The constants a and A can be found using the equations,

$$A = \frac{f_{\min}}{1-k}, \tag{2.54}$$

$$a = \log_{10}(\frac{f_{\max}}{A} + k), \tag{2.55}$$

where $f_{\min}$ and $f_{\max}$ are frequency range of the species.

Warping is done in the same way as that of MFCCs described in section 2.2.2, in that the vocalizations are framed using a sliding window with length appropriate for the

vocalizations. If the frequency of vocalization is higher and has a rapidly varying temporal pattern, small window size such as 2 – 5ms is used. Vocalizations with lower frequencies and slowly varying temporal patterns uses larger window size (Clemins 2017). Once the vocalizations are framed, the magnitude spectrum is calculated using the short-term Fourier transform and warped to Greenwood scale using filter bank analysis. Log magnitude spectrum is calculated by taking the logarithm of sum of coefficients after multiplying each filter with the magnitude spectrum of the frame. The next step is to apply a discrete cosine transform for a compact representation of general shape of speech spectrum in cepstral domain. The resulting features are called Greenwood Function Cepstral Coefficients.

Another feature extraction model for bioacoustic signal processing is generalized Perceptual Linear Prediction coefficients (gPLP), a generalized form of Perpetual Linear Prediction (PLP) designed to include the frequency perception of the any animal species. In gPLP, the filter banks are mapped to an equal loudness normalization curve to suit the animals hearing frequencies. Then filter bank energies are compressed and processed with low order all pole filter to solve for coefficients and converted to cepstral domain using direct recursion. A detailed description of gPLPs can be found in the paper by Clemins and Johnson (Clemins and Johnson 2006).

2.3.2.2. Gaussian Mixture Model (GMM)

Gaussian mixture models are used in various tasks in bioacoustics studies as a statistical model for representation of animal vocalizations, just as described in Section 2.2.3.1 for human speech. In individual identification using GMMs, each individual is modeled as GMM speaker model with sufficient number of mixtures to model the data. GMMs can be used as a classifier to discriminate between the classes using a maximum

likelihood classification. GMMs offer a computationally efficient method to classify individuals when compared to the HMMs.

Individual identification in Killer whales using GMMs has shown a 75% overall accuracy for vocalizations from four whales using single call type (Brown, Smaragdis et al. 2010). This study compared both HMMs and GMMs for individual identification and both methods showed little difference in accuracy. A study in Mashona mole-rats using GMM -Universal Background Model (UBM) showed an 76.7% of identification accuracy from mating calls of individuals (Dvorakova, Ptacek et al. 2017). In a study using Norwegian ortolan bunting data Adi et al. uses clustering of GMMs to identify the individuals from their vocalizations (Adi, Johnson et al. 2010).

2.3.2.3. Hidden Markov Model (HMM)

As described in Section 2.2.3.2, HMMs are able to model both temporal characteristics and spectral complexity of vocalizations. The temporal characteristics of a vocalization are modeled through the time-evolution of states and the spectral characteristics through the state distributions, typically GMMs in most of the speech recognition systems.

Speaker and gender identification problems are similar to isolated word recognition of human speech recognition, since they use individually pre-segmented vocalizations. In isolated word recognition using HMMs, there is one HMM learned for each vocalization, with a number of states determined according to the temporal characteristics of the vocalization type. After deciding appropriate number of states HMMs are initialized and trained using the Baum-Welch algorithm as explained in section 2.2.3.2, which is an expectation maximization algorithm used to find the optimal parameters to represent each

model. In the next step Viterbi algorithm is used to find likelihood with each of HMMs and then classification is done using the maximum likelihood classifier (Clemins 2017).

A study in red deer stag using MFCCs and HMMs revealed individual identity present in the common roars and yielded accuracy of 93% in individual identification (Reby, Andre-Obrecht et al. 2006). Another study in African elephants using the rumble vocalization showed 82.5% accuracy (Clemins, Johnson et al. 2005). The study in 4 Killer whales using N2 type of vocalization resulted in 75% accuracy (Brown, Smaragdis et al. 2010). Individual identification of a protected species, Asian small clawed otters using chirp vocalizations, showed 91% accuracy between two individuals.

2.3.3. Species under study

In this study, mice (*Mus musculus*) and meerkats (*Suricata suricatta*) vocalizations are used to conduct speaker identification and gender identification experiments. In mice five types of calls are used for speaker identification and six types of calls were used for gender identification. For meerkats, only one type of call, called Close calls, were used since they are abundantly available. In this section a brief description about the species and their vocal repertoire is discussed.

2.3.3.1. Mice

Mice ultrasonic vocalizations has been studied for decades. In 1956 Zippelius and Schleidt found that mice pups emit distress calls ranging from 70 – 80 KHz from birth until the age at which their eyes open (Zippelius and Schleidt 1956). In 1972, Sales G D discovered that adult mice emit ultrasonic vocalizations in various social situations (Sales 1972). Since then there have been many studies of mice vocalization patterns.

Mice emit sonic (Whitney 1970) and ultrasonic vocalizations during social interactions, ranging from 3 to 110 KHz (Holy and Guo 2005, Heckman, McGuinness et al. 2016). Pups produce isolation calls to gain the attention of mothers (Ehret 1992), while adults mainly produce vocalizations during courtship (Holy and Guo 2005) or territorial dispute (Hammerschmidt, Radyushkin et al. 2012). Female mice also produce these USVs when alone, searching for pups, or in the presence of other females (Portfors 2007). In this work we focus on ultrasonic vocalizations by adult mice.

Mice vocalizations have a song like structure containing syllables arranged in a sequential pattern, ranging from 30 to 200ms in duration (Holy and Guo 2005). Vocalizations have been categorized USVs into 9 types of syllables based on length, bandwidth, and overall shape syllables using spectrogram analysis (Hanson and Hurley 2012).

These categories include:

- Short syllables were less than 10 ms in duration.
- Flat syllables had less than 5 KHz of modulation.
- Harmonic syllables contained at least one segment with at least one harmonic (most of these also had breaks in frequency).
- Jump syllables contained at least one break in frequency with no break in intensity (and no harmonics).
- Up syllables increased in frequency (sweep.5 KHz).
- Down syllables decreased in frequency (sweep.5 KHz).
- Arc syllables increased and then decreased in frequency, with the highest frequency reaching .5 KHz above the beginning and end frequencies.

- U syllables decreased and then increased in frequency, with the lowest frequency reaching .5 KHz below the beginning and end frequencies.
- Complex syllables contained two or more directional changes in frequency and .5 KHz modulation of frequency.

A few types of vocalizations are shown below:

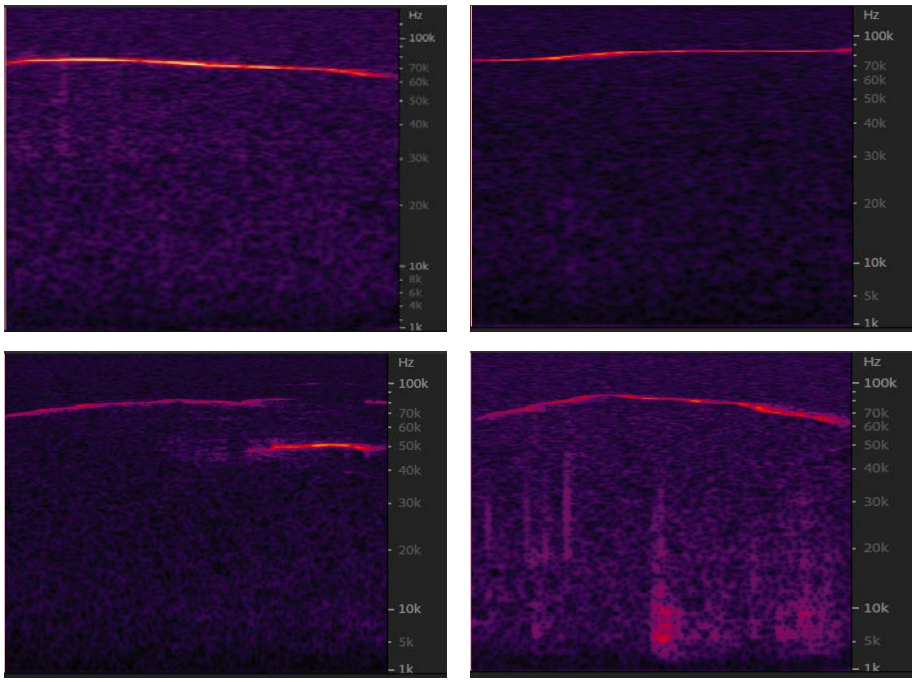

Figure 7: A few calls from Mice repertoire
7a: Down Sweep, 7b: Up Sweep
7c: Jump, 7d: Inverse Chevron

2.3.3.2. Meerkats

Meerkats are cooperatively breeding mongoose species that live in groups of 3 to 50 individuals. Each group has a dominant male and female who is responsible for most of the breeding and rest of the group members are helpers. The helpers are responsible for pup care, the male helpers stay in the group for one or two years and then they disperse into other groups. Female helpers either inherit the dominant position and stay in the group or get expelled from the group by the dominant female (Clutton-Brock, Russell et al. 2005).

Meerkats are highly social and territorial animals that spend most of their time foraging in groups for non-vertebrate preys. While foraging they keep their head to the ground which obstructs their vision and makes them prone to predator danger. Biologists believe that this foraging behavior is the reason for highly developed acoustic communication in meerkats (Reber, Townsend et al. 2013). Meerkats have around 30 different call types (Collier, Townsend et al. 2017), the most common call type is the Close calls which is believed to be used for group cohesion (Townsend, Hollén et al. 2010). It has been found that Close calls are individually distinctive, used for identifying the group members (Townsend, Allen et al. 2012) and plays an important role in territorial defense (Young and Monfort 2009). The Close calls ranges in frequency from 600-1000Hz and can travel up to 20 meters (Townsend, Hollén et al. 2010).

Meerkats produce aggression calls when other individuals approach them while eating or digging (Mausbach, Braga Goncalves et al. 2017). Lead calls are emitted when the individual changes the foraging patch which can facilitate the entire group to move. Move calls are emitted when an individual wants to change the foraging patch but require a minimum number of individual in favor for a change in forage patch to happen (Gall, Strandburg-Peshkin et al. 2017). Another call type is alarm call, emitted when spotting a predator, which can convey the information about type of predator and level of urgency (Townsend, Rasmussen et al. 2012).

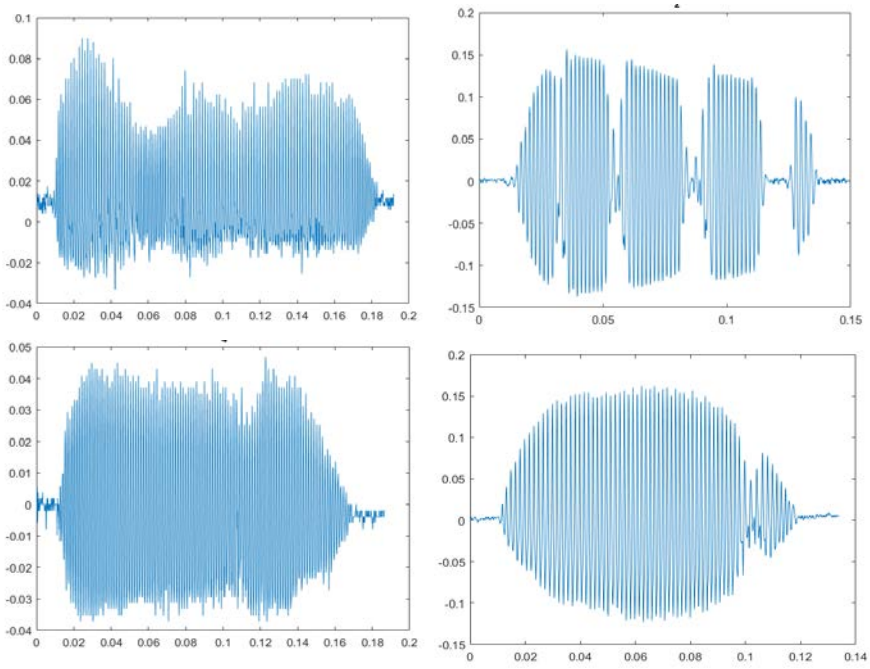

Figure 8: Meerkat call types
8a: Lead call, 8b: Close call
8c: Move call, 8d: Alarm call

2.4. Summary

The speech processing, feature extraction and statistical modeling techniques used for human speech have been discussed in this chapter, with emphasis on speaker and gender identification which is the main focus of this thesis. The applicability of human speech processing techniques along with modified feature extraction models for bioacoustic signals have been presented. In the next chapter the techniques explained in this chapter are implemented to identify the speaker and gender in mice from their vocalizations.

Chapter 3: Speaker and gender identification in Mice

3.1. Overview

This chapter demonstrates using a Gaussian Mixture Model (GMM) approach with Greenwood Function Cepstral Coefficient (GFCC) features for speaker identification and gender identification in mice using ultrasonic vocalizations. The vocalizations are segmented to extract voice activity and GFCC features are extracted from the segmented data, which in turn is used to train GMMs and to evaluate the effectiveness of the speaker and gender identification models.

3.2. Data Collection

Data for this study included vocalizations from 40 mice, 20 Male and 20 Female, collected at the University of Buffalo – SUNY, Psychology department (Burke, Screven et al. 2017). These study subjects are used for research in production and perception of ultrasonic vocalization in mice. Vocalizations were recorded using an ultrasonic microphone in individual sound-attenuated enclosures, in both isolated and visually paired conditions, including same-gender and opposite-gender pairing, for 1-hour periods. Vocalizations were recorded in segments of 5 minutes duration and used two channels, labeled Mouse A and Mouse B. The sampling rate of the vocalizations was 300KHz. Recordings were categorized as Female-Female, Male-Male, Male-Female, Non-Exposure or Pre-Exposure according to the nature of exposure. Calls were manually labeled into 9 different vocalization categories (Hanson and Hurley 2012): Chevron, Chirp, Complex, Down-sweep, Flat, Harmonic, Inverse Chevron, Jump and Upsweep. The recordings had timestamps with fields of start time, end time, channel number, name of the mouse, peak

frequency, duration and category of vocalizations, all of which were manually labeled by the researchers at SUNY by visual analysis of spectrograms after the recordings were made. The typical frequency range of the ultrasonic vocalizations is between 30 and 125 kHz, with an average frequency around 70KHz.

3.3. Experimental setup

The steps involved in classifying the vocalizations according to speaker or gender in this work include segmentation, dividing vocalizations into training and test sets, extracting features from vocalizations using GFCC, training the GMMs using the labeled training set and classifying the vocalizations in test set according to classification criteria. This work has been implemented using MATLAB 2017b. The process flow is shown in the block diagram below:

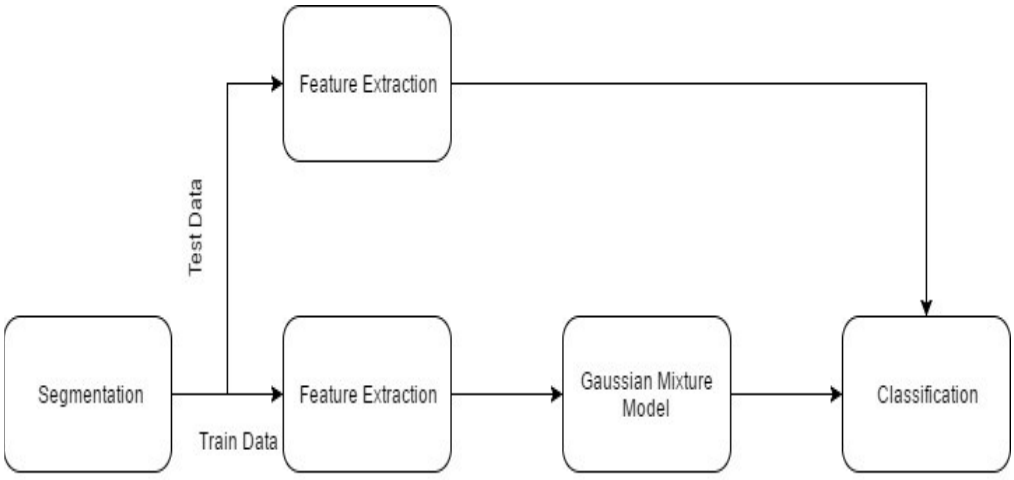

Figure 9: Work flow block diagram

Segmentation used the timestamp fields start time, end time, channel number and call category. A MATLAB script was used to extract these fields from the timestamps and each recorded audio file from all exposure categories was segmented for the voice activity

according to the extracted fields. The segmented files were named according to the name of the individual subject vocalizing. Segment duration varied from 2ms to 190ms. The segments were sorted according to speaker and gender for further processing. The call distribution for each category according to the call types are given in Table 1.

Table 1: Call Type occurrences over exposure categories

Category / Call Type	Female - Female	Male - Female	Male - Male	Non-Exposure/ Pre-Exposure	Total
Chevron	2	7	1	0	10
Chirp	33	35	114	13	195
Complex	16	55	23	7	101
Down Sweep	5	116	86	7	214
Flat	9	52	25	12	98
Harmonic	1	21	17	2	41
Inverse Chevron	16	34	33	36	119
Jump	62	21	64	20	167
Upsweep	210	36	31	125	402

The test and training data sets were created using an M fold cross validation approach. In M-fold cross validation, the data is divided randomly into M evenly sized partitions called folds. M experimental runs were conducted, with each consisting of (M-1) partitions being used for training and 1 partition used for testing. This ensured that each vocalization segment was used as a test segment one time.

With the low end of the frequency range at 30kHz, the minimum frame size to include at least 2-3 full cycles of any target frequency is about .1ms. In addition, since the minimum duration of vocalizations was approximately 2ms, a 2ms window size was the

maximum window size that could be used for framing. Based on this, in this work a window size of 1ms was used in order to capture the spectral complexity of mice ultrasonic vocalizations. A step size of half the window size was used.

GFCC features were extracted from the vocalization frames and GMMs for each individual were trained using the GFCC features from all vocalizations in the training set for the individual. Thirty-two mixtures were used for training the GMMs for speaker identification based on the amount of data available. Gender identification, which had a larger amount of data for each category, used 64 mixtures for the GMMs. Classification was done using a maximum-likelihood approach, by applying GMMs to determine the likelihood of the data for each category and choosing the highest likelihood category as the outcome.

Results are displayed in confusion matrix with the rows of the matrix representing the known class and columns representing the predicted class. Diagonal entries of the confusion matrix represent correct classification of each class. Accuracy of the classification is calculated as the ratio of sum of diagonal entries to sum of all entries. Another variable included is the chance accuracy, the measure of how well the classifier would have performed by chance, measured by taking the ratio of maximum number of entries for a known class to sum of total entries.

Each predicted class is also evaluated according to their Signal to Noise Ratio(SNR) calculated as

$$\frac{SNR_{s+n} - SNR_n}{SNR_n} \tag{3.1}$$

where SNR_{s+n} is the SNR of the vocalization and SNR_n is the SNR of the background noise, extracted from 100ms of the audio waveform before and after a vocalization.

3.3.1. Feature extraction

Greenwood Function Cepstral Coefficients (GFCC), as described in section 2.3.2.1, were used as features in this experiment. The GFCC features were calculated on individual windows, placed across each vocalization segment. Since the vocalizations have relatively high frequency content, in the ultrasonic range, only a small window size is needed to frame the vocalizations, on the order of 1ms. The vocalizations themselves are of short duration, so when the window size increases the number of frames generated from each vocalization segments decreases, which will in turn decrease the number of examples and therefore the number of mixtures that can be used to model the speaker using GMMs.

The Greenwood frequency warping constants were found based on the published frequency range of 30KHz – 125KHz (Holy and Guo 2005). To make sure of the use of all possible lower ultrasonic vocalizations, the minimum frequency was extended downward to 25KHz. 12 GFCC coefficients are extracted from each frame. The constants for calculating the GFCCs are calculated as follows,

$$k = 0.88, \tag{3.2}$$

$$A = \frac{f_{\min}}{1-k} = \frac{25000}{1-0.88} = 113636.7, \tag{3.3}$$

$$a = \log_{10}\left(\frac{f_{\max}}{A} + k\right) = \log_{10}\left(\frac{125000}{113636.7} + 0.88\right) = 0.297. \tag{3.4}$$

The SPEFT MATLAB tool box designed to extract speech features for bioacoustics such as GFCCs was used for feature extraction (Li 2007). Additional parameters included delta, delta – deltas and short-term energy from each frame to augment the feature vector using methods described in section 2.2.2.

3.3.2. Model training

The resulting features from GFCC extraction were used to train GMMs using the training data as explained in section 2.2.3.1 for each class. GMMs are trained using the built-in function *fitgmdst.m* in matlab. The algorithms involved in *fitgmdist.m* are Gaussian Mixture Model likelihood optimization with the k-means^{++} algorithm for initialization (McLachlan and Peel 2004). The Gaussian Mixture Model likelihood optimization algorithm uses an iterative Expectation Maximization (EM) algorithm to optimize the likelihoods of GMMs. The k-means^{++} algorithm is used to initialize the parameters of EM algorithm for GMMs. The k-means^{++} algorithm assumes a specific number of clusters to be calculated based on the number of mixtures, with an equal probability assigned for each cluster. The covariance is selected as diagonal and identical. In the next step a first initial center μ_1 taken uniformly from each data point in train set. Then other centers, $j = 1.....k$ at random from X for $m = 1.....n$, $p = 1.....j-1$ are calculated using,

$$\frac{d^2(x_m, \mu_p)}{\sum_{h, x_h \in M_p} d^2(x_h, \mu_p)}, \tag{3.5}$$

where $d^2(x_m, \mu_p)$ is the distance between observation m, μ_p and M_p is the set of all observations closest to centroid μ_p and x_m belongs to M_p. Then the Mahalanobis distance of each data point from the centers are calculated and is assigned to the closest center.

3.3.3. Identification

Once the GMMs were trained, the likelihoods of each test segment is calculated for all GMMs. The overall log likelihood of the segment was calculated by combining the likelihoods of all frames then choosing a single category as the predicted class based on the maximum log likelihood as discussed in section 2.2.4. The predicted class is then compared with the known class and results are displayed in a confusion matrix format.

3.4. Speaker identification

Speaker identification was implemented using most common call types: Up Sweep, Down Sweep, Chirp and Jump. Vocalizations were sorted for each speaker according to call types and the GFCC features were extracted from both train and test sets using the methods explained in section 3.3.1. The GMMs were trained for each speaker using a cross-validation process as described in section 3.3. The test set was evaluated by calculating GMM likelihoods and selecting a predicted class as described in section 3.3.3.

3.4.1. Subjects

Although there are 40 speakers in the dataset, due to the highly isolated conditions of the cages, not all individuals produced vocalizations. Because of this, there is a high variance in the call types and number of calls available for each speaker for speaker identification experiments. Based on the distribution of data, individuals with at least 9 calls for each call type were used for speaker identification. The call type distribution used for the speaker identification experiments is given in Table 2 below.

Table 2: Call distributions for each speaker for speaker identification

2a: Up Sweep call

Speaker	Up Sweep
34313	13
34334	60
Captain	9
Darby	124
Jackie	23
Luna	45
Minerva	14
OJ	45

2b: Down Sweep calls

Speaker	Down Sweep
Captain	25
OJ	13
Bob	17
Brutus	20
Cedric	10
Jamie	10
Ralph	17
Seifer	17
Squall	9

2c: Chirp calls

Speaker	Chirp
34312	17
34322	15
34332	48
34334	9
Darby	15
Jamie	17

2d: Jump calls

Speaker	Jump
34313	9
34312	22
34322	16
34332	22
Darby	52
Minerva	9
Shadowcat	12

3.4.2. Results

As described in Section 3.3, speaker identification was implemented using the Jump, Up Sweep, Down Sweep and Chirp call types. Vocalizations were framed using a window size of 1ms and GFCC features were extracted. Each individual was modeled using 32 mixtures and identification was implemented using maximum likelihood classifier. The overall results, compared to chance accuracy, for each individual call type as well as for all call types grouped together in a single experiment, are given in Table 3.

Table 3: Accuracy of speaker identification using different call types

Call Type	Accuracy	Chance
Jump (7 individuals)	**78.3**	36.36
Up Sweep (8 individuals)	58.9	37.2
Down Sweep (9 individuals)	33.3	39.7
Chirp (6 individuals)	50.4	36.6
All calls (27 individuals)	46.3	17.1

The highest accuracy was for the Jump calls, which had 7 callers with a sufficient number of this call-type, with an accuracy of 78.3% for individual identification. The second highest accuracy was for Up Sweeps with 8 individuals, an accuracy of 58.9%. Speaker identification using all call type had an accuracy of 46.3%, which was not as high as several of the individual call types but does indicate that it is possible to differentiate individuals to some extent without first segmenting into individual call types.

For the initial experiment, Jump calls were selected because most of the mice literature using ultrasonic vocalizations are based on the pitch jumps in the vocalizations analogous to the jump calls in this work (Holy and Guo 2005, Hoffmann, Musolf et al. 2012). The resulting confusion matrix for speaker identification for Jump calls is shown in Figure 10.

Overall accuracy for Jump calls was 78.3%. Higher accuracies are shown by individuals that have a larger number of data instances to train the speaker model, higher energy in the frequency bands, higher SNR or higher call duration. For example, Darby and Shadowcat had the highest SNRs, with means of 30.1 and 32.8 respectively. But

Shadowcat had vocalizations with higher duration than Darby, and Darby had more data to create the speaker model. Several individuals show specific error patterns. For example, the individuals Minerva and Darby seem to have somewhat similar vocalizations, with SNR and duration in the same range. This could be the possible reason that more than half of the vocalizations for Minerva are classified as Darby.

	C_F_Darby	C_F_Minerva	C_F_Shadowcat	C_M_34312	C_M_34313	C_M_34322	C_M_34332
F_Darby	51	1	0	0	0	0	0
F_Minerva	7	3	0	0	0	0	0
F_Shadowcat	0	0	12	0	0	0	0
M_34312	1	0	1	17	0	0	3
M_34313	0	0	0	3	2	1	3
M_34322	1	0	0	2	0	12	1
M_34332	1	0	0	6	0	0	15

Figure 10: Speaker identification for Jump calls with 7 individuals (Accuracy 78.3%)

Speaker identification results with Up Sweep calls are shown below in Figure 11. Up Sweep calls had the second highest accuracy of 58.9%. For Up Sweeps Minerva shows same pattern of error to Jump calls.

	C_34313	C_34334	C_Captain	C_Darby	C_Jackie	C_Minerva	C_OJ	C_luna
34313	12	0	0	0	0	0	0	1
34334	0	26	0	3	2	2	22	5
Captain	0	1	2	3	0	0	1	2
Darby	0	6	0	91	5	1	14	7
Jackie	0	2	0	5	8	0	0	8
Minerva	0	2	0	4	0	5	2	1
OJ	0	14	0	3	1	1	24	2
luna	0	4	0	5	4	2	2	28

Figure 11: Speaker identification for Up Sweep calls with 8 individuals (Accuracy 58.9%)

The Down Sweep calls had the least accuracy, 33.3% which was less than chance, as shown in Figure 12. Only a few speakers showed good accuracy, and the errors are widely distributed, suggesting that perhaps there is less individually identifying information in these calls.

	C_F_OJ	C_M_Bob	C_M_Brutus	C_M_Captain	C_M_Cedric	C_M_Jamie	C_M_Ralph	C_M_Seifer	C_M_Squall
F_OJ	4	1	3	3	1	0	0	0	1
M_Bob	2	4	0	4	1	0	0	6	0
M_Brutus	3	4	2	6	0	0	3	2	0
M_Captain	1	0	4	14	0	0	2	4	0
M_Cedric	0	0	1	0	7	0	1	1	0
M_Jamie	0	1	1	5	0	0	1	2	0
M_Ralph	1	0	1	2	0	0	8	5	0
M_Seifer	1	2	0	4	0	0	3	7	0
M_Squall	0	0	1	5	0	0	1	2	0

Figure 12: Speaker identification for Down Sweep calls with 9 individuals (Accuracy 33.3%)

	C_F_34334	C_F_Darby	C_M_34312	C_M_34322	C_M_34332	C_M_Jamie
F_34334	9	1	0	0	0	1
F_Darby	2	6	0	1	5	1
M_34312	0	0	2	3	10	2
M_34322	0	0	1	4	9	1
M_34332	0	0	2	4	41	1
M_Jamie	3	1	2	4	7	0

Figure 13: Speaker identification for Chirp calls with 6 individuals (Accuracy 50.4%)

Speaker identification using calls from all 27 individuals with at least 10 vocalizations had an accuracy of 46.3%, compared to a chance of 17.1%. The results for speaker identification using all calls are shown in confusion matrices in Figure 14. Results vary significantly across individuals but don't suggest a broad pattern.

	C_Squall	C_Tyrd	C_f34	C_fAph	C_fDar	C_fEle	C_fGam	C_fHer	C_fJac	C_fMin	C_fOJ	C_fScat	C_fluna	C_m12	C_m13	C_m15	C_m22	C_m32	C_mAres	C_mBob	C_mBrts	C_mCapt	C_mCedr	C_mJamie	C_mRal...	C_mSeifer	C_mWolv
Squall	11	0	0	0	0	1	1	0	0	0	0	0	1	0	0	0	2	0	0	4	0	5	0	0	0	0	0
Tyrd	0	5	0	0	4	0	0	0	2	0	1	1	1	0	0	0	0	0	0	0	0	2	1	0	0	0	0
f34	0	0	48	0	3	0	0	0	7	0	23	1	3	0	0	0	0	0	0	0	1	0	1	0	0	0	0
fAph	0	0	0	1	0	0	1	0	0	0	1	0	1	1	0	0	0	0	0	0	1	3	1	0	0	4	0
fDar	1	1	10	1	143	0	0	1	7	7	24	0	11	1	0	0	1	0	0	0	2	1	2	1	0	1	0
fEle	1	0	0	0	0	1	0	0	3	1	0	0	0	0	0	0	0	1	0	0	0	0	2	0	6	0	1
fGam	0	0	0	2	2	0	1	0	0	2	0	0	0	0	0	0	0	0	0	1	0	4	0	0	1	0	0
fHer	0	0	0	0	0	0	0	0	1	1	2	0	0	1	0	0	1	0	2	1	1	1	0	0	4	0	0
fJac	0	0	2	0	5	0	0	0	17	1	0	0	4	0	0	0	0	0	0	0	1	1	2	0	0	0	0
fMin	1	0	2	0	6	0	0	1	3	13	1	0	1	0	0	0	0	1	1	2	0	1	5	0	12	1	0
fOJ	0	0	20	1	5	0	0	0	3	3	27	0	1	0	2	0	0	0	1	2	2	8	2	0	1	0	0
fScat	0	0	1	0	1	0	0	0	0	0	0	16	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
fluna	0	1	5	0	9	0	0	0	7	3	4	1	28	0	0	0	0	0	0	1	1	2	0	0	0	2	0
m12	0	0	0	0	0	0	0	0	1	0	1	0	0	35	6	0	4	11	0	0	3	3	1	0	0	2	1
m13	0	0	0	0	0	0	0	0	0	0	0	0	0	11	42	0	4	9	0	0	0	0	0	0	0	1	1
m15	0	0	0	0	0	0	0	0	0	0	0	0	0	6	1	2	1	3	0	0	0	0	0	0	0	0	0
m22	2	0	0	0	1	0	0	0	0	2	1	0	0	3	3	0	26	5	0	2	4	0	2	0	4	0	0
m32	0	0	0	0	1	1	0	0	0	0	0	0	0	17	5	0	14	42	0	1	1	0	0	0	1	0	0
mAres	0	0	2	0	0	0	0	0	0	0	2	0	0	0	0	0	1	0	3	0	2	2	1	0	5	0	0
mBob	0	0	0	0	3	0	0	0	0	2	0	0	2	1	0	0	1	1	1	11	0	8	1	0	2	5	0
mBrts	0	0	2	0	2	1	0	4	0	0	2	0	0	2	0	0	5	0	3	3	12	6	1	2	4	1	0
mCapt	0	0	2	0	2	0	0	1	1	0	1	0	0	0	0	0	0	0	0	6	5	42	0	0	3	0	1
mCedr	0	0	0	0	1	0	0	0	1	1	1	0	0	0	0	0	1	1	1	0	0	0	7	0	0	0	0
mJamie	2	0	0	0	2	0	0	1	0	2	2	0	1	4	0	0	2	0	1	2	4	6	0	0	0	2	0
mRalph	0	0	0	0	0	2	0	0	0	1	1	0	1	0	0	0	1	0	1	3	4	3	1	1	26	3	0
mSeifer	0	0	0	0	3	0	0	0	0	0	1	0	0	1	0	0	2	0	1	4	0	9	0	0	3	23	0
mWolv	0	0	0	0	0	0	0	0	0	0	1	0	0	0	1	0	0	0	3	3	0	2	0	0	1	1	0

Figure 14: Speaker identification for all calls grouped together with 27 individuals (Accuracy 46.3%)

Short term energy, delta and delta – deltas were used to augment feature vector for speaker identification with Jump calls. A window size of 1ms and step size of 0.5ms were used to frame the vocalizations and GFCCs, delta, delta – delta and short-term energy were extracted from each vocalization. Speaker identification using GFCCs and short-term energy had an accuracy of 75.5%. Speaker identification using GFCCs, short term energy and delta along with 64 mixtures to incorporate the increased vector space had an accuracy of 76.9%. Speaker identification using GFCCs, short-term energy, delta and delta – deltas had an accuracy of 76.2%. The confusion matrices for the speaker identification using short term energy, delta and delta- delta are shown below.

	C_F_Darby	C_F_Minerva	C_F_Shadowcat	C_M_34312	C_M_34313	C_M_34322	C_M_34332
F_Darby	51	1	0	0	0	0	0
F_Minerva	7	3	0	0	0	0	0
F_Shadowcat	0	0	12	0	0	0	0
M_34312	1	0	1	15	1	0	4
M_34313	0	1	0	3	2	1	2
M_34322	1	0	0	2	0	12	1
M_34332	0	1	0	7	0	1	13

Figure 15: Speaker identification for Jump calls using GFCC and Short-term energy (Accuracy 75.5%)

	C_F_Darby	C_F_Minerva	C_F_Shadowcat	C_M_34312	C_M_34313	C_M_34322	C_M_34332
F_Darby	51	1	0	0	0	0	0
F_Minerva	6	4	0	0	0	0	0
F_Shadowcat	0	0	12	0	0	0	0
M_34312	1	0	1	16	0	1	3
M_34313	0	0	0	3	2	1	3
M_34322	1	0	0	1	0	13	1
M_34332	1	0	0	9	0	0	12

Figure 16: Speaker Identification for Jump calls using GFCC, Short-term energy and delta (Accuracy 76.9%)

	C_F_Darby	C_F_Minerva	C_F_Shadowcat	C_M_34312	C_M_34313	C_M_34322	C_M_34332
F_Darby	50	2	0	0	0	0	0
F_Minerva	5	5	0	0	0	0	0
F_Shadowcat	0	0	12	0	0	0	0
M_34312	1	0	1	14	0	0	6
M_34313	0	0	0	3	2	1	3
M_34322	1	0	0	0	1	14	0
M_34332	0	1	0	8	0	1	12

Figure 17: Speaker Identification for Jump calls using GFCC, Short term energy, Delta and Delta - Delta (Accuracy 76.2%)

3.5. Gender identification

Gender identification was implemented using call types Up Sweep, Down Sweep, Chirp, Jump and Inverse Chevron. Each call type was sorted into Male and Female class and GFCC features were extracted from each frame of the test and train sets using the methods described in section 3.3.1. Once the models were trained test samples were classified using the methods explained in section 3.3.3.

3.5.1. Subjects

Individuals with the specified call types were selected for gender identification experiments. Although data was separated into specific call types to test the gender data present in the vocalizations, gender identification was also tried using all vocalizations from all individuals to check dependency of gender data on vocalization category. Call distribution for gender identification using single call type for both male and female class is given in Table 4.

Table 4: Call distribution for gender identification

Gender / Call type	Female	Male
Jump	87	78
Chirp	39	47
Up Sweep	115	61
Down sweep	51	125
Inverse chevron	47	62
All calls	605	622

3.5.2. Results

As described in Section 3.3, gender classification was implemented using 10-fold cross validation to sort test and train sets. Vocalizations were framed using a window size of 1ms and GFCC features were extracted. Each gender was modeled using 64 mixtures and identification was implemented using maximum likelihood classifier. The overall results, as compared to chance accuracy, for each individual call types as well as for all call types grouped together in a single experiment, is given in Table 5.

Table 5: Accuracy and chance of Gender classification using different call types

Call Type	Accuracy	Chance
Jump	**93.2**	52.7
Chirp	87.2	54.6
Up Sweep	90.9	65.3
Down Sweep	62.9	56.1
Inverse Chevron	84.4	56.9
All calls	88.9	54.5

Jump calls yielded the highest accuracy of 93.2%. The confusion matrix for Jump calls is shown in Figure 18.

	C_female	C_male
female	84	2
male	9	68

Figure 18: Gender identification for Jump calls (Accuracy 93.2%)

Identification using Jump calls had higher accuracy for female vocalizations (97.7%) than that for male vocalizations (87.2%), probably because of the higher spectral energy in female vocalizations. Females had a Signal to Noise Ratio (SNR) higher than the males, with a mean of 28 and standard deviation of 10.2, with 90% of the females had SNR higher than 10.2 whereas males had an SNR with a mean of 4.8 and standard deviation of 6.9. The two misclassified females have the lowest SNR, less than 5, among the females. The 5 out of the 10 males that were misclassified had a higher SNR, ranging from of 11 to 22 with only 10 males being in that SNR range. A comparison between male and female SNR is given in Figure 19.

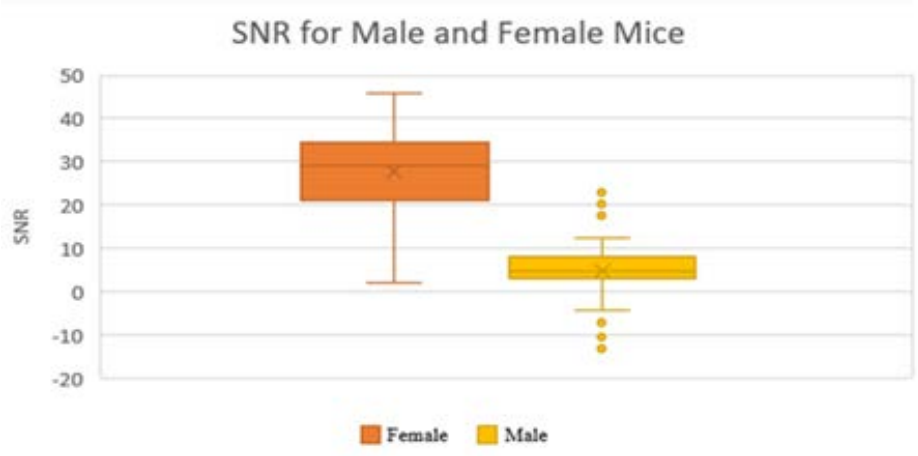

Figure 19: SNR comparison for Male and Female mice for Jump calls

Gender identification with Up Sweep calls had an accuracy 90.9%. The confusion matrix for gender identification using Up Sweep calls is given in Figure 20.

	C_female	C_male
female	113	2
male	14	47

Figure 20: Gender identification for Up Sweep calls (Accuracy 90.9%)

Up Sweep calls had the same pattern of errors with 13 out of 17 males were misclassified which had a SNR with mean 17.8 and standard deviation 4.0, with only 20 males in that SNR range. Females shows the same pattern of SNR as of the Jump calls with 93% of females had a mean SNR of 22.9 and standard deviation of 6.43.

Gender identification was also tested using Down Sweeps, Chirps and Inverse Chevron calls which yielded an accuracy more than that of chance except for Down Sweep. Down Sweeps had the lowest accuracy, 62.9%. For Down Sweeps the females with a lower SNR was classified as males and males with higher SNR were classified as females. The confusion matrix for Down Sweeps are shown in Figure 21.

	C_female	C_male
female	30	16
male	23	36

Figure 21: Gender identification for Down Sweep calls (Accuracy 62.9%)

	C_female	C_male
female	33	6
male	5	42

Figure 22: Gender identification for Chirp calls (Accuracy 87.2%)

	C_female	C_male
female	46	1
male	16	46

Figure 23: Gender identification for Inverse Chevron calls (Accuracy 84.4%)

Gender identification results using all the calls from all individuals had an accuracy of 88.9% with males and females having almost same accuracy and it followed same error pattern as of the Jump and Up Sweep calls. The confusion matrix for identification is given in Figure 24.

	C_female	C_male
female	501	67
male	49	426

Figure 24: Gender identification for all calls (Accuracy 88.9%)

3.6. Summary

This chapter has discussed speaker and gender identification experiments on ultrasonic mice vocalizations using GFCC features and GMM statistical classification. The results for individual identification and gender identification show that the Jump, Chirp

and Up Sweep calls contain individual and gender specific clues, with results significantly higher than chance. In the next chapter Speaker identification using meerkat Close calls will be discussed.

Chapter 4: Speaker identification in Meerkats

4.1. Overview

This chapter investigates using a Hidden Markov Model (HMM) - Gaussian Mixture Model (GMM) approach with Mel Frequency Cepstral Coefficients (MFCC) features for speaker identification in meerkats using Close calls. The vocalizations are segmented to extract voice activity and MFCC features are extracted from the segmented data, which in turn is used to train HMM – GMM and for testing purposes.

4.2. Data collection

Data for this study included vocalizations from 6 meerkats, 1 female and 5 males collected at the Kalahari Research Center, located in Kuruman River Reserve in Northern South Africa, during July - September of 2017. The data collection took place in the context of a long-term study on meerkats, the Kalahari Meerkat Project. These study subjects are used for research in movement coordination and determining the factors that influence the movement decisions as a group by Department of Evolutionary Biology and Environmental Studies, University of Zurich, Zurich, Switzerland. Vocalizations were recorded using collars attached to each meerkat that recorded GPS and audio, for 21 days of 3 hour-long sessions at a sample rate of 8 kHz. The audio recordings were manually labelled by analyzing the spectrograms using the software Adobe Audition CC 2018, which can embed labels into audio waveforms. Voice activity was extracted from all recordings using Adobe Audition and sorted according to the call type for each individual. There were multiple call types, including Close calls, Alarm, Move, Lead, and Aggression calls, in which Close calls were abundantly available for all individuals.

4.3. Experimental setup

Steps involved in classifying the vocalizations according to speaker include dividing vocalizations into training and test sets, extracting features from vocalizations using MFCC, training the HMM – GMMs using the labeled training set and classifying the vocalizations in test set according to classification criteria. This work is implemented using the Recognition Toolkit (RTK) (RTK 2004) and the Hidden Markov Model Toolkit (HTK) (Young, Evermann et al. 2002).

HTK is a commonly used toolkit for building HMMs. Although HTK is designed for recognizing human speech, in this work it has been adapted for speaker recognition tasks in meerkats using Close calls. HTK uses Baum – Welch algorithm to train each speaker model and uses Viterbi algorithm for classification of test data as explained in section 2.2.3 and 2.2.4.

RTK is a MATLAB graphical user interface to HTK which can help with sound recognition in animals. RTK can categorize the data, create labels for the data, create configuration files and prepare the vocalizations so that it can be trained and classified by HTK. The screen shot for graphical user interface of RTK is given in Figure 25.

Create category is used to organize the vocalizations according to the task being implemented. In this work vocalizations are categorized according to each individual. HTK requires Master Label Files (MLF) to train the vocalizations. Create label is used to automatically create the MLFs. Configuration is used to create the config file that specifies the type of feature extraction to be used, window size, step size, maximum frequency, minimum frequency, number of filters, number of cepstral coefficients and pre-emphasis value. The configuration file can also specify the features to be extracted as short-term

energy, delta, acceleration and zero crossing rates. The configuration files are used to customize HTK for a specific task.

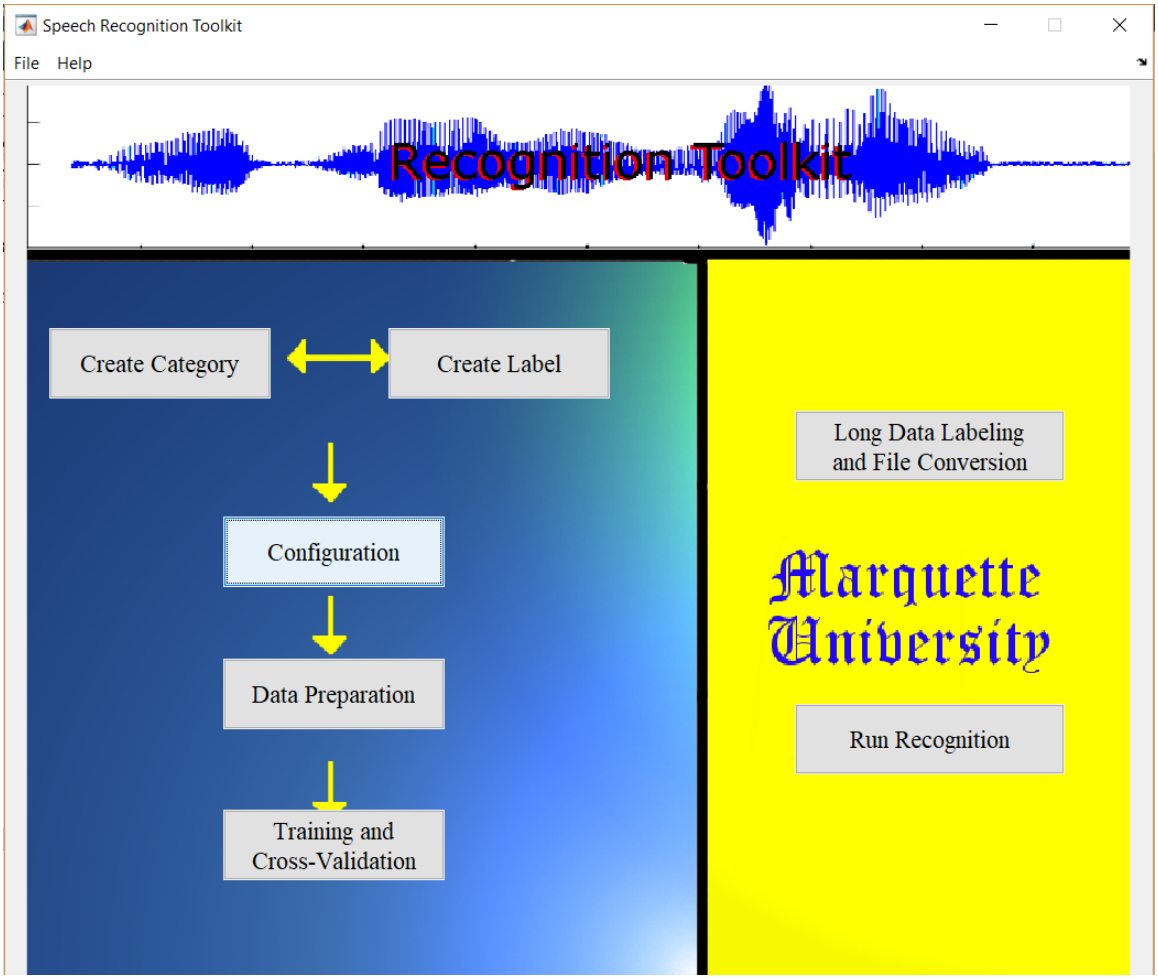

Figure 25: Recognition Toolkit(RTK) user interface

Data preparation extracts GFCC features from the vocalizations according to the specifications given in the configuration file. The training and cross validation trains the data by specifying the number of states and GMM mixtures that should be used in the modeling of speakers using HMMs. Recognition is used to recognize the speaker from the test data using the trained models for each speaker.

The speaker identification with Close calls was conducted using both GMMs and HMMs for 6 individuals. Both HMM and GMMs were implemented using RTK and HTK.

In this work a 10-fold cross validation is used to select test and training sets because of good amount of data available for each speaker. Since the duration of Close calls available for this work ranges from 38ms to 322ms and Close calls have a low frequency range, a window size of 5ms was used. The step size was selected as half the window size. GMMs were implemented using a single state HMM with 32 mixtures. For HMMs 16 states and 16 mixtures were used for training each speaker model. For both GMMs and HMMs short term energy is used to augment the feature vector.

4.3.1. Feature extraction

Greenwood Function Cepstral Coefficients (GFCC), as described in section 2.3.2.1, were used as features in this experiment.

The Greenwood frequency warping constants were found within the frequency range of 600 to 1000 Hz (Townsend, Hollén et al. 2010) for the Close calls. To make sure the maximum use of species frequency range, the minimum and maximum frequency was set as 400Hz and 1200Hz respectively. 12 GFCC coefficients are extracted from each frame. The constants for calculating the GFCCs is found as follows,

$$k = 0.88, \tag{4.1}$$

$$A = \frac{f_{\min}}{1-k} = \frac{400}{1-0.88} = 3333.33, \tag{4.2}$$

$$a = \log_{10}\left(\frac{f_{\max}}{A} = k\right) = \log_{10}\left(\frac{1200}{3333.33} + 0.88\right) = 1.24. \tag{4.3}$$

A MATLAB interface RTK designed for HTK was used for feature extraction as described in the previous section 4.3. The additional feature vectors used were short-term energy, delta and delta-deltas which is explained in section 2.2.2.

4.3.2. Model training

The features from each vocalization are extracted as GFCCs as shown previously in Section 2.2.3.2. By using Baum – Welch re-estimation the model parameter for HMMs are calculated that optimizes the likelihood of the training set of each speaker.

4.3.3. Identification

A maximum likelihood classifier for HMMs as explained in Section 2.2.4 was used in this work, implemented using HTK via the RTK toolkit. The likelihood of observation vector from each vocalization given the speaker model for each speaker was calculated and the speaker with maximum likelihood was selected as the predicted speaker. In HTK, the Viterbi decoder is used for finding the maximum likelihood.

4.4. Speaker identification in Meerkats

Speaker identification was implemented using Close calls because of the large number of data available. Vocalizations were sorted according to each speaker and the GFCC features were extracted from both train and test sets using methods explained in Section 4.3.1. The model with maximum likelihood is selected as the predicted class as explained in previous section 4.3.3. The results are displayed using a confusion matrix.

4.4.1. Subjects

Six meerkats, 1 female and 5 males, with Close calls ranging from 67 to 337 in number were used in this experiment. Although there were other types of calls, only a very few calls were not labeled as Close calls. The call distribution of each meerkat for Close calls is given in the Table 6 below. HMB is oldest and dominant male, HRT and HTB are

oldest subordinates, LT and RT are youngest subordinates. The female meerkat is designated HTB.

Table 6: Call distribution of Close calls

Individual	Number of calls
HMB	101
HRT	67
HTB	188
LT	440
PET	207
RT	77

4.4.2. Results

As described in Section 4.3 speaker identification was implemented for 6 meerkats with Close calls using GMMs and HMM-GMMs. Vocalizations were framed using a window size of 5ms and GFCC features were extracted as explained in Section 4.3.1. For Identification using GMMs each individual was modeled using 32 mixtures and identification was implemented using maximum likelihood classifier as explained in Section 4.3.3. Even though there were more vocalizations for training each speaker model, since meerkat vocalizations were less spectrally complex the number of mixtures required to model each individual was less than that of mice data.

Results of speaker identification using GMMs are given below in Figure 26. The confusion matrix for identification using HMMs is shown in Figure 27.

	C_HMB	C_HRT	C_HTB	C_LT	C_PET	C_RT
HMB	97	0	0	0	4	0
HRT	0	52	13	1	0	0
HTB	0	11	173	0	3	1
LT	0	4	5	390	4	37
PET	0	1	1	1	202	2
RT	0	0	2	7	1	67

Figure 26: Speaker identification in Meerkats with Close calls using GMMs (Accuracy 90.8%, Chance 40.7%)

	C_HMB	C_HRT	C_HTB	C_LT	C_PET	C_RT
HMB	98	0	0	0	3	0
HRT	0	51	13	2	1	0
HTB	0	9	174	0	5	0
LT	0	1	3	420	1	12
PET	0	1	0	1	204	1
RT	0	1	1	4	1	70

Figure 27: Speaker identification in Meerkats with Close calls using HMM - GMMs (Accuracy 94.4%, Chance 40.6%)

Speaker identification using GMMs had an accuracy of 90.8% and using HMMs had an accuracy of 94.4%. The results suggest two pairs of confusable individuals, HRT and HTB as well as LT and RT. Errors between these pairs represented more than half the

identification errors for the HMM results, totaling 38 of the 60 errors made. HRT and HTB are the oldest subordinates and LT and RT are the youngest subordinates. Each of these confusable pairs have similar social status within the group, which suggests the possibility that call similarity could also be connected to dominance or role within the social group structure. The dominant male HMB shows a higher accuracy even though vocalizations for that individual was quite noisy with most of the vocalizations having a low Signal to Noise Ratio (SNR).

When speaker identification was implemented using various state and mixture combination, there was a gradual increase in the accuracies, as shown in Table 7. Although the state mixture combination of 16 and 16 were selected for the speaker classification, the general trend was that accuracy continued increasing both with number of mixtures and number of states. The increase with mixtures is expected, due to increased spectral resolution of the state distributions, but the increase corresponding to a larger number of states suggests that there is additional temporal or timing information that is individually identifying as well.

Table 7: HMM number of states vs number of mixtures

No of mixtures \ number of states	4	6	8	10	12	14	16	18
2	90.7	91.4	91.3	90.6	92.2	92.1	92.6	92.2
4	91.3	92.3	92.5	92.5	92.5	92.9	93.5	93.6
8	91.1	92.5	92.5	93.3	93.8	94.1	94.1	93.4
16	92.9	92.8	93.9	93.7	93.4	93.6	94.4	94.3
32	91.9	93.2	93.9	93.4	94.1	94.0	94.1	93.7

4.5. Summary

This chapter has discussed speaker identification experiments on meerkat vocalizations using GFCC features and HMM-GMM statistical classification with slightly higher accuracy for identification using HMMs. The results for individual identification shows that Close calls contain speaker specific cues, both in spectral and temporal domain of the vocalizations.

Chapter 5: Conclusion and future work

5.1. Overview

This work has focused on the identification of speaker and gender from bioacoustic data sets, using vocalizations from mice and meerkats. A feature extraction technique explicitly developed for animal vocalization analysis, Greenwood Function Cepstral Coefficients (GFCC), is used to extract the features from vocalizations, which are modeled and classified using the statistical modeling techniques Gaussian Mixture Models (GMM) and Hidden Markov Models (HMM).

5.2. Summary of contribution and significance

The main contribution of this work is the application of human speech technology to bioacoustic data sets, for the tasks of speaker and gender identification. This study examines the extendibility of GFCC feature extraction, GMMs and HMMs for individual and gender identification for animal vocalizations.

In the study using mice data set, the presence of speaker and gender cues in ultrasonic vocalizations is supported by the classification results. Although mice vocalizations have been studied for decades, the incorporation of speech processing methods such as GFCCs and GMMs may help better understand and easy to analyze the communication and behavior in this species.

In the study using meerkat Close calls, presence of speaker specific cues was found in Close calls. GFCCs and HMM implementations both indicated that the temporal sequence of Close calls contains speaker specific data. The comparison of speaker

identification using GMMs and HMMs has shown a slight improvement in results for HMMs.

5.3. Future work

One potential experiment to extend this work would be to automatically detect the vocalizations from the long recording data using HMMs, rather than using manual timestamps, for further implement speaker and gender identification. This can be done by modeling noise data and all call types using HMMs and connecting them in a parallel loop so that the detection runs through the whole recording.

Another potential experiment, for the meerkat data in particular, could be identifying individuals using collar recordings from another individual. This would allow exploration of the relationship between SNR and identification accuracy in much noisier vocalizations. Being able to do this automatically could be a further support the application of speaker identification methods to problems in bioacoustics such as acoustic censusing.

Bibliography

Adi, K., et al. (2010). "Acoustic censusing using automatic vocalization classification and identity recognition." J Acoust Soc Am 127(2): 874-883.

Benesty, J., et al. (2007). Springer Handbook of Speech Processing, Springer-Verlag New York, Inc.

Boughman, J. W. and G. S. Wilkinson (1998). "Greater spear-nosed bats discriminate group mates by vocalizations." Anim Behav 55(6): 1717-1732.

Brown, J. C. and P. J. Miller (2007). "Automatic classification of killer whale vocalizations using dynamic time warping." J Acoust Soc Am 122(2): 1201-1207.

Brown, J. C., et al. (2010). "Automatic identification of individual killer whales." J Acoust Soc Am 128(3): El93-98.

Burke, K., et al. (2017). "Exposure-induced changes in laboratory mouse ultrasonic vocalizations." J Acoust Soc Am 142(4): 2596-2596.

Carlson, G. and C. H. Trost (1992). "Sex Determination of the Whooping Crane by Analysis of Vocalizations." The Condor 94(2): 532-536.

Charrier, I. and R. G. Harcourt (2006). "Individual Vocal Identity in Mother and Pup Australian Sea Lions (Neophoca cinerea)." Journal of Mammalogy 87(5): 929-938.

Clemins, M. T. J. a. P. J. (2017). Hidden Markov Model Signal Classification. Comparative Bioacoustics: An Overview, Bentham Science: 358-414.

Clemins, P. J. and M. T. Johnson (2006). "Generalized perceptual linear prediction features for animal vocalization analysis." J Acoust Soc Am 120(1): 527-534.

Clemins, P. J., et al. (2005). "Automatic classification and speaker identification of African elephant (Loxodonta africana) vocalizations." J Acoust Soc Am 117(2): 956-963.

Clemins, P. J., et al. (2006). Generalized Perceptual Features for Vocalization Analysis Across Multiple Species. 2006 IEEE International Conference on Acoustics Speech and Signal Processing Proceedings.

Clutton-Brock, T. H., et al. (2005). "'False feeding' and aggression in meerkat societies." Animal Behaviour 69(6): 1273-1284.

Collier, K., et al. (2017). "Call concatenation in wild meerkats." Animal Behaviour 134: 257-269.

Davis, S. and P. Mermelstein (1980). "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences." IEEE Transactions on Acoustics, Speech, and Signal Processing 28(4): 357-366.

Deller, J. R., et al. (2000). Discrete-time processing of speech signals. New York, Institute of Electrical and Electronics Engineers.

Dvorakova, V., et al. (2017). "Mashona Mole-Rat Automatic Individual Identification Based on the Mating Calls." bioRxiv.

Eda-Fujiwara, H., et al. (2004). "Sexual dimorphism of acoustic signals in the oriental white stork: non-invasive identification of sex in birds." Zoolog Sci 21(8): 817-821.

Ehret, G. (1992). Categorical perception of mouse-pup ultrasounds in the temporal domain.

Favaro, L., et al. (2015). Vocal individuality cues in the African penguin (Spheniscus demersus): a source-filter theory approach.

Fischer, J. and K. Hammerschmidt (2011). "Ultrasonic vocalizations in mouse models for speech and socio-cognitive disorders: insights into the evolution of vocal communication." Genes Brain Behav 10(1): 17-27.

Gales, M. and S. Young (2007). "The application of hidden Markov models in speech recognition." Found. Trends Signal Process. 1(3): 195-304.

Gall, G. E. C., et al. (2017). "As dusk falls: collective decisions about the return to sleeping sites in meerkats." Animal Behaviour 132: 91-99.

Garland, E. C., et al. (2015). "Beluga whale (Delphinapterus leucas) vocalizations and call classification from the eastern Beaufort Sea population." J Acoust Soc Am 137(6): 3054-3067.

Green, S. M. (1981). "Sex Differences and Age Gradations in Vocalizations of Japanese and Lion-tailed Monkeys (Macaca fuscata and Macaca silenus)1." American Zoologist 21(1): 165-183.

Greenwood, D. D. (1990). "A cochlear frequency-position function for several species--29 years later." J Acoust Soc Am 87(6): 2592-2605.

Hahn, A. H., et al. (2013). "Female song in black-capped chickadees (Poecile atricapillus): acoustic song features that contain individual identity information and sex differences." Behav Processes 98: 98-105.

Hammerschmidt, K., et al. (2012). "The Structure and Usage of Female and Male Mouse Ultrasonic Vocalizations Reveal only Minor Differences." PLOS ONE 7(7): e41133.

Hanson, J. L. and L. M. Hurley (2012). "Female Presence and Estrous State Influence Mouse Ultrasonic Courtship Vocalizations." PLOS ONE 7(7): e40782.

Heckman, J., et al. (2016). "Determinants of the mouse ultrasonic vocal structure and repertoire." Neurosci Biobehav Rev 65: 313-325.

Hermansky, H. (1990). "Perceptual linear predictive (PLP) analysis of speech." J Acoust Soc Am 87(4): 1738-1752.

Hinton, G., et al. (2012). "Deep Neural Networks for Acoustic Modeling in Speech Recognition: The Shared Views of Four Research Groups." IEEE Signal Processing Magazine 29(6): 82-97.

Hoffmann, F., et al. (2012). "Spectrographic analyses reveal signals of individuality and kinship in the ultrasonic courtship vocalizations of wild house mice." Physiol Behav 105(3): 766-771.

Holy, T. E. and Z. Guo (2005). "Ultrasonic Songs of Male Mice." PLOS Biology 3(12): e386.

Huang, X., et al. (2001). Spoken Language Processing: A Guide to Theory, Algorithm, and System Development, Prentice Hall PTR.

Janik, V. M., et al. (2006). "Signature whistle shape conveys identity information to bottlenose dolphins." Proc Natl Acad Sci U S A 103(21): 8293-8297.

Ji, A., et al. (2013). "Discrimination of individual tigers (Panthera tigris) from long distance roars." J Acoust Soc Am 133(3): 1762-1769.

Khanna, H., et al. (1997). "DIGITAL SPECTROGRAPHIC CROSS-CORRELATION: TESTS OF SENSITIVITY." Bioacoustics 7(3): 209-234.

Kogan, J. A. and D. Margoliash (1998). "Automated recognition of bird song elements from continuous recordings using dynamic time warping and hidden Markov models: a comparative study." J Acoust Soc Am 103(4): 2185-2196.

LePage, E. L. (2003). "The mammalian cochlear map is optimally warped." J Acoust Soc Am 114(2): 896-906.

Li, X. (2007). SPEech Feature Toolbox (SPEFT) Design and Emotional Speech Feature Extraction, Marquette University.

Li, X., et al. (2007). Stress and Emotion Classification using Jitter and Shimmer Features. 2007 IEEE International Conference on Acoustics, Speech and Signal Processing - ICASSP '07.

Litvin, Y., et al. (2007). "Rat 22kHz ultrasonic vocalizations as alarm cries." Behavioural Brain Research 182(2): 166-172.

Mausbach, J., et al. (2017). "Meerkat close calling patterns are linked to sex, social category, season and wind, but not fecal glucocorticoid metabolite concentrations." PLOS ONE 12(5): e0175371.

McLachlan, G. and D. Peel (2004). Finite Mixture Models, Wiley.

Mellinger, D. K. and C. W. Clark (1997). "Methods for automatic detection of mysticete sounds." Marine and Freshwater Behaviour and Physiology 29(1-4): 163-181.

Mellinger, D. K. and C. W. Clark (2000). "Recognizing transient low-frequency whale sounds by spectrogram correlation." J Acoust Soc Am 107(6): 3518-3529.

Mitani, J. C. and J. Gros-Louis (1995). "Species and sex differences in the screams of chimpanzees and bonobos." International Journal of Primatology 16(3): 393-411.

Murry, T. and S. Singh (1980). "Multidimensional analysis of male and female voices." J Acoust Soc Am 68(5): 1294-1300.

Musolf, K., et al. (2010). Ultrasonic courtship vocalizations in wild house mice, Mus musculus musculus.

Pollard, K. A. and D. T. Blumstein (2011). "Social group size predicts the evolution of individuality." Curr Biol 21(5): 413-417.

Portfors, C. V. (2007). "Types and functions of ultrasonic vocalizations in laboratory rats and mice." J Am Assoc Lab Anim Sci 46(1): 28-34.

Rabiner, L. R. (1989). "A tutorial on hidden Markov models and selected applications in speech recognition." Proceedings of the IEEE 77(2): 257-286.

Reber, S. A., et al. (2013). "Social monitoring via close calls in meerkats." Proc Biol Sci 280(1765): 20131013.

Reby, D., et al. (2006). "Cepstral coefficients and hidden Markov models reveal idiosyncratic voice characteristics in red deer (Cervus elaphus) stags." J Acoust Soc Am 120(6): 4080-4089.

Ren, Y., et al. (2009). "A Framework for Bioacoustic Vocalization Analysis Using Hidden Markov Models." Algorithms 2(4).

Rendall, D., et al. (2004). "Sex differences in the acoustic structure of vowel-like grunt vocalizations in baboons and their perceptual discrimination by baboon listeners." J Acoust Soc Am 115(1): 411-421.

Rendall, D., et al. (1996). "Vocal recognition of individuals and kin in free-ranging rhesus monkeys." Animal Behaviour 51(5): 1007-1015.

Reynolds, D. A. (1995). "Speaker identification and verification using Gaussian mixture speaker models." Speech Commun. 17(1-2): 91-108.

Reynolds, D. A., et al. (2000). "Speaker Verification Using Adapted Gaussian Mixture Models." Digit. Signal Process. 10(1): 19-41.

Roch, M. A., et al. (2007). "Gaussian mixture model classification of odontocetes in the Southern California Bight and the Gulf of California." J Acoust Soc Am 121(3): 1737-1748.

RTK (2004). "http://speechlab.eece.mu.edu/dolittle/proj_reu.html."

Sakoe, H. and S. Chiba (1978). "Dynamic programming algorithm optimization for spoken word recognition." IEEE Transactions on Acoustics, Speech, and Signal Processing 26(1): 43-49.

Sales, G. D. (1972). "Ultrasound and aggressive behaviour in rats and other small mammals." Animal Behaviour 20(1): 88-100.

Shapiro, A. D. (2010). Chapter 11.4 - Recognition of individuals within the social group: signature vocalizations. Handbook of Behavioral Neuroscience. S. M. Brudzynski, Elsevier. 19: 495-503.

Stevens, S. S. and J. Volkmann (1940). "The Relation of Pitch to Frequency: A Revised Scale." The American Journal of Psychology 53(3): 329-353.

Tao, J., et al. (2008). "Acoustic model adaptation for ortolan bunting (Emberiza hortulana L.) song-type classification." J Acoust Soc Am 123(3): 1582-1590.

Terry, A. M., et al. (2005). "The role of vocal individuality in conservation." Frontiers in Zoology 2(1): 10.

Titze, I. R. (1989). "Physiologic and acoustic differences between male and female voices." J Acoust Soc Am 85(4): 1699-1707.

Townsend, S. W., et al. (2012). "A simple test of vocal individual recognition in wild meerkats." Biol Lett 8(2): 179-182.

Townsend, S. W., et al. (2010). "Meerkat close calls encode group-specific signatures, but receivers fail to discriminate." Animal Behaviour 80(1): 133-138.

Townsend, S. W., et al. (2012). "Flexible alarm calling in meerkats: the role of the social environment and predation urgency." Behavioral Ecology 23(6): 1360-1364.

Trawicki, M. B., et al. (2005). Automatic Song-Type Classification and Speaker Identification of Norwegian Ortolan Bunting (Emberiza Hortulana) Vocalizations. 2005 IEEE Workshop on Machine Learning for Signal Processing.

Trifa, V. M., et al. (2008). "Automated species recognition of antbirds in a Mexican rainforest using hidden Markov models." J Acoust Soc Am 123(4): 2424-2431.

Volodin, I., et al. (2009). The technique of noninvasive distant sexing for four monomorphic dendrocygna whistling duck species by their loud whistles.

Volodin, I. A., et al. (2011). "Nasal and Oral Calls in Juvenile Goitred Gazelles (Gazella subgutturosa) and their Potential to Encode Sex and Identity." Ethology 117(4): 294-308.

Volodin, I. A., et al. (2015). "Gender identification using acoustic analysis in birds without external sexual dimorphism." Avian Research 6(1): 20.

Whitney, G. (1970). "Ontogeny of sonic vocalizations of laboratory mice." Behavior Genetics 1(3): 269-273.

Wu, K. and D. G. Childers (1991). "Gender recognition from speech. Part I: Coarse analysis." J Acoust Soc Am 90(4 Pt 1): 1828-1840.

Yang, C., et al. (2007). "Static and Dynamic Spectral Features: Their Noise Robustness and Optimal Weights for ASR." IEEE Transactions on Audio, Speech, and Language Processing 15(3): 1087-1097.

Young, A. J. and S. L. Monfort (2009). "Stress and the costs of extra-territorial movement in a social carnivore." Biol Lett 5(4): 439-441.

Young, S., et al. (2002). The HTK book.

Zippelius, H.-M. and W. M. Schleidt (1956). "Ultraschall-Laute bei jungen Mäusen." Naturwissenschaften 43(21): 502-502.

Vita

Neenu Jose, Master Student

Department of Electrical and Computer Engineering, University of Kentucky

---

EDUCATION

Master of Science - Electrical and Computer Engineering, May 2018
University of Kentucky
Thesis: Speaker and Gender identification using Bioacoustic datasets.
Director of Thesis: Dr. Michael T Johnson

Bachelor of Science – Biomedical Engineering, June 2007
Calicut University.