Modeling Lemur vocalizations from a signal processing perspective

Lorenzo Porcaro

MASTER THESIS UPF / 2015

Master in Sound and Music Computing

Master thesis supervisors:
Dr. Jordi Bonada
Department of Information and Communication Technologies
Universitat Pompeu Fabra, Barcelona

Dr. Marco Gamba
Department of Animal and Human Biology
University of Torino, Torino

UNIVERSITAT
POMPEU FABRA

Copyleft: 2015 Lorenzo Porcaro. This is an open-access document distributed under the terms of the Creative Commons Attributions License 3.0 Unported, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited

Acknowledgments

The work done during last months has been very exciting, but at the same time very stressful, and without the help of several persons, probably it could not be done.

Firstly, I would like to thank Dr. Jordi Bonada for the huge amount of knowledge shared, fundamental for continuing to have a high-level of motivation also in the darkest moments. Hoping that this work might help future research. Thanks to Dr. Marco Gamba for the support concerning the biological part and for having shared the database, fruit of his researches, the base of the whole study. Also, many thanks to Dr. Merlijn Blaauw for helping me for taking the first steps with HTK/HTS/SPTK toolkits.

Secondly, I would like to thank Prof. Xavier Serra and all the Music Technology Group (MTG), for introducing me into the world of scientific research and for giving me a lot of motivations and knowledge which hopefully will help me in my future career. In particular, thanks to Prof. Emilia Gómez and Dr. Jordi Janer, whose suggestions and advices have helped us students during all year long.

In conclusion, thanks to my family, to all my always present friends, and in general to who is still continuing to support me unconditionally. My work is mostly for you. Especially, thanks to every single *Renoizers*, who teach me what music means and what “work with music” means. Time spent together is a big part of who I am, and part of who I want to be, not only musically speaking.

Abstract

Most of the synthesis models for the generation of the animal vocalization until now have been studied with a non-probabilistic approach. Current studies are based on physical models, which have a large number of restrictions and limits for the creation of a realistic synthesis. In this thesis, we have created acoustic representations of black and white ruffed lemur (*Varecia variegata)* vocalizations, basing on signal processing techniques, and we have used them for achieving high-level synthesis. Afterwards, we have introduced Hidden Markov Models framework, approach which has been very successful in the context of speech synthesis. The outcome of the project has been evaluated by means of listening tests with humans, where original and synthetic vocalizations have been compared.

Contents

List of Figures

Chapter 1
INTRODUCTION

Animal acoustic communication is an interdisciplinary field involving research from various backgrounds, from biologists to linguists, to psychologists with a particular focus on cognitive sciences. The recent development of more powerful engineering techniques has helped the progress of scientific research, with increased focus on the signal processing perspective. As a result, interest in the study of Bioacoustics has increased, with progress and results in different contexts.

In this sense, we can identify the main perspectives of these studies: (1) the comparison of human and animal peculiarities in communication, which can provide insight into the evolutionary history of the human language; (2) analysis of animal vocalizations, which can lead to a better comprehension of animal behavior, as well as help recognition and classification tasks; (3) synthesis, which can improve the human-animal interaction, and at the same time the naturalness of playback experiments, widely used in biological research

The latter of the three is the main focus of this thesis and the goal of this work is to achieve a synthesis of animal vocalizations based on Hidden Markov Models. However, it is fundamental to review the works related to the other two categories presented before, for having a complete vision of the general framework of Bioacoustics research.

Chapter 2
STATE OF THE ART

The master thesis “Animal Vocalization Analysis and Synthesis” done by Graugés [Graugés, 2014] provides a good starting point for this study. It presents an extensive review of Lemur vocalization and the literature related to analysis and synthesis techniques, both in the case of speech and in animal vocalizations. Grauges has also covered some of the main concepts in depth and a reading of his works are thoroughly encouraged.

Consequently, in the following sections we have chosen to follow a different line, for outlining other common grounds between Biology and Signal Processing research. Before we will review the works related to comparative approaches between human and animal oral communication. After that, we will introduce general framework of analysis and synthesis techniques, focusing on techniques utilizing Hidden Markov Models.

2.1 Comparative approach

A comparative analysis of animal vocalizations and speech can be considered as a key point for understanding the evolution of the human language. In particular, identifying similarities and differences between non-human primate and our species can shed light on the evolution of modern day language and communications. Two main directions can be considered: analysing peripheral differences, hence how mechanisms of sound production and perception are based, and analysing neural differences, related to the perception and cognition of language. In both cases, it's possible to find relations between animals, not only primates, and humans [Fitch, 2000b].

2.1.1 Peripheral differences

Human speech production is a topic that has been widely studied during the last decades. The source-filter theory is the most affirmed model that describe it. Basically, it considers speech as a combination of a source function, a pulsating airflow passing through the larynx, and a vocal-tract filtering process, which conveys a large part of the information in speech [Fant, 1981].

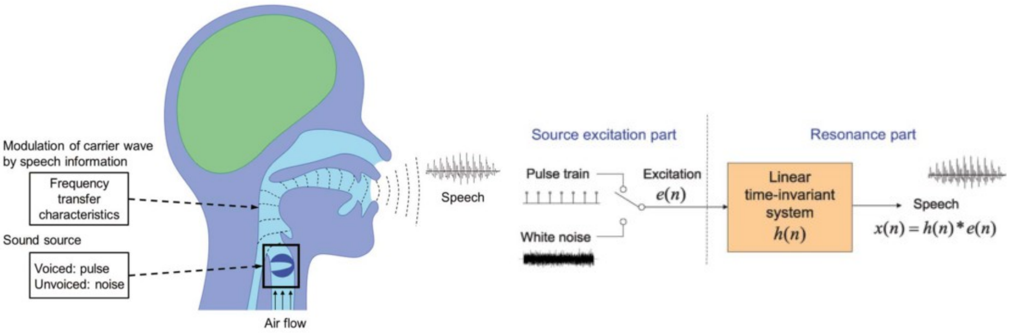


Figure 2.1: Overview of human speech production (left) and source-filter model (right) [Tokuda et al, 2013].

Although for a long time the human vocal productions system (VSP) has been considered unique, several studies have demonstrated the presence of common characteristics with other animal systems. In particular, three main component are shared in the VSP of tetrapods: (1) a respiratory system with lungs; (2) a larynx that has primarily evolved to protect the lungs, and can actively produce sounds; and (3) a supralaryngeal vocal tract (or only “vocal tract”), that filters this sound before its release into the environment. In addition, there are other elements which historically have been considered unique in human, which recently has been discovered that are shared with different animals. For example the larynx position, which is permanently "lowered" in the human vocal tract. A study by Fitch and Reby [2001] threw light on the fact that the descent of the larynx is not uniquely human. In facts, other mammals, such as red and fallow deer, may lower their larynx during vocalization. Also movements of vocal tract articulators (lips, tongue, jaw, velum, larynx), which characterize formant variations, have been observed both in humans and animal production of sound [Fitch, 2000a].

Despite all these factors, widespread among humans and animals, it is evident that humans can produce sounds which animals do not, and vice versa. Formants, which have a key role in both human and animal acoustic communication [Gamba, 2014], provide great insight into this difference. Formant frequencies are correlated with the length and shape of the vocal tract, and thanks to this each species can produce different vocalizations. However, the ability to produce rapid and precise movements of vocal tract articulators, together with the characteristic of the decedent larynx, allows humans to produce a much wider range of formant patterns than other mammals, which characterize speech communication [Fitch, 2000b].

Nevertheless, it is evident that taking into account solely these results cannot be sufficient for explaining how speech communication has been evolved in a thus different way. Hence, it is necessary to analyze other differences based on neural aspects.

2.1.2 Neural differences

Syntax can be considered an exclusive aspect of human language. The human ability to create with a limited set of phonemes, a lot of words and combine them in infinite different mode creating meaningful sentences has not equal in the animal world [ten Kate and Okanoya, 2012]. Nevertheless, the basic structure of animal vocalizations presents some common features. Units of production, which can be considered as the small part in which we can segment vocalizations, are species-specific, and this is a consequence of the anatomical differences presented in the previous section. These units are combined in different vocalizations, which can convey different meanings, depending on how they are used. Two vocalizations can be used for creating different high-level structures, creating calls that convey different meanings [Arnold and Zuberbühler, 2006]. Hence, we can affirm that a fundamental difference between human and animal language is not the presence or not of an high-level structures of vocalizations, but it is more related to the complexity of that. [Berwick et al, 2011].

Another interesting observation is related to the function of vocal mimicry, which is fundamental to the learning process of human language. This characteristic, which is not common in mammals, is present in other animals, as birds, its influence on the evolution of speech remains unclear [Fitch, 2000b].

In conclusion, we can affirm that using a comparative approach several similarities in the acoustic communication of animals and humans can be found. This suggests the possibility to use the knowledge accumulated in studying speech to interpret animal calls, with due adaptations. Taking into account these observations, in the next section we will present several studies, showing the application of different techniques for the analysis of animal vocalizations. Afterwards, we will focus on the main topic of this thesis, the black and white ruffed lemur (*Varecia variegata*) vocal repertoire.

2.2 Analysis of animal vocalizations

The analysis of animal acoustic communication is one of the main tasks in the field of Bioacoustics. Bioacoustics is a powerful tool for understanding behaviour both at intra- and inter-specific level. Moreover, as a consequence of this analysis, more sophisticated tasks such as automatic recognition and classification, can achieve better results. However, considering the degree of diversity in the animal world, analysis methods can not be totally generalized, and specific studies have to be fine-tuned according to the specific case.

Below, we will introduce a methodological framework, resulting from a 2013 workshop entitled, 'Analyzing vocal sequences in animals' [Kershenbaum et al., 2014]. Then we will present some representative examples of the literature about the analysis of animal vocalization. They will provide hints regarding how different methods that have been used led to someway different results.

2.2.1 Methodological framework

The key point of the analysis of animal vocalizations is to define what it is a vocalization. We can consider it as a series of acoustic elements, or basic units, which can convey several information. Nowadays, the meanings of part of these vocalizations have been understood, but the function of a part of them it is still unclear. The general process proposed for validating hypothesis regarding meanings of different vocalization, as represented in Figure 2, is composed of four steps:

(1) Collecting the species vocalizations. In this passage, basic signal processing analysis is done, including pre-processing, filtering, time-frequency and time series analysis.

(2) Identifying basic units of vocalizations. Starting from the previous step, the aim is to find how a vocalization is built.

(3) Characterization of the vocalization. Obtaining several vocalizations, we identify particular features of each of them.

(4) Identifying the meaning of each vocalization. In this last step, the analysis of how vocalization are used, leads to validate or not hypothesis regarding the meaning.

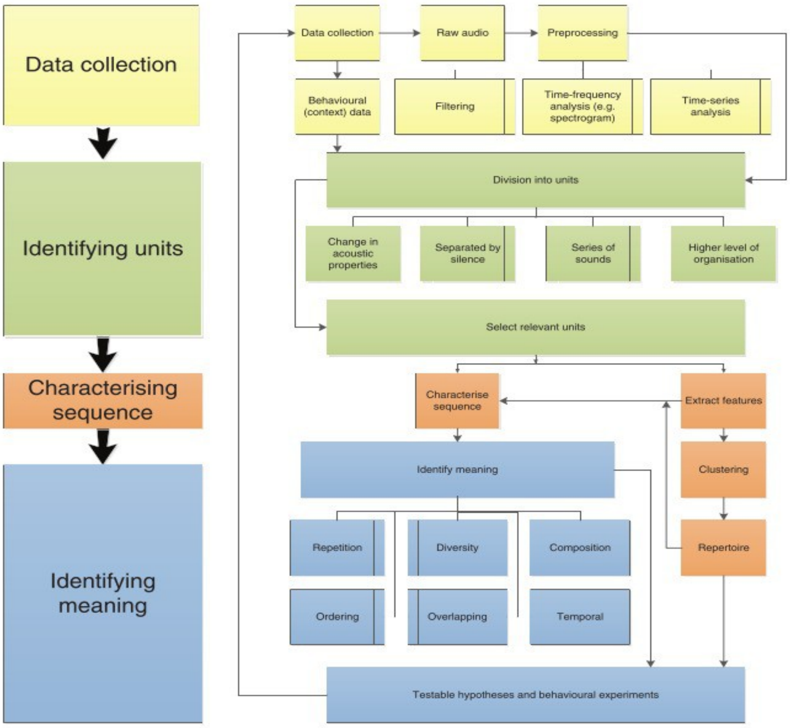


Figure 2.2: Flowchart showing a typical analysis of animal acoustic sequences [Kershenbaum et al., 2014].

This framework can be considered as a reference, however specific-cases adaptations have to be implemented to obtaining sounding results, mainly in the first two steps. From a perspective of signal processing, the enormous differences that can be found between vocalizations of different species, have led to applying several techniques depending on the particular case. In the next section, we will provide some examples that can give an idea of this variety.

2.2.2 Overview of specific-cases analysis

The features of animal acoustic communication vary dramatically between species, even if common patterns can be identified. This variability is evident in the process of production, which involves the morphology and the functional anatomy of the vocal tract. We can have species uttering in the infrasound like elephants and whales, and species emitting ultrasounds, like rodents and bats [Fitch, 2006]. Because of that, it is critical to consider different approaches basing on species-specific features.

In ornithology the presence in several species of rapid modulations and moreover the presence of numerous vocal units (often overlapped) led to the use of different methods for the analysis [Baker and Logue, 2003]. More recently innovative techniques have been applied to perform an automatic tracking of birdsong [Stowell et al., 2013]. Instead, for marine mammals, approaches not based on the Fourier analysis showed better results [Adam, 2006]. Classification tools coming from the music information retrieval field [Ness, 2013] were also useful. Regarding non-human primates, quantitative methods have started to be considered in addition to qualitative ones, which have been widely used historically [Gamba and Giacoma, 2007].

Apparently, the studies mentioned above do not represent the whole picture of animal communication research, but can give an idea of the multitude of approaches used. Below, we will present several kind of analysis of animal vocalization based on Hidden Markov Models.

2.2.3 Analysis of Animal Vocalization using Hidden Markov Models

Interest in Hidden Markov Models have grown in the field of Bioacoustics during last years. Mainly because of the huge versatility of this probabilistic approach, several results have been accomplished thanks to it. In tasks such as call-type classification, individual identification, and assessment of correlation between vocalization patterns and specific social or behavioral contexts, other analysis techniques have been used, such as multivariate feature analysis, spectrogram cross-correlation, matched filtering, neural networks, dynamic time warping and others. However, HMM-based systems have been shown to have a bigger percentage of accuracy [Ren et al., 2009]. A big advantage of HMMs, as statistical classification models, is that they are able to use any frame-based feature vector. Hence, it has been possible to adapt features in several specific contexts. For instance, instead of using Mel-Frequency Cepstral Coefficients (MFCCs) or Perceptual Linear Prediction (PLP) coefficients, widely affirmed in the human speech analysis, different features have been developed. Thanks to the work of Greenwood [Greenwood, 1961], generalization of the previous features have been created, called Greenwood Function Cepstral Coefficients (GFCCs) and generalized perceptual linear prediction (gPLP) coefficients, for being used in species-specific case of study [Clemins and Johnson, 2006]. In the next section, we will focus on the analysis of the study species of this thesis, the black and white ruffed lemur.

2.2.4 The Black and White Ruffed Lemur (*Varecia variegata)* vocalizations

We will introduce a brief description of the different vocalizations that characterize white and black ruffed lemur acoustic communication. For each call, we will use labels adopted in [Pereira et al., 1988]. In addition, further analysis has been done [Gamba and Giacoma, 2006], which will helps us understand the peculiarities of each vocalization, leading to a more precise categorization.

The recognized repertoire of white and black ruffed lemur is composed of 16 vocalizations. A first step is to subdivide calls into three sets: high-amplitude, moderate-amplitude, and low-amplitude.

High-amplitude calls

Roar/shriek Chorus

This is characterized by the presence of two vocalizations, shriek, and roar. Roar is a wide-band noisy sound while shriek is the frequency modulated narrow-band component. Duration can vary from 5 to 30 second. It is produced at the same time by several adult lemurs. It is used to maintain spacing between groups of free-ranging ruffed lemurs.

Abrupt roar

It comprises a rapid series of 2-5 short sound pulses. It is used commonly in presence of large birds predators.

Growl-snort

This is a low-pitched emission, in which the first part is emitted with mouth closed (growl), followed by an explosive expulsion of air (snort) including low-frequency sound energy. It is used when facing terrestrial predators or perceived threats from the ground.

Pulsed squawk

This is composed by a series of acoustic pulses with harmonic overtones with a duration of 2 to 4 seconds. Emitted synchronously by several members of a group, it can be given in the presence of a carnivore or can be emitted by males during the breeding season.

Wail

The last part of the pulsed squawk, when the rate of pulse decreases and pulses become more tonal and very elongated, the sound is called wail and is characterized by a richness of harmonic overtones. This may denote urgency for re-aggregation.

Moderate-amplitude calls

Growl

This is composed by a series of low-frequency laryngeal vibrations, and it can last from 1 to 4 seconds. It is commonly emitted in context of generalized low-level disturbance.

Chatter

This is a series of brief, high-pitched, wide-band units whose duration can be highly variable. Social subordinates usually directed chatter towards dominants, but also dominants can use chatters in the context of an inter-individual conflict.

Whine

It varies from wide-band noise with tonal undertones to warbled frequency modulation with numerous harmonics and reduced noise. It has been observed during the breeding season. Can also be uttered by the juveniles and the infants.

Brays

It exhibits a widespread of low-frequency energy during exhalation. It is produce by adult male lemurs, and it is observed only in conjunction at the roar/shriek chorus.

Quacks

Similar to brays, in addition it have a leading portion of tonal energy with harmonics overtones. Also in this case, the use is associated at the Roar/shriek chorus.

Low-amplitude calls

Grunts

These consist of a low-frequency tonal element presented as a train of pulses. It is emitted in context of mild disturbance.

Huffs

Huffs are produced by the passage of air through the nasal tract. It covers a wide frequency range and can be given in a rapid sequence. Occasionally, the duration can be unusually long and frequency range more narrow. It can appear to accompany decreasing arousal.

Mew

This is a tonal call, whose duration is about 0.8-1.0 second. It is often present a slow rise in pitch. The function is related to mother-offspring communication and serves to maintain contact between conspecifics in general.

2.3 Synthesis based on Hidden Markov Models

The literature related at the synthesis of animal vocalization not as extensive as its human counterpart. Probably because the interests in the field of Bioacoustics have been mainly oriented at the analysis of vocalizations, as we have presented before.

Nevertheless, it is possible to find examples of animal sound synthesis based on various approaches. As in [Clark et al, 1983], where a FFT-based software has been created for analyzing and then synthesizing bird vocalizations. Another synthesis based on spectral features is presented in [Dhar et al, 2010], which focuses on whale sounds. In a more creative way, in [Marino, 2000] where animal-like vocalizations are modeled using MAX/MSP, a sound design software. On the contrary, examples of synthesis generated on probabilistic models, as the one presented in this thesis, have not been found. This kind of synthesis has been very successful within the context of human voice, leading to notable results, one of the reason for our choice to use HMMs for achieving a new kind of synthesis.

Consequently, in the next section we will present a series of concepts extracted from [Tokuda et al, 2013], which summarizes the framework related to speech synthesis based on Hidden Markov Models. This will be the key to the full understanding of this thesis.

2.3.1 Speech Synthesis

Voice production

The source-filter model, already introduced in the previous sections, is capable of modeling the human voice based on few features.

With the fundamental frequency (f0) and a spectral envelope, it is possible to reconstruct speech waveforms. The aim of HMM is to predict the parameters related to these features, basing on an initial input. A classic case where HMM can be applied are Text-to-Speech (TTS) system, where the initial input is, specifically, written text.

Hidden Markov Model

An N-state hidden Markov model λ, is characterized by sets of initial-state probabilities $\{\pi_i\}_{i=1}^{N}$, state transition probabilities $\{a_{ij}\}_{i,j=1}^{N}$ and state-output probability distributions $\{b_i(\cdot)\}_{i=1}^{N}$ are typically assumed to be single multivariate Gaussian distribution :

$$\begin{aligned} b_i(\boldsymbol{o}_t) &= \mathcal{N}(\boldsymbol{o}_t; \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) \\ &= \frac{1}{\sqrt{(2\pi)^d |\boldsymbol{\Sigma}_i|}} \exp\left\{ -\frac{1}{2} (\boldsymbol{o}_t - \boldsymbol{\mu}_i)^\top \boldsymbol{\Sigma}_i^{-1} (\boldsymbol{o}_t - \boldsymbol{\mu}_i) \right\} \end{aligned}$$

where $\boldsymbol{\mu}_i$ is a mean vector, $\boldsymbol{\Sigma}_i$ is a covariance matrix, and $\boldsymbol{o}_t$ is an observation vector, which in speech case is formed by concatenating spectral and excitation parameter vectors.

The training of HMMs and synthesis from HMMs can be written as follows:

$$\begin{aligned} \text{Training: } \lambda_{\max} &= \arg\max_{\lambda} p(\boldsymbol{O} | \lambda, \mathcal{W}) \\ p(\boldsymbol{O} | \lambda, \mathcal{W}) &= \sum_{\forall \boldsymbol{q}} \pi_{q_0} \prod_{t=1}^{T} a_{q_{t-1} q_t} b_{q_t}(\boldsymbol{O}_t) \\ \text{Synthesis: } \boldsymbol{o}_{\max} &= \arg\max_{\boldsymbol{o}} p(\boldsymbol{o} | \lambda_{\max}, w) \end{aligned}$$

where $\boldsymbol{q} = \{q_1, q_2, \ldots, q_T\}$ is a state sequence. $\boldsymbol{O}$ and W are a set of speech parameters and corresponding linguistic specifications (such as phoneme labels) to be used for the training of HMMs, respectively, and $\boldsymbol{o}$ and w are speech parameters and corresponding linguistic specifications that we want to generate at synthesis time.

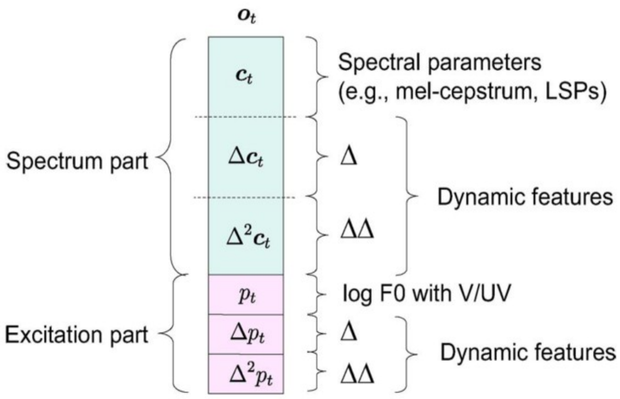


Figure 2.3: Example of an observation vector at each frame [Tokuda et al, 2013].

Speech Parameter Generation From HMM

The idea behind the generation of the speech parameters is that the most probable speech parameter vector sequence, given a set of HMMs and an input to be synthesized, is determined as solution of a maximization problem. Furthermore, to increase the naturalness of the synthesis, the speech parameter generation algorithm introduces the use of first- and second-order time derivatives of speech parameters as a part of the observation vector, which is a powerful mechanism for capturing time dependencies within the HMM framework.

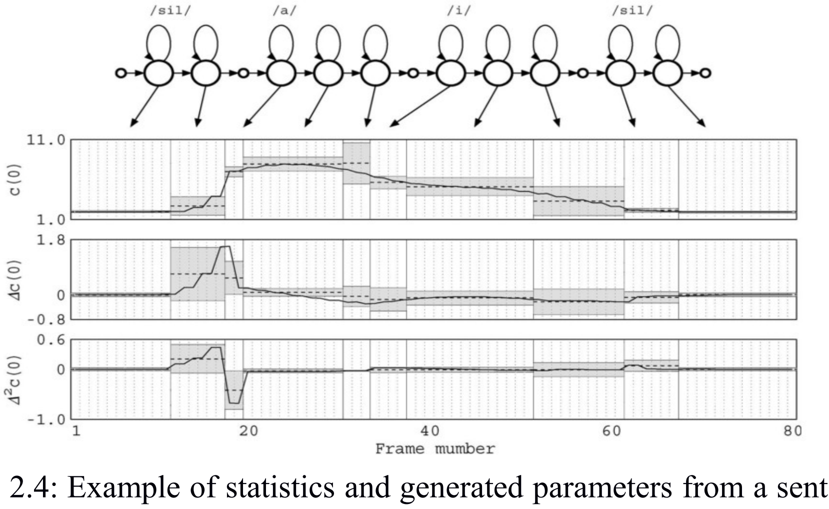


Figure 2.4: Example of statistics and generated parameters from a sentence-level HMM composed of phoneme-level HMMs for /a/ and /i/. The dashed line and shading show the mean and standard deviation, respectively, of a Gaussian pdf at each state [Tokuda et al, 2013].

Training part

The training part is where is performed the maximum-likelihood estimation of the HMM parameters. Apart from excitation and spectral features, other characteristics of speech are taken into account. For example, durations are considered for modeling the temporal structure of the speech parameter sequence. other information related to more general linguistic context, such as lexical stress, pitch accent, tone etc. can be introduced into the model, affecting prosodic and duration parameters.

Synthesis

In the synthesis part a sequence of speech parameters including spectral and excitation parameters is determined so as to maximize its output probability using the speech parameter generation algorithm. After that a speech waveform is synthesized directly from the generated spectral and excitation parameters by using a speech synthesis filter.

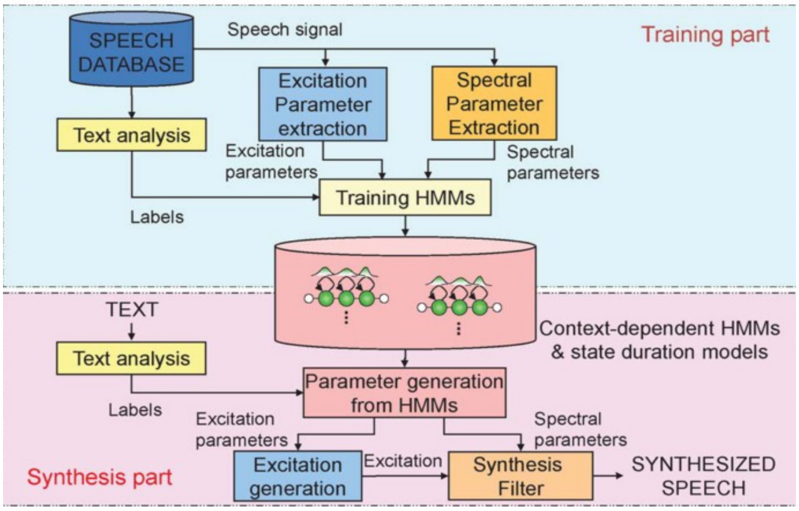


Figure 2.5: Overview of the HMM-based speech synthesis system [Tokuda et al, 2013].

2.4 Open Source Toolkits

Open source toolkits have been consistently used during this work, and they have been necessary for the development of several parts. In details, Hidden Markov Model Toolkit (HTK) and HMM-based Speech Synthesis System (HTS) are toolkits for building and manipulating hidden Markov models. The demo which has been used in a step of the work, has been created by the HTS working group. Furthermore, the Speech Signal Processing Toolkit (SPTK), which is a suite of speech signal processing tools, has been fundamental in the last part of the process, while generating the synthesis. In conclusion, also the use of Sound eXchange (SoX) has been necessary for performing several kind of sound processing.

Chapter 3
METHODOLOGY

The analysis and synthesis process of Lemurs vocalizations has been adapted starting from the general framework used in the human speech context. However, given the nature of these particular sounds, several considerations have to be done.

First, the lack of precedent similar works in this field has meant that the creation of an acoustic representation for each vocalization had not been straightforward. The procedure for extracting the features in several cases presented incompatibility with the ones normally used in speech-context. In these cases, it has been necessary starting from scratch, leading to specific representations.

Furthermore, the two vocalizations considered in this study, the Mew and the Growl, present quite different acoustic characteristics. It has meant that for each one the process of features extraction has been particularized. It is important to underline that the choice to focus on these two has been mainly arbitrary, but in part because of their vowel like acoustic structure [Gamba and Giacoma, 2006].

The next sections are organized as follow: before it will be introduced the dataset, with its strengths and weaknesses. Afterwards, we will identify the general strategy of analysis and synthesis which has been used for both the vocalizations, focusing the on the details of each steps done for achieving the final results.

3.1 Lemur vocalization database

The collaboration with Dr. Gamba has made possible this work, considering the huge database of Lemur vocalizations that he furnished us. The recordings have been done during a period of almost 10 years, in several places around the world. More than 50 specimen have been studied, with a range of age which start from newborn arriving at Lemurs 7 years old. In total, the whole database is composed by more than 8 thousand examples.

For our purposes and our not extended knowledge of Lemur vocalization, a fundamental tool has been the annotations related to this database which Dr. Gamba gave us. In this table, for each vocalization has been annotated: the specimen, the sex, the birth date, the date of recording, the context of recording, the type of vocalization, the age of the specimen during the recording and a number ID of the vocalization. Having these informations, we have been able to optimize our research within the database.

The only limitation of this database has been the quality of most of the recordings, as we will see the reasons in detail later. Indeed, for our kind of analysis approach, vocalizations too much obfuscated from other sources, as noises, other animals, and degradation of audio quality derived from echoes or distortion due to extreme proximity of the source, have meant that a large part of recordings could not be used.

3.2 General strategy

The first aim has been to create an optimal acoustic representation which could lead us to a good quality synthesis, regardless from using HMMs model. This means that we had to find features which could represent peculiarities of the signal for both vocalization. Furthermore, it is important to point up that a relevant characteristic of the feature set, which we had to keep in mind during the features selection, was that since we wanted to arrive at an HMM based statistical model, we were looking for rather stationary features, that evolved slowly over time, at least over the states of the HMM. As we will see, for Mew it did not present extreme difficulties, while for Growl it has been necessary a more complex research. Once obtained a robust acoustic representation, we used the selected feature set for generating a first synthesis.

After that, we have proceeded introducing HMMs framework. Our second aim, has been to achieve a HMM-based synthesis for single vocalizations. For this step, we have based our work on the general structure used for HMMs-based speech synthesis. We have used part of the script of the "Speaker dependent training demo (English-normal version)" released by the HTS Working Group. Details will be presented later.

Achieved this first HMM-based synthesis, last step has been to use several vocalizations for creating a model and then generating a synthesis. Difficulties had emerged in the selection of the examples for the training part of HMMs, which had required a deepened analysis of the behaviors of the features for each vocalization. Specific cases will be discussed. Surpassed this problem, we have obtained the final synthesis based on HMMs.

What described before can be considered the general strategy which has been used in both the vocalizations taken into account. However, specific characteristics in each case have been observed, and it has been necessary to model them for achieving a good quality synthesis. In next sections, before we will analyze the acoustic representation created for both vocalizations. After that, we will deepen the HMMs framework used, and in conclusion we will focus on the synthesis part.

3.3 Acoustic representation

The research and extraction of an adequate feature set for each vocalizations has been the hardest task. Starting from the assumption that also for Lemur vocalizations it can be applied the source-filter model, basically we needed to find two types of features: one for describing the excitation part and one for the spectrum part. In the field of speech synthesis, usually the fundamental frequency (f0) is used for creating the excitation signal, and spectral parameters, as Mel-Generalized Cepstral Coefficients (MGCCs) [Tokuda et al, 1994], are used for filtering the excitation and achieving the synthesis [Tokuda et al, 2013].

Therefore, we have chosen to maintain this pattern, even if we have encountered two kind of difficulties:

- algorithms for feature extraction designed for speech signals, often do not work correctly in presence of other kind of signals, as in our case.
- using only two features, f0 for the excitation signal and MGCCs for the spectral part, cannot be enough for modeling properly some kinds of Lemur vocalizations.

For solving these problems, it has been necessary to deepen the low-level analysis of the considered signals, and basing on empirical observations, create from scratch specific algorithms for extracting specific features.

3.3.1 Mew

The Mew has been considered a good starting point for its acoustic characteristics. As presented before, it is a tonal call where duration hardly exceeds 1 sec, and it has a slow increase in pitch. Furthermore, the presence of a great number of examples of this vocalization in the database has led us to choose it for starting our work.

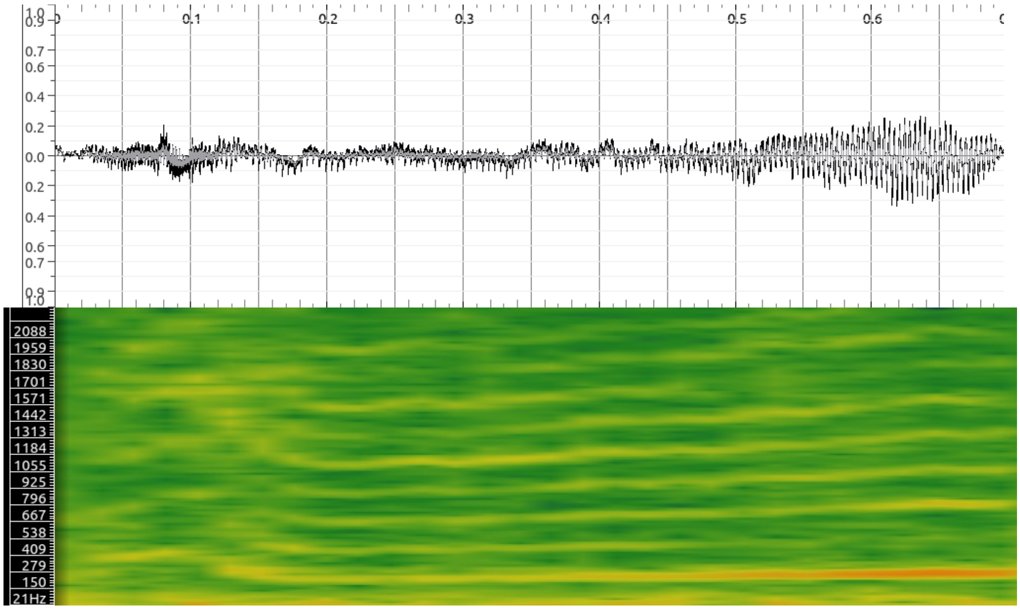


Figure 3.1: Example of Mew vocalization, waveform (top) and spectrogram (bottom)

a. F0 estimation and data normalization

The first problem has been encountered during the f0 extraction. In fact, the two algorithms implemented in the SPTK toolkit, RAPT [Talkin et al, 1995] and SWIPE [Camacho, 2007], are not extremely accurate with this kind of signals. This led to a synthesis where f0 trajectory was not well defined. For solving this problem, we chose to try with SAC algorithm, described in [Gómez and Bonada, 2013], thought for dealing with automatic transcription of flamenco music recordings, more specifically a cappella singing. Thanks to this, we have improved considerably the accuracy of f0 extraction.

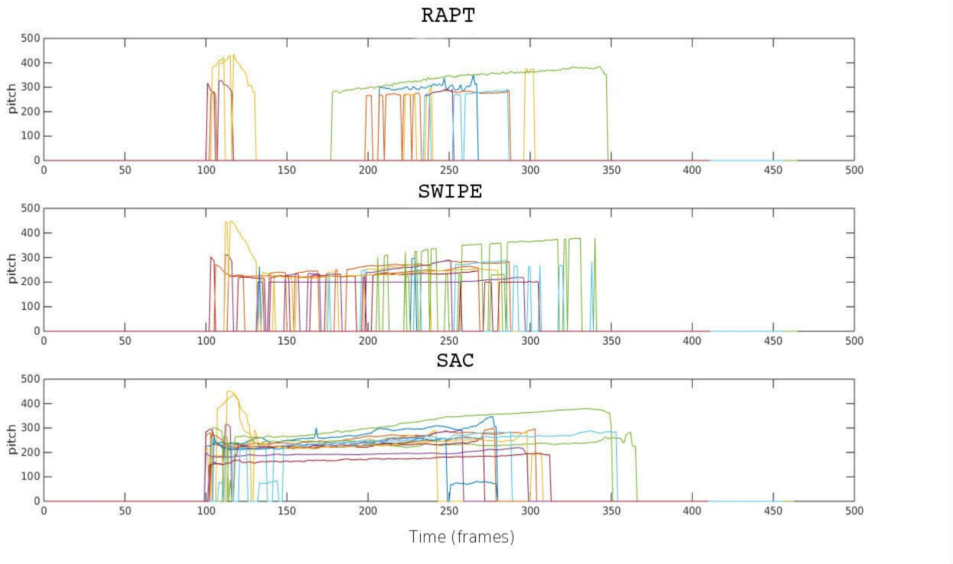


Figure 3.2: F0 estimation of several vocalizations performed with different algorithms. RAPT (top), SWIPE (center), SAC (bottom)

Once obtained a good f0 estimation, we had to deal with the big variety of the sounds. Indeed, behavior of f0 presents several differences between Mew vocalizations. A root of this diversity is that not all the vocalizations belong to the same individual and this can be a cause for differences in f0. This could be a problem when trying to use several vocalizations together, because mixing completely different behaviors into HMM may make misleading variations at the Gaussian models created. Hence, it has been necessary to analyze in deep how was substantial the difference between f0 trajectories.

For this reason, we have chosen to work with normalized data. With data which are quite different, using the absolute f0 values can lead to a really strong over smoothing. Furthermore, with the normalization it is possible to make the f0 feature frequency shift invariant and also frequency scale invariant. With this choice we achieved a more compact visualization of the data and a more reliable comparison between vocalizations.

We can describe the normalization process adopted as follow: first we selected the parts of the f0 estimation different from zero, achieving a vector f0 with only positive values. Namely, we considered only the voiced parts of the vocalization. After that, we computed a histogram of the elements in vector f0, sorted into 100 equally spaced bins along the x-axis between the minimum and maximum values of f0. Following, we performed a cumulative sum, achieving a minimum and maximum frequency considering only a range between the 10% and the 90% of the total frequency range of the vocalization. This passage has been fundamental for avoiding the influence of extremes behaviors of f0 trajectories in the normalization, due sometimes also by computational errors.

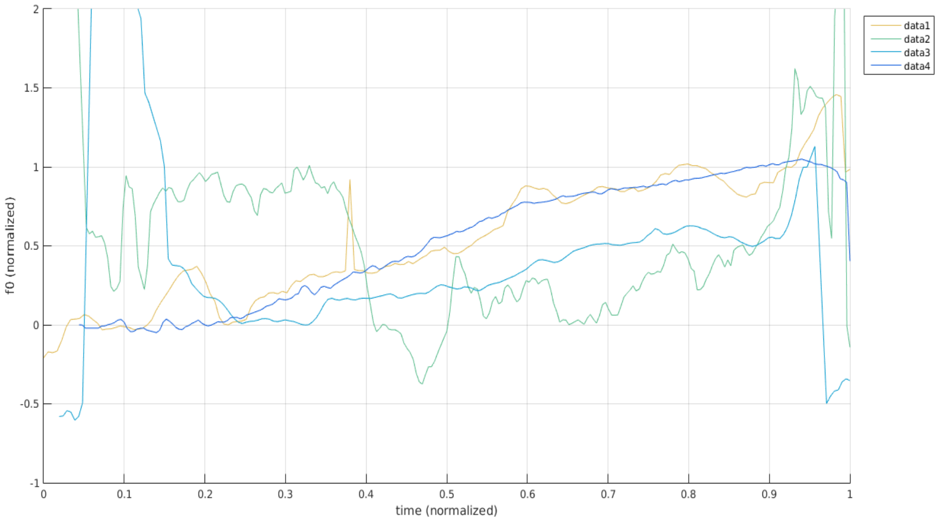


Figure 3.3: Plot of several examples of Mew f0 trajectories. In data1 we can observe a quasi-regular rise, however some computational errors are evident observing the unnatural peak around 0.4. Data2 presents is a case of very irregular trajectory of f0. In data3 is present an initial prominent peak which characterize some Mew. Data4 is a case where there is no presence of the initial peak and the rise of the trajectory is regular.

In conclusion, we obtained normalized data applying

$$f_{0,norm} = \frac{f_0 - f_{0,min}}{f_{0,max} - f_{0,min}}$$

b. Classification

The analysis of normalized f0 data has pointed out several kinds of trajectories within the considered set of Mew. Hence, our second purpose was to perform a valid classification. This passage has been necessary, because when putting together vocalizations with behaviors widely different in the training dataset for HMMs, the resultant synthesis loses its sense.

To begin with, we were able to discriminate between two main behaviors. In fact, in the range of 0-0.2 in the axis of time-normalized, some of the Mew present an initial prominent peak in the f0 trajectory. It characterizes the begin of the vocalization and it is clearly distinguishable when listening the vocalizations. Hence, a first subdivision of Mew was between examples with initial peak and without initial peak. For achieving this, it has been performed a comparison between the area of the first part of f0 estimation, 0-0.2 in time-normalized, and a constant area used as threshold for discriminating between the trajectories with or without initial peak. Simply, if the area of f0 was bigger than the threshold, the Mew were classified as with peaks, and if not without peaks. The threshold as been fixed after analysis of several examples of this vocalization.

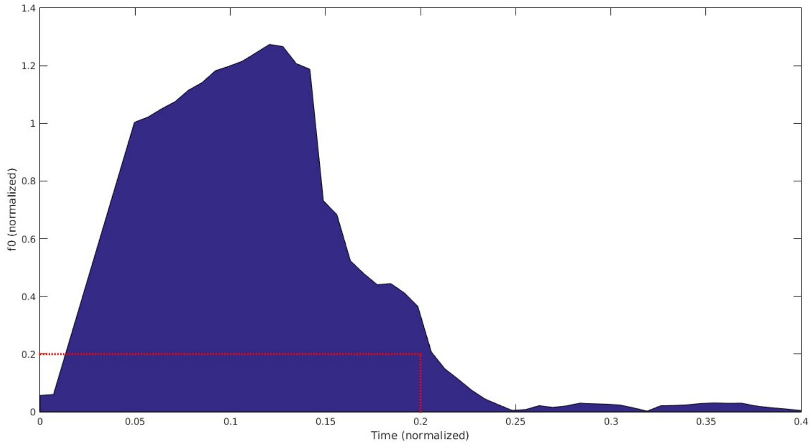


Figure 3.4 : Example of Mew classified as "with initial peak". The dotted red line delimits the threshold area considered.

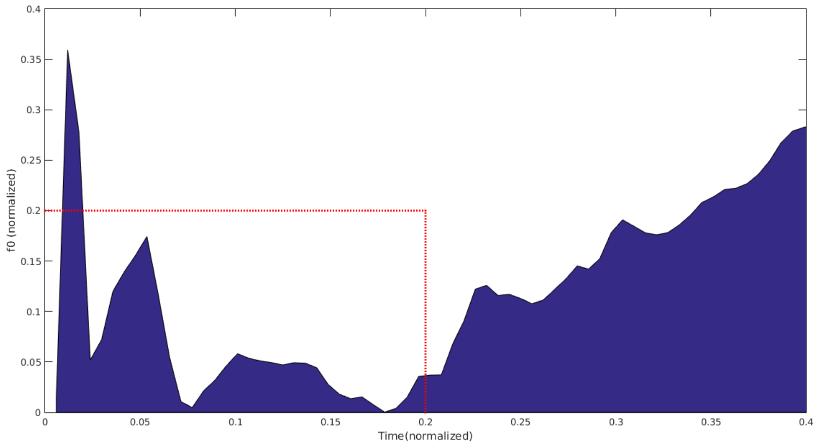


Figure 3.5: Example of a Mew classified as "without initial peak". The dotted red line delimits the threshold area considered .

Afterwards, the task was to analyze the part after the initial peak until the end of the vocalization. In this case, the problem was that even if it is present a slow rise in f0 in almost all the Mew, the evolution of this rise can vary a lot. In some cases, it follows a sort of regular pattern, in other it is very irregular, presenting random valleys and peaks. Hence, we tried to recognize if there were regular patterns repeated along different vocalizations and how they could be described.

Our approach was to confront the part of f0 trajectory comprised in 0.2-0.9 in time normalized, with its polynomial approximations of degree from 1 to 4. Thanks to this, it was easy to locate and discard examples too much irregular for our purpose. In fact, if the difference between the f0 of the vocalization and at least one of its polynomial approximations was under a fixed threshold, this vocalization could be marked as regular. If not, it was considered irregular, then discarded.

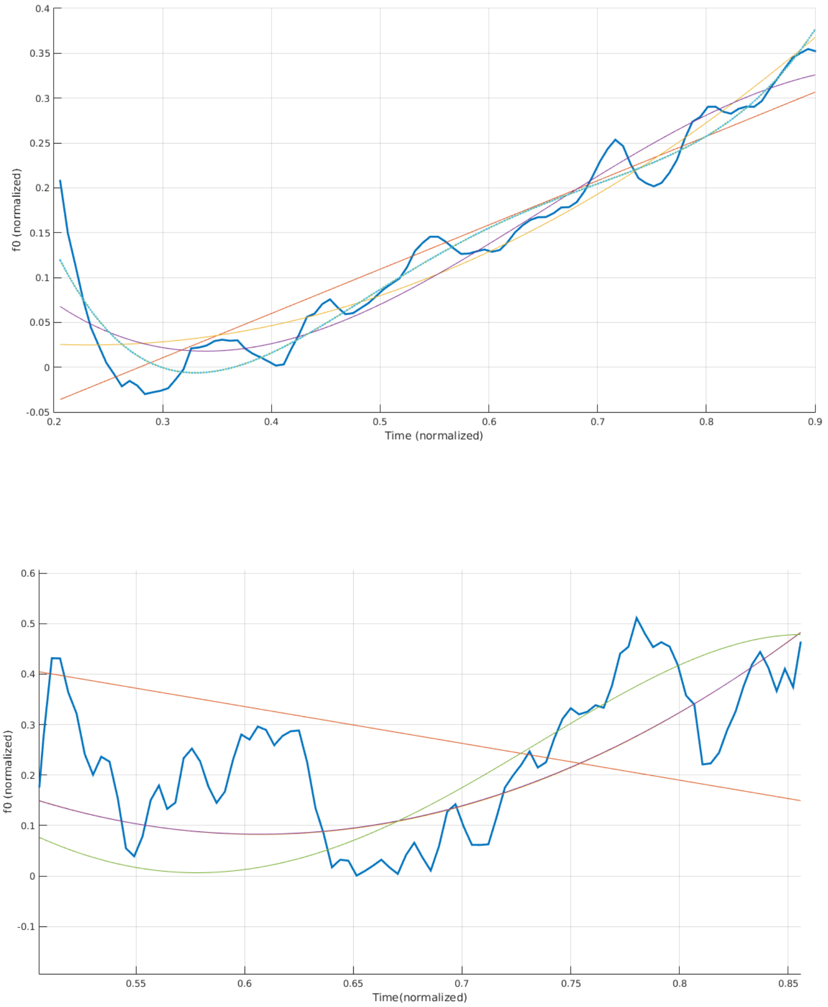


Figure 3.6: Example of Mew classified as regular (top) and irregular (bottom). The blue line represents the f0 estimation, the others are the polynomial approximations.

Once discarding the overly irregular trajectories, another attempt to improve the classification was to separate the regular vocalization basing on the convexity or concavity of the f0 trajectories. However, later we have noticed that this difference was not so perceptually relevant when listening the resulting synthesis, hence we discarded this passage.

In conclusion, achieved this classification we were able to select two sets of vocalization for the training part of HMMs and this led us to re-synthesize faithfully the different behaviors of Mew.

c. Spectral part

The selection and extraction of timbre features, necessary for filtering the excitation signal and then obtaining the synthesis, has not been a crucial point in this work. Observed that the MGCCs were an enough valid spectral representation for our purposes, we used the implementation present in the SPTK toolkit for extracting them.

3.3.2 Growl

The characteristics of Growl are quite different from the ones of Mew, therefore we have decided to continue with this vocalization for trying different approaches in the construction of the acoustic representation. Similar to the growling of dogs, it comprises a long series of low-frequency laryngeal vibrations and wide-band sound frequency striations. Duration is typically from 1 to 4 sec [Pereira et al, 1988].

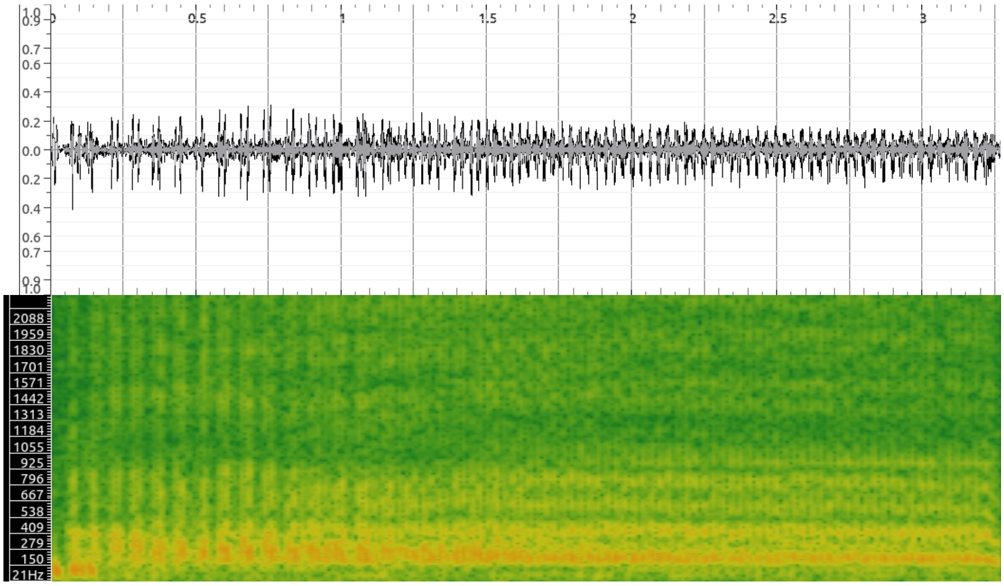


Figure 3.7: Example of Growl vocalization, waveform (top) and spectrogram (bottom)

The main difficulty with the growl samples was that they are highly non-stationary. This means that with a short window of just a few periods of signal it is possible to observe strong amplitude and frequency modulation (AM and FM).

Furthermore, as we mentioned before, there is plenty of noise and ambient effects in the samples. The big advantage of the Growl is that we can observe a clear temporal structure. This led us to analyze it in time-domain rather than in frequency-domain. Thanks to this, it has been quite simple to observe behaviors of amplitude modulations and then designing algorithm to represent

Basing on that, our strategy for having an optimal acoustic representation has been to create in a supervised way the excitation signal, taking into account the different peculiarities. After that, we have filtered it with the timbre features. Hence, our first step has been to locate what features were adequate for replicating the excitation in a realistic way.

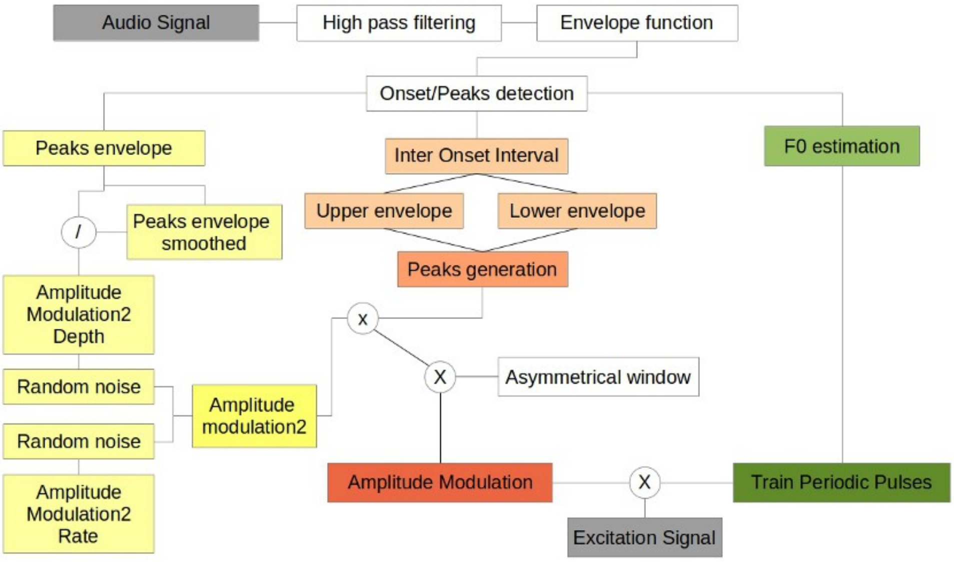


Figure 3.8: Block diagram of the process for creating the Growl excitation signal

After several analysis, for creating our excitation model of Growl, we have reputed satisfactory to consider four features: one for describing the f0 evolution, and three for the amplitude modulation. It is important to underline that the extraction of these features is based on the analysis of the temporal structure of the vocalization. Indeed, in all the four cases, the first step has been to locate the peak of the amplitude of the signal.

With these, we were able to determine how the Growl evolves in time domain and what are the point where we can find more information relative at the vocalization.

a. Peak detection

The temporal structure of Growl, which is easily recognizable by ear, is characterized by a series of low-frequency vibrations rapidly performed. For capturing useful information from the signal, detecting the peak of the amplitude, normalized previously, has been the first step. Feature estimation is more reliable where Signal-to-Noise ratio (SNR) is higher, and then strategy has been estimating features in the location of these peaks and afterwards interpolating the values.

Hence, we built a specific algorithm for detecting this point, basically based on finding possible candidates and confronting with surrounding points. The idea is straightforward: find a local maximum with a value over a fixed threshold, and then comparing it with surrounding minima, for being sure that it is effectively an useful peak. After that, check if there are other peaks too near, and in case find the maximum values.

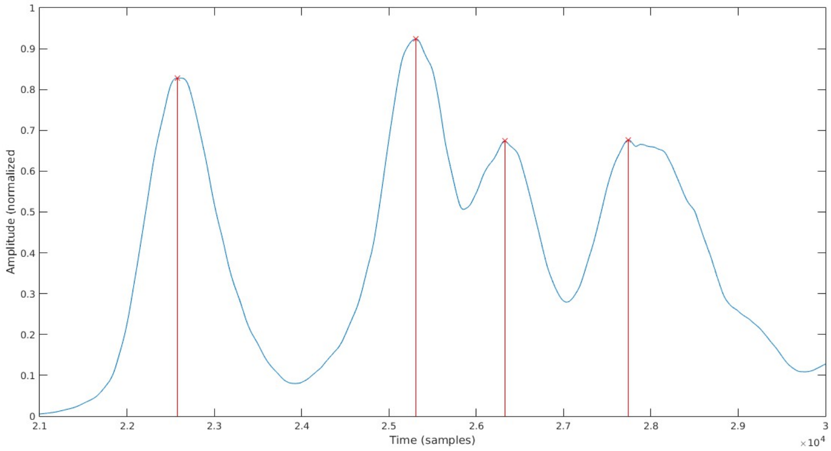


Figure 3.9: Example of detected peaks of the amplitude

Applying an 125 Hz high-pass filter at the selected vocalizations before performing the peaks detection has been useful for improving the performance of this algorithm.

In fact considering the conditions of recording, filtering the audios before to analyze them has helped to eliminate disturbances of various environment noises. As result, it has been available how peaks were distributed in time, and we were able to continue in the process of feature extraction.

b. F0 estimation

It has been necessary to implement a specific procedure for estimating f0 values in the case of Growl vocalizations, considering that attempts with algorithms previous used didn't achieve satisfactory results.

In the beginning, we started analyzing different vocalizations, trying to understand the behavior of f0 trajectories, and trying to locate a reference value valid for all the examples. After several spectral analysis, we have fixed this value at 220 Hz. Afterwards, we have proceeded extracting f0 in the peaks location previously estimated. First, centering a Blackman-Harris window in the position of the peaks, with a windows size equal to four times the period of the reference f0 and then, computing FFT, with an FFT size equal to four times the window size. At this point, having available the spectrum, we have found the maximum point within a range of 200 Hz around our reference value, 220 Hz. This maximum has been selected as f0 values relative at the selected peak.

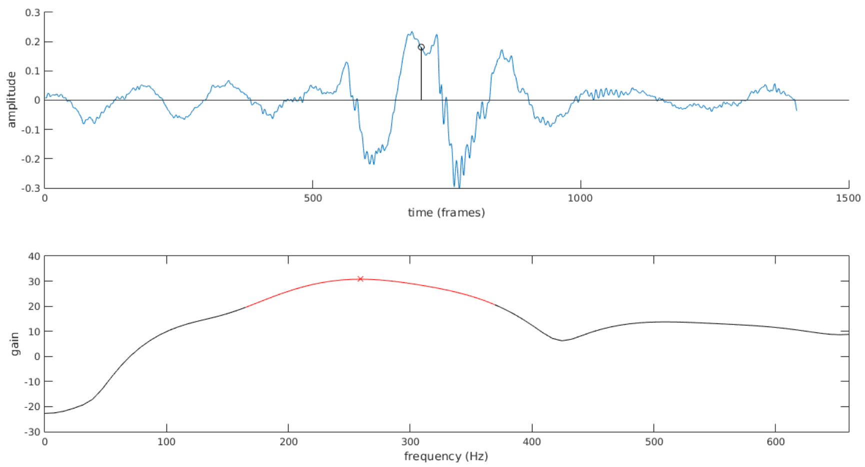


Figure 3.10: Example of windowed signal centered in the peak location, the black stem, (top); computed spectrum with selected candidate as f0, the red cross, in the considered range, highlighted in red (bottom).

Once obtained values for each peak location, we had to interpolate these for obtaining an f0 estimation each ~5ms, achieving a f0 trajectory usable later for training HMMs.

c. Amplitude modulation

Modeling the amplitude modulation of Growl has been a complex process, however it has been fundamental for creating an acoustic representation which could lead us to a good synthesis.

First, we needed to have an idea of general trends of the temporal distribution of the peaks of amplitude, previous estimated within the whole vocalization. Hence, we started analyzing the distance between peak locations. Thanks to this, we have been able to detect the presence of regular or irregular behavior of the distribution of peaks. Indeed, using the mean and standard deviation of this distribution, we have been able to construct an upper and lower envelope, which gave us a sort of measure of regularity. If the distance between these two envelopes was relatively small, it meant that in those points the distance between peaks was enough regular. At the contrary, it meant that there was an irregular behavior of the distribution.

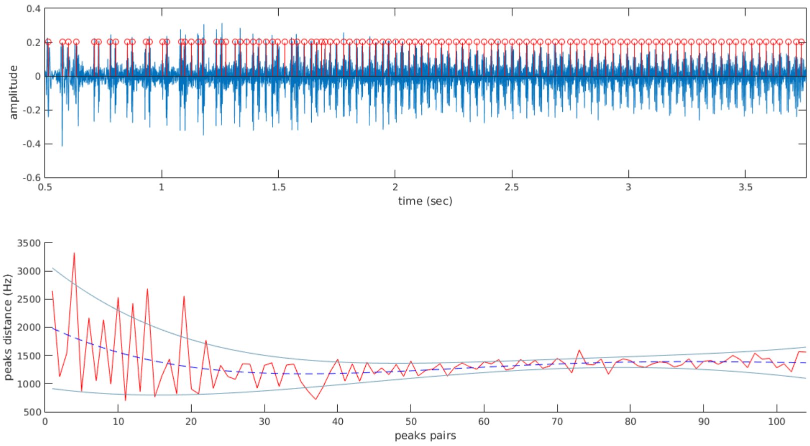


Figure 3.11: Example of Growl waveform, blue line, with peaks detected, red stem (top); plot of trend of peak distance, red line, mean, dashed line, and standard deviation, gray lines (bottom).

This has been fundamental for modeling the locations of pulses were later applying the amplitude modulation. These envelopes have been the first two features which we have selected to model the amplitude modulation, because, thanks to them, we were able to construct peak locations, hence the temporal structure of Growl.

After that, our first attempting for generating amplitude modulation was based on convolving the pulse location obtained before with an Hanning window with a fixed size of 1227 frames, a value deemed valid after several analysis of the shape of the signal. However, the synthesis based on these features was not totally satisfactory. In fact, even if the excitation signal presented a behavior in time similar to the original Growl, there was no presence of substantial variations in the amplitude modulation. This is because in this first step we have achieved to generate the AM pulse locations, but we do not had any control on their amplitude.

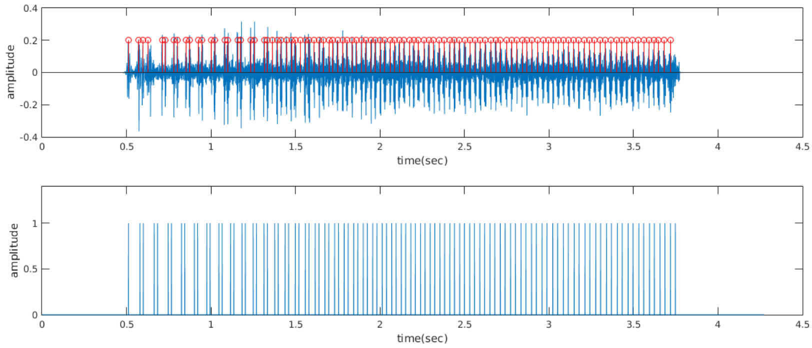


Figure 3.12: Example of Growl waveform, blue line, with peaks detected, red stem (top); pulse location (bottom).

Again, we returned to analyze the normalized amplitude of the signal, searching a way to model the amplitude modulation. Our first observation has been regarding how HMMs are able to model the evolution of the energy along the utterance. Indeed, the output of HMMs is a sort of smoothed version of the amplitude envelope of the vocalization.

This information is generally contained in the first (or 0th) coefficient of the timbre representation. However, computing the amplitude envelope basing on the pulse locations, it was possible to observe a behavior which was clearly different from the hypothetic one deriving from HMMs. Furthermore, with HMMs we will not be able with a few number of states to represent the differences between HMMs derived and real envelope. Since there are relevant variations we decide to model them with two features, depth and rate.

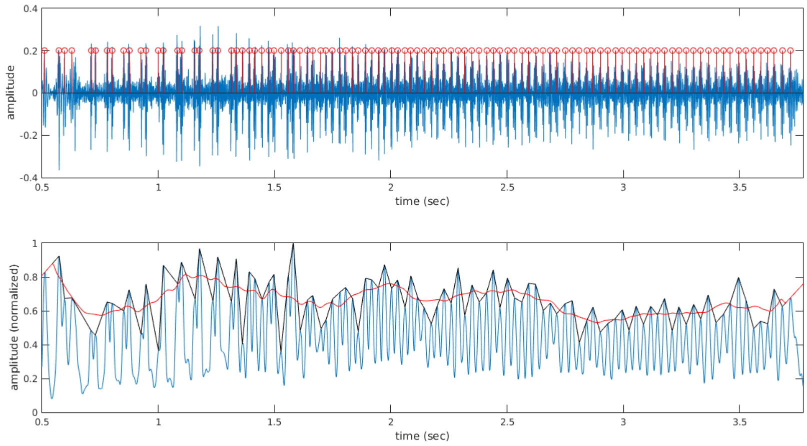


Figure 3.13: Example of Growl waveform, blue line, with peaks detected, red stem (top); normalized amplitude of the signal, blue line, evolution of the energy, red line, pulse amplitude envelope,black line (bottom)

The idea has been that observing some periodicities in the difference signal, it makes sense for modeling it, to consider an oscillator with depth and rate modulation features with some noise. Thus, we obtain another feature for the amplitude modulation, that models actually this difference.

In conclusion, we have modified the pulse onset signal adding variations based on the previous observations, obtaining a signal consisting of deltas with different amplitudes. Then we have convolved it with the pulse shape. In this passage, after further observations, we have chosen to not use an Hanning window, but instead to create an asymmetrical window with a sudden start and a slow decays. For achieving it, we have merged a flat top window and an Hanning window, both with a windows size equal to 1227.

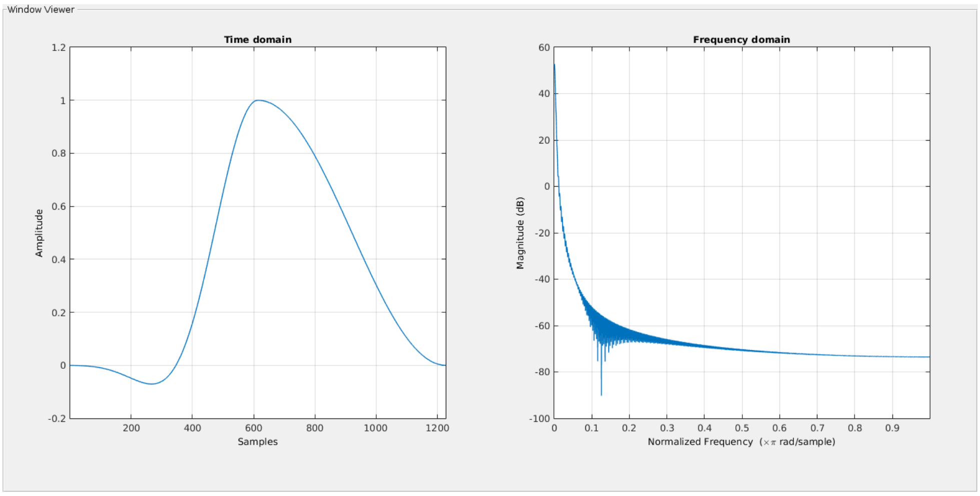


Figure 3.14: Asymmetrical window used for modeling the amplitude modulation

This has been useful also for emphasizing the perception of f0 at the energy modulation rate. The results of the whole process has been the final amplitude modulation. Another way of looking the procedure done for creating the excitation signal is:

- with the analysis of the pulse locations, we have been able to replicate the temporal structure of the Growl, creating a series of uniform deltas.
- with the analysis of the normalized amplitude, we have been able to model for each pulse locations the amplitude, creating a series of non uniform deltas.
- with the convolution with an asymmetrical window, we have been able to replicate the pulse shape, and achieving the final amplitude modulation.

Once available a representation of the amplitude modulation, the last step for creating the excitation signal has been to convolve it with a train of periodic pulses generated with the f0 estimation.

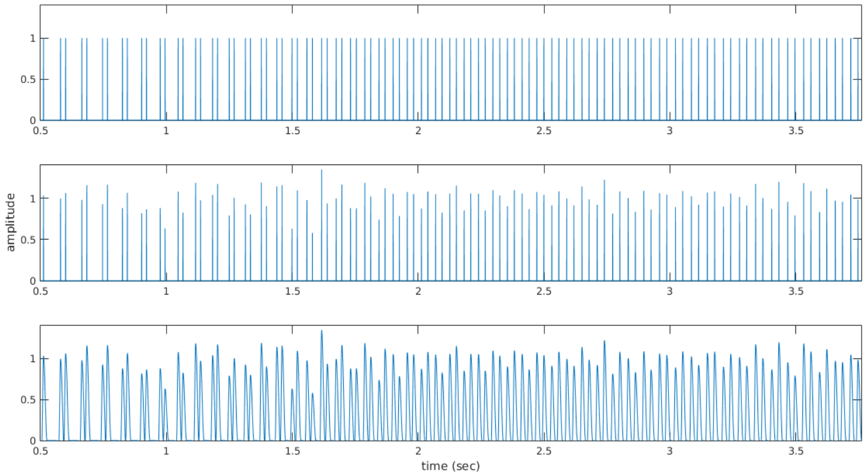


Figure 3.15: Uniform deltas in pulse locations (top); deltas with amplitude variations (center); obtained amplitude modulation (bottom).

d. Spectral part

Similar considerations at the ones done for the Mew have been done for extracting the spectral features. We have not considered particular needs for using a different spectral representation, hence also in this we have used MCG analysis for obtaining timbre features of the vocalization.

However in this case, given the acoustic structure of Growl, we have tried to improve the spectral features. In fact, considering the vocalization as a sequence of pulses, with noise between each one, what we tried to achieve was to catch the timbre information corresponding at the pulses, and discarding others.

Considering the whole MGC representation, extracting the features only in the location of pulses, and finally smoothing it, we have obtained a pulse-centered smoothed version of the timbre representation. Nevertheless, the results achieved have not improved the final synthesis in a way perceptually relevant. At the end, we have chosen to discard this approach.

3.4 Hidden Markov Models framework

The creation of an acoustical representation for the two vocalizations has been the preliminary but fundamental step, previous the introduction of the HMMs framework.

It is important to point out that the choice to use HMMs is mainly derived by two factors. First, since we want to model statistically the feature trajectories, with HMMs we have a measure of likelihood of synthetic trajectories, and this allows us to generate variations of samples with a natural sound and behavior. Second, to learn which context labels are relevant for explaining the variations found among the different vocalizations.

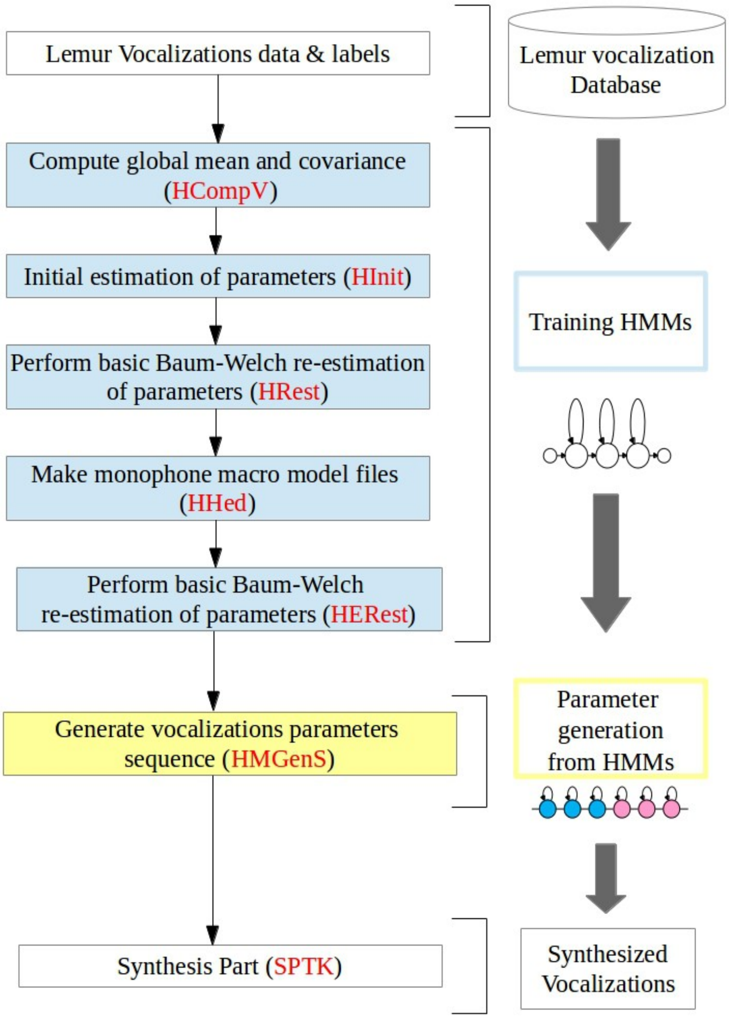
As already mentioned, for testing the validity of our representation we have generated a first synthesis based only on the features extracted. Once the quality was relatively high, we have proceeded to create a second synthesis of a single vocalization, but this time based on HMMs. This second step has been important for observing how the features selected for each vocalization would have behaved. After that, the final passage has been selecting several vocalization to be used in the training dataset for HMMs, training HMMs, and as consequence obtaining a probabilistic model for each vocalization.

In this part, it has been necessary to understand the use of some commands of the toolkits presented before (HTK, HTS and SPTK) and to adapt at our exigences the script of the demo provided by the HTS working group. We can summarize the whole process as following:

- Calculate global mean and covariance of a set of training data (*HCompV*)
- Provide initial estimates for the parameters of a single HMM using a set of observations sequences, repeatedly using Viterbi algorithm (*HInit*)
- Perform basic Baum-Welch re-estimation of the parameters of a single HMM using a set of observations sequence (*HRest*)
- Make monophone macro model files (*Hhed*)

- Perform a single re-estimation of the parameters of a set of HMMs, using an embedded training version of the Baum-Welch algorithm (*HERest*)
- Generate speech parameter sequence (*HMGenS)*
- Synthesize waveform basing on the parameters generated by the previous step (*SPTK*)

where in brackets are indicated the related commands or toolkits used for each specific passage. In the next sections, we will describe step by step the procedure done for achieving synthesis based on HMMs.



Lemur Vocalizations data & labels
Lemur vocalization Database
Compute global mean and covariance (HCompV)
Initial estimation of parameters (HInit)
Training HMMs
Perform basic Baum-Welch re-estimation of parameters (HRest)
Make monophone macro model files (HHed)
Perform basic Baum-Welch re-estimation of parameters (HERest)
Generate vocalizations parameters sequence (HMGenS)
Parameter generation from HMMs
Synthesis Part (SPTK)
Synthesized Vocalizations

Figure 3.16: Block diagram of the process done with HTK/HTS/SPTK toolkits

3.4.1 Data Preparation

Difficulties have arisen when dealing with HMMs, in particular in the process of adaptation of data with the used script. In fact, the demo is thought for working with human speech, and a series of parts do not match with our purpose.

The most relevant difference is features extraction of training data, f0 and MGC, which is automatic in the demo, while we had to use our specific f0 estimation, both with Mew and Growl, and with the second, also adding amplitude modulation features. Furthermore, speech audios are automatic labeled, which means that duration of each phoneme is automatically recognized and information is saved for being used later. Given that we have dealt with Lemur vocalizations, this automation cannot be used, hence we have manually annotated each label. An example of a label of Mew is:

```
       0      3963816  pau
 3963816     12287829  mew
12287829     16251645  pau
```

Where the first two numbers denote the start and the end in time of the vocalization, 'pau' denote the pause or silence, hence no presence of signal, and 'mew' is the kind of vocalization, referred to the audio relative at the label.

While labeling the data, we had to choose how to manage vocalizations. The simpler choice was to consider each vocalization has a single phoneme, and it was the approach that we chose. Mainly it was because even if considering a vocalization as a series of "phonemes" probably can lead to better synthesis, we have chosen to start with a basic approach. We have done some test trying to separate a vocalization into different units, as for example in the Mew considering the initial peak as a phoneme and the rest of the vocalization as another phoneme. However, our decision has been to remain with the model "vocalization=phoneme", leaving as future work different kind of labeling approaches, as we will discuss later.

Last step for preparing the data was to convert to raw the audio of the database, originally in aiff format, and to change the sample rate at 48000 Hz. For this passage we have used the SoX toolkit.

3.4.2 Training Part

Once having data ready to be used, we have continued setting the training part of your system. Basically, we have not modified the original script in this part, because for our aim it was enough to exclude some parts related at high-level context which we have not taken into account during this study. Indeed, our process of analysis and synthesis was focused on each case separately, hence we have not used any kind of relationship between vocalizations and a general context where they are used. Surely, it has been a limitation both for the improvement of the synthesis quality and considering that one point of strength of HMMs is to be able to manage different level of relationships between phonemes, syllables, words, phrases etc. This is possible because of the syntax is well defined in human languages, but with Lemurs, and in general animals, the approach cannot be exactly the same. Also here, we will discuss in future work different ones which could be adopted. Considering the previous observations, we have used default values of the original script for performing the training part. The only change has been done for the Growl, because of the presence of addition features for modeling the amplitude modulation.

3.5 Synthesis

The synthesis is the last step of the whole process described in the previous sections. However, several attempts have been done during work, which in most of the cases have led at results with a poor quality. At the beginning, the lack of correct acoustical representations has been the cause of synthesis which did not replicate the original vocalizations faithfully. Afterwards, we had to deal with the difficulty on finding valid examples to be used in the training part. Furthermore, the number of vocalizations with an enough quality for our purposes, has been not so massive.

However, working on three different kind of synthesis, we had the possibility to evaluate results separately, taking into account in each case the constrictions below which we had to work. As already mentioned before, we have achieved three types of synthesis:

1) Synthesis of one example of vocalization not using HMMs, which probably has been the most important. Thanks to this, we have obtained an immediate response of the validity of the acoustic representation and of the quality of the features extracted for the vocalization.
2) Synthesis of one example of vocalization using HMMs. In this passage, we have started to have an idea of how synthesis behaved using HMMs, and as consequence, where we started to have a decrease in naturalness.
3) Synthesis of several examples of vocalization using HMMs. In this last type, the whole process has been as the one used in speech-context. We have been able to create a model for each vocalization, based on probabilistic analysis of the several examples used in the training part. Specifically, for the Mew we have used 10 examples in the case with initial peak and 25 in the case without initial peak. While for the Growl we have found only 6 vocalizations which had enough quality to be used.

3.5.1 Post-Processing

The introduction of Hidden Markov Models in the process of synthesis immediately has lead to a notable decrease of the naturalness of results. Several reasons could be considered: lack of useful examples, validity of choices of feature extraction, analysis and synthesis parameters, approaches used, etc. All the issues treated in the previous sections have influenced partially the final results.

Because of that, before to perform the evaluation it has been necessary to post-process the synthesis. As we will see in the next chapter, the evaluation has been based on comparison between original and synthesized vocalizations, hence having a synthesis that clearly sounded artificial would not led to interesting results.

However, only in the case of the second type of synthesis, the one of a single vocalization based on HMMs, it has been performed this further step of post-processing. Regarding f0, it has been enough to add noise at a fixed percentage of frames, about 10%, uniformly distributed in time, and afterwards smoothing the result.

In the case of MGC, the idea has been to take the feature of original vocalization and the one generated by HMMs, and to do statistical analysis of the several differences. Once having an idea of how much differed the behaviors, automatic variations have been generated and then applied at the features extracted with HMMs. In details, first the difference between original MGC and HMM derived MGC has been computed, frame by frame. After that, using the mean of the difference values for each frame, noise has been generated, and variations has been computed in the same way of f0 estimation. In conclusion, we have used this modified features for generating the synthesis.

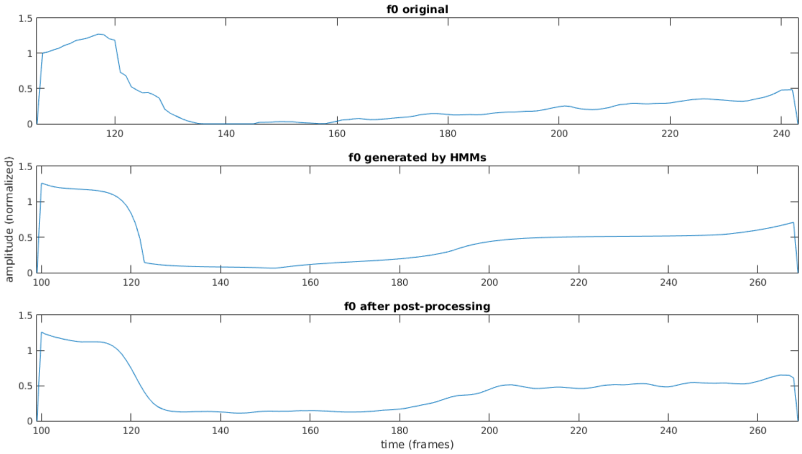


Figure 3.17: Plots of f0 initial estimation (top), f0 generated by HMMs (center), final f0 after post-processing (bottom).

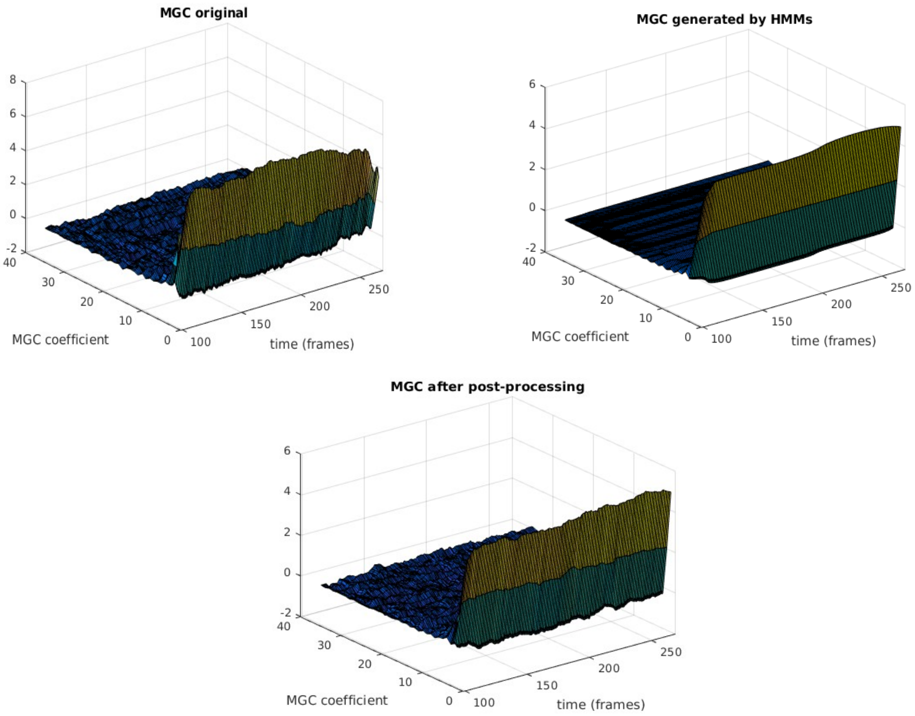


Figure 3.18: Plots of MGC initial estimation (top), MGC generated by HMMs (center), final MGC after post-processing (bottom).

In this way, it has been possible to have features based on HMMs, but with slightly variations based on statistical analysis, choice which has enhanced enough the quality of the synthesis.

The last step necessary before to start the evaluation has been to add background at synthesis. Several fragments have been selected within the database, and it has enhanced the naturalness of the artificial signals. Indeed, in the synthesis generated obviously there is no presence of noises and echoes derived from the environment. Hence, we have due recreate these ambient effects before to perform the evaluation.

Chapter 4
EVALUATION AND RESULTS

The final part of this work has been to evaluate the outcomes of the project by means of listening tests with humans, which have been performed via a web-based form. In the next section we will introduce the general schema of the evaluation process and several characteristics of the participant population and the conditions during the experiment. Afterwards, we will examine the results of each vocalization separately.

4.1 Evaluation

4.1.1 Listening test

While designing the experiment for evaluation of the synthesis, several issues have been taken into account.

The first consideration has been that probably most of the participants would not had a background in the field of Bioacoustics or Signal Processing. Considering that it has not been selected a specific population for performing this test, the variety of participants has been previously hypothesized, and later it has been confirmed. In addition, even if with some experience in fields related to this work, it is not very likely to encounter persons which have already listened lemurs vocalization during their life. Furthermore, if we consider that in our case vocalizations are treated out-of-context, as single signals, to be familiar with this kind of sounds becomes very unlikely.

Consequently, the first step has been to present several examples of the original vocalizations. Thanks to this, the participants had the possibility to have clear acoustic references on how these vocalizations are structured, which are the main peculiarities, significant variations etc. After that, the core of the test has been consisting of two parts, which have been presented randomly in each test for avoiding possible bias.

First, it has been presented an A/B testing, where the aim was to compare two audios and to recognize the nature of each one, hence if they were original or synthesized versions. For each kind of vocalization, six couples of audios have been presented, where there were four possibilities: or the first audio were original and the second synthesized, or the contrary, or both original, or both synthesized. In this section, the first and the second type of synthesis have been evaluated, those based on single vocalization. Moreover, this part has been also useful for having a criteria of judgment of the reliability of the decisions taken by the participant. Indeed, apart from recognizing the ability in differentiating between original and not, analyzing the results, it has been possible to know if the participant was able to recognize if the same audio was repeated two times.

Second, it has been required to identify, within a set of audios, which ones were synthesized vocalizations and which ones were original. This part has been used for evaluating the synthesis created by using Hidden Markov Models with several examples in the training part, hence the third type of synthesis. Performing and A/B testing did not make sense in this case, considering that the audios artificially generated were based on the analysis of several examples used together. As before, this part it has been useful for observing if the participants were able to recognize the original vocalizations, hence for having an idea of their ability.

4.1.2 Population participant

The web-form created for performing the listening test has been open without any kind of restrictions, hence it has meant that we have not performed any kind of selection within the participants. The population of this experiment has been composed by 22 participants, mixed gender, with a range of age between 22 and 29 years with some exceptions over the 30. Furthermore, more than the 85% never had experiences in the fields of Biology, Bioacoustics, Signal Processing and Audio Engineering. In addition, the conditions under which the test have been done are various: about 50% have used low-quality speakers while the rest have used headphones.

These aspects surely have influenced in part the results of the experiment, hence it is fundamental to take them into account analyzing the responses.

4.2 Results

The analysis of the responses of listening test confirmed several considerations already taken into account at the end of the synthesis process, but at the same time it has pointed out various relationship between vocalizations and their artificial version.

The A/B testing gave us an evaluation of the first two synthesis, both derived from a single vocalization. A first important consideration is that more the users have been able to recognize two identical vocalizations within the couple, more they were able to recognize between original and synthesized vocalizations. It is a validation that responses have not been totally random. As a consequence, we can also affirm that the synthesis of Mew without initial peak is the one which is less recognizable within the three computed, while the Growl has been clearly perceived as artificial when compared to the original. Also, it is interesting that in the case of Mew, the synthesis generated without using HMM has been more difficult to recognize. This fact is not completely unexpected, on the contrary it has confirmed that in some cases using HMMs has not improved the naturalness of the synthesis.

	O – S1	**O – S2**	**S1 – S2**	**O – O**	**S1 – S1**	**S2 – S2**
Growl	59,10%	50.00%	36,40%	68.20%	54.50%	40.90%
Mew N.P.	27.30%	45.50%	45.50%	27.30%	27.30%	54.50%
Mew W.P.	45.50%	59,10%	31.80%	40.90%	40.90%	36,40%

Table 4.1: Percentage of correct responses of the A/B testing. Legend: O = original vocalization. S1 = synthesis generated without using HMMs. S2 = synthesis generated using HMMs. Mew N.P. = Mew without initial peak. Mew W.P. = mew with initial peak.

Regarding the other part of the test, the identification task, it has helped to evaluate the level of naturalness of the synthesis generated using several vocalizations together as training dataset for HMMs. In this case, the percentage of correct identifications of original examples has not been so high, between 60% and 70%. At the contrary of the previous results, it is interesting to notice that in this case Growl is the vocalization better synthesized, while Mew has been detected as artificial in more than 18 cases over 22, both with and without initial peak.

	Original	HMM-based
Growl	60.23%	50.00%
Mew (N.P.)	68.18%	81.82%
Mew (W.P.)	67.04%	81.82%

Table 4.2: Percentage of correct responses of the identification testing.

Chapter 5
CONCLUSIONS AND FUTURE WORK

The work presented in this thesis is nothing but a first little step into a new way for creating artificially animal vocalizations. Several directions could be taken both in the specific case of Lemurs and, in general, in animal context. The aim of this research is to show that, even if there are not specific information about the morphology and the functional anatomy of the vocal tract, it is possible to model and then synthesize animal vocalizations. Indeed, all the results have been obtained mainly thanks to signal processing techniques. Obviously, for achieving high-quality synthesis it is fundamental a collaboration between biologists and engineers, because apart from having a good modelization of the signals, it is necessary to have a profound knowledge of the "semantic part" concerning the animal communication.

5.1 Contributions

The main contribution of this work is to have created an acoustic model of two vocalizations within the Lemur repertoire, the Mew and the Growl, only based on signal analysis. Starting from several acoustic features, we have synthesized these vocalizations, achieving three types of synthesis: one based only on the feature extracted, one based on HMMs but created starting from a single vocalization, and then a last one where HMMs have been used as normally in the speech-context, with due adaptations.

Apart from the obtained results, we have introduced in the context of animal vocalization a different approach for achieving high-quality synthesis, which is widely used in the field of speech synthesis. Surely, other approaches could be considered, but we are confident that Hidden Markov Models could potentially lead to optimum results.

5.2 Future work

Considering this work as a first attempt of using probabilistic based models for the creation of animal vocalization synthesis, directions that could be taken are various. A list of possible future work, focusing on Lemurs vocalization, could be the following:

- Try different labeling for each vocalization while using HMMs. In this passage, it is fundamental to analyze in deep each vocalization, both from a signal processing and a biological point of view, for locating basic units of each vocalizations. In this manner, more precise characterization could be done, leading to create more specific models in each case, improving the quality of the synthesis. Indeed, one of the basic problem is to identify what could be considered as a phoneme within a vocalization, if it has sense to talk about phonemes in the animal oral communication and issues related. Even if mainly it can be treated from a biological point of view, with the help of signal processing techniques, it could be easier to achieve better results.

- Extend the kind of analysis done in this work at all the vocalizations within the Lemur vocalization repertoire, leading to have a specific acoustic representation in each case. In addition, it could be try to use the ones presented for the Mew and the Growl with other vocalizations, observing if they can be adapted.

- Once having specific synthesis for all the repertoire, try to add more information of the high-level context into the Hidden Markov Models. In fact, one of the point of strength of HMMs is that it can use context-related information for modeling the sound, improving the quality of the synthesis generated. Also in this part, it is necessary a work both in the field of biology and in signal processing. Analyzing how vocalizations are used, the correlation between each other, repetitions and other kind of relationship, important information could be extracted which could improve the final results.

- Done all the steps before, a final experiment could be to try to record enough vocalizations of a specific specimen and built a dependent-speaker model. Indeed, using vocalizations of exemplars with different sex, age, size etc. adds non-sense variations into the probabilistic model, which could be misleading.

However, it is important to underline that the approach which has been adopted in this thesis with Lemurs, could be extended, at least, at all the animals which sound production can be represented with the source-filter model. Obviously, in each case the species-specific acoustic representations of vocalization within the repertoire have to be adapted and customized.

Hopefully, in the future approaches where are mixed together biological researches and audio engineering techniques could lead to improve notably results, helping in the discovery of new knowledges, fundamental in understanding animal behaviors and improving human-animal interactions.

Bibliography

[Adam, 2006] O. Adam, “Advantages of the Hilbert Huang transform for marine mammals signals analysis.,” *J. Acoust. Soc. Am.*, vol. 120, no. 5 Pt 1, pp. 2965–73, Nov. 2006.

[Airaksinen, 2012] M. Airaksinen, “analysis/synthesis comparison of vocoders utilizyed in statistical parametric speech synthesis,” Master Thesis,2012.

[Amador et al, 2013] A. Amador, Y. S. Perl, G. B. Mindlin, and D. Margoliash, “Elemental gesture dynamics are encoded by song premotor cortical neurons.,” *Nature*, vol. 495, no. 7439, pp. 59–64, Mar. 2013.

[Arnold and Zuberbühler, 2006] K. Arnold and K. Zuberbühler, “Language evolution: semantic combinations in primate calls.,” *Nature*, vol. 441, no. 7091, p. 303, May 2006.

[Babu and Rao, 2011] G. Babu and R. Rao, “Enhancement of animal vocalization using various algorithms,” *Int. J. Eng. Sci.,* vol. 3, no. 3, pp. 2298–2307, 2011.

[Baker-Médard et al, 2013] M. S. A. Baker-Médard, M. C. Baker, and D. M. Logue, “Chorus Song of the Indri (Indri indri: Primates, Lemuridae): Group Differences and Analysis of Within-group Vocal Interactions,” *Int. J. Comp. Psychol.*, vol. 26, pp. 241–255, 2013.

[Baker and Logue, 2003] M. C. Baker and D. M. Logue, “Population differentiation in a complex bird sound: A comparison of three bioacoustical analysis procedures,” *Ethology*, vol. 109, pp. 223–242, 2003.

[Bardeli et al, 2010] R. Bardeli, D. Wolff, F. Kurth, M. Koch, K.-H. Tauchert, and K.-H. Frommolt, “Detecting bird sounds in a complex acoustic environment and application to bioacoustic monitoring,” *Pattern Recognit. Lett.*, vol. 31, no. 12, pp. 1524–1534, Sep. 2010.

[Beecher, 1988] M. D. Beecher, “Spectrographic Analysis of Animal Vocalizations: Implications of the ‘Uncertainty Principle,’” *Bioacoustics*, vol. 1, no. 2–3, pp. 187–208, 1988.

[Berenzweig et al, 2003] A. Berenzweig, D. Ellis, B. Logan, and B. Whitman, “A Large Scale Evaluation of Acoustic and Subjective Music Similarity Measures,” in *Proc. of International Symposium on Music Information Retrieval*, 2003.

[Boucher et al, 2012] N. Boucher, M. Jinnai, and A. Smolders, “Improvements in an automatic sound recognition system using multiple parameters to permit recognition with noisy and complex signals such as the dawn chorus,” *Acoust. 2012 Nantes*, no. April, pp. 2447–2453, 2012.

[Camacho, 2007] A. Camacho, “SWIPE: A Sawtooth Waveform Inspired Pitch Estimator,” in *The Journal of the Acoustical Society of America,* no. 124, vol. 3, pp 1638-52, October 2008.

[Chung, 1967] K. Chung, Markov chains With Stationary Transition Probabilities, Springer US, 1967.

[Fant, 1981] G. Fant, “The source filter concept in voice production(Quarterly Progress and Status Report),” *Stl-Qpsr*, vol. 1, pp. 21–37, 1981.

[Fitch, 2000a] W. T. Fitch, “The phonetic potential of nonhuman vocal tracts: comparative cineradiographic observations of vocalizing animals.,” *Phonetica*, vol. 57, pp. 205–218, 2000.

[Fitch, 2001] W. T. Fitch and D. Reby, “The descended larynx is not uniquely human.,” *Proc. Biol. Sci.*, vol. 268, no. 1477, pp. 1669–1675, 2001.

[Fitch, 2000b] W. T. Fitch, “The evolution of speech: a comparative review,” vol. 6613,*Trends Cogn Sci* ,no. 4, vol 7, pp. 258–267, 2000.

[Fitch and Hauser, 2004] W. T. Fitch and M. D. Hauser, “Computational Constraints on Syntactic Processing in a Nonhuman Primate,” *Science (80-.).*, vol. 303, no. January, pp. 348–351, 2004.

[Fitch et al, 2002] W. T. Fitch, J. Neubauer, and H. Herzel, “Calls out of chaos: the adaptive significance of nonlinear phenomena in mammalian vocal production,” *Anim. Behav.*, vol. 63, no. 3, pp. 407–418, Mar. 2002.

[Fitch, 2006] W. Fitch, “Production of Vocalizations in Mammals,” *Encycl. Lang. Linguist.*, no. 1972, pp. 115–121, 2006.

[Gamba and Giacoma, 2007] M. Gamba and C. Giacoma, “Quantitative acoustic analysis of the vocal repertoire of the crowned lemur,” *Ethol. Ecol. Evol.*, vol. 19, no. 4, pp. 323–343, Oct. 2007.

[Gamba, 2014] M. Gamba, “Vocal tract-related cues across human and nonhuman signals,” *Reti, saperi, linguaggi*, pp. 49–66, 2014.

[Gamba and Giacoma, 2003] M. Gamba and C. Giacoma, “Monitoring the vocal behaviour of ruffed lemurs in nest-box,” *EAZA News*, no. 43, pp. 28–30, 2003.

[Gamba and Giacoma, 2010] M. Gamba and C. Giacoma, “Key issues in the study of primate acoustic signals , an update,” *J. Anthropol. Sci.*, vol. 88, pp. 215–220, 2010.

[Gamba and Giacoma, 2006] M. Gamba and C. Giacoma, “Vocal tract modeling in a prosimian primate: the black and white ruffed lemur,” *Acta Acust. united with Acust.*, vol. 92, pp. 749–755, 2006.

[Geissmann and Mutschler, 2006] T. Geissmann and T. Mutschler, “Diurnal distribution of loud calls in sympatric wild indris (indri indri) and ruffed lemurs (varecia variegata): Implications for call functions,” *Primates*, vol. 47, no. 4, pp. 393–396, 2006.

[Gómez et al, 2013] E. Gómez, J. Bonada, and G. Emilia, “Towards Computer-Assisted Flamenco Transcription : An Experimental Comparison of Automatic Transcription Algorithms as Applied to A Cappella Singing Towards Computer-Assisted Flamenco Transcription : An Experimental Comparison of Automatic Transcription A,” *Comput. Music J.*, vol. 37, pp. 73–90, 2013.

[Graugés, 2014] A. Graugés, “Animal Vocalization Analysis and Synthesis,” *Master Thesis*, 2014.

[Juang and Rabiner, 1986] B. H. Juang and L. R. Rabiner, “An introduction to Hidden Markov Models.” 1986.

[Jun et al, 2010] H.-S. Jun, P. K. Dhar, C.-H. Kim, and J.-M. Kim, “Baleen Whale Sound Synthesis using a Modified Spectral Modeling,” *KIPS Trans.*, vol. 17B, no. 1, pp. 69–78, 2010.

[Kershenbaum et al, 2014] A. Kershenbaum, D. T. Blumstein, M. a. Roch, Ç. Akçay, G. Backus, M. a. Bee, K. Bohn, Y. Cao, G. Carter, C. Cäsar, M. Coen, S. L. DeRuiter, L. Doyle, S. Edelman, R. Ferrer-i-Cancho, T. M. Freeberg, E. C. Garland, M. Gustison, H. E. Harley, C. Huetz, M. Hughes, J. Hyland Bruno, A. Ilany, D. Z. Jin, M. Johnson, C. Ju, J. Karnowski, B. Lohr, M. B. Manser, B. McCowan, E. Mercado, P. M. Narins, A. Piel, M. Rice, R. Salmi, K. Sasahara, L. Sayigh, Y. Shiu, C. Taylor, E. E. Vallejo, S. Waller, and V. Zamora-Gutierrez, “Acoustic sequences in non-human animals: a tutorial review and prospectus,” *Biol. Rev.*, p. n/a–n/a, 2014.

[King et al, 2010] L. E. King, J. Soltis, I. Douglas-Hamilton, A. Savage, and F. Vollrath, “Bee threat elicits alarm call in African elephants,” *PLoS One*, vol. 5, no. 4, 2010.

[Lieberman, 1968] P. Lieberman, “Primate vocalizations and human linguistic ability.,” *The Journal of the Acoustical Society of America*, vol. 44. pp. 1574–1584, 1968.

[Maretti et al, 2010] G. Maretti, V. Sorrentino, A. Finomana, M. Gamba, and C. Giacoma, “Not just a pretty song: an overview of the vocal repertoire of Indri indri.,” *J. Anthropol. Sci.*, vol. 88, pp. 151–65, Jan. 2010.

[Martino, 2000] R. Martino, “Synthasaurus : An Animal Vocalization Synthesizer,” Master Thesis, 2000.

[McCowan et al, 1999] B. McCowan, S. Hanser, and L. Doyle, “Quantitative tools for comparing animal communication systems: information theory applied to bottlenose dolphin whistle repertoires.,” *Anim. Behav.*, vol. 57, no. 2, pp. 409–419, 1999.

[Mcgregor, 2000] P. K. Mcgregor, “Playback experiments: design and analysis,” pp. 3–8, 2000.

[Mitan, 1992] J. C. Mitan, H. Julie, P. Marler, and R. Byrne, “Dialects in Wild Chimpanzees ?,” vol. 243, pp. 233–243, 1992.

[Nadhurou et al, 2015] B. Nadhurou, M. Gamba, N. V. Andriaholinirina, a. Ouledi, and C. Giacoma, “The vocal communication of the mongoose lemur (*Eulemur mongoz*): phonation mechanisms, acoustic features and quantitative analysis,” *Ethol. Ecol. Evol.*, no. June, pp. 1–20, 2015.

[Ness, 2013] S. Ness, “The Orchive : A system for semi-automatic annotation and analysis of a large collection of bioacoustic recordings,”, PhD Thesis, 2013.

[Pereira et al, 1988] M. E. Pereira, M. L. Seeligson, and J. M. Macedonia, “The behavioral repertoire of the black-and-white ruffed lemur, Varecia variegata variegata (Primates: Lemuridae).,” *Folia Primatol. (Basel).*, vol. 51, no. 1, pp. 1–32, 1988.

[Rainey et al, 2004] H. J. Rainey, K. Zuberbühler, and P. J. B. Slater, “Hornbills can distinguish between primate alarm calls.,” *Proc. Biol. Sci.*, vol. 271, no. 1540, pp. 755–759, 2004.

[Reby et al, 2006] D. Reby, R. André-Obrecht, A. Galinier, J. Farinas, and B. Cargnelutti, “Cepstral coefficients and hidden Markov models reveal idiosyncratic voice characteristics in red deer (Cervus elaphus) stags.,” *J. Acoust. Soc. Am.*, vol. 120, no. 6, pp. 4080–4089, 2006.

[Ren et al, 2009] Y. Ren, M. T. Johnson, P. J. Clemins, M. Darre, S. S. Glaeser, T. S. Osiejuk, and E. Out-Nyarko, “A framework for bioacoustic vocalization analysis using hidden Markov models,” *Algorithms*, vol. 2, pp. 1410–1428, 2009.

[Stowell et al, 2013] D. Stowell, S. Muševič, J. Bonada, and M. D. Plumbley, “Improved multiple birdsong tracking with distribution derivative method and Markov renewal process clustering,” Feb. 2013.

[Suzuki, 2005] R. Suzuki, J. R. Buck, and P. L. Tyack, “The use of Zipf’s law in animal communication analysis,” *Anim. Behav.*, vol. 69, no. 1, pp. 9–17, 2005.

[Talkin et al, 1995] D. Talkin, W. B. Kleijn, and K. K. Paliwal, “A Robust Algorithm for Pitch Tracking(RAPT),” *Speech Coding and Synthesis, The Eds. Ams-terdam, Netherlands:Elsevier*. pp. 495–518, 1995.

[Tao, 2009] J. Tao, “Acoustic model adaptation for automatic speech recognition and animal vocalization classification,” *Word J. Int. Linguist. Assoc.*, 2009.

[ten Cate and Okanoya, 2012] C. ten Cate and K. Okanoya, “Revisiting the syntactic abilities of non-human animals: natural vocalizations and artificial grammar learning.,” *Philos. Trans. R. Soc. Lond. B. Biol. Sci.*, vol. 367, no. 1598, pp. 1984–94, Jul. 2012.

[Terry, 2005] A. M. R. Terry, T. M. Peake, and P. K. McGregor, “The role of vocal individuality in conservation.,” *Front. Zool.*, vol. 2, no. 1, p. 10, 2005.

[Tokuda et al, 2002] I. Tokuda, T. Riede, J. Neubauer, M. J. Owren, and H. Herzel, “Nonlinear analysis of irregular animal vocalizations.,” *J. Acoust. Soc. Am.*, vol. 111, no. 6, pp. 2908–2919, 2002.

[Tokuda et al, 1994] K. Tokuda, T. Kobayashi, T. Masuko, and S. Imai, “Mel-Generalized Cepstral Analysis - a Unified Approach To Speech Spectral Stimation,” *Proc. Int. Conf. Spok. Lang. Process.*, no. 2, pp. 1043–1046, 1994.

[Tokuda et al, 2013] K. Tokuda, Y. Nankaku, T. Toda, H. Zen, J. Yamagishi, and K. Oura, “Speech Synthesis Based on Hidden Markov Models,” *Proc. IEEE*, vol. 101, no. 5, pp. 1234–1252, May 2013.

[Young et al, 2006] S. J. Young, G. Evermann, M. J. F. Gales, et al., “The HTK Book”, version 3.4 , 2006

[Wang and Tao, 2015] Y. Wang and J. Tao, “Implementation of Parameter Generation in HMGenS,” pp. 1–24, 2015.

[Watkins, 1968] W. A. Watkins, “The harmonic interval: fact or artifact in spectral analysis of pulse trains,”, Marine Bioacoustic, vol. 2, Pergamon Press, New York, 1967.

[Zen et al, 2009] H. Zen, K. Oura, T. Nose, J. Yamagishi, and S. Sako, “Recent development of the HMM-based speech synthesis system (HTS),” *SSW*, pp. 294–299, 2009.