The Raymond and
Beverly Sackler Faculty
of Exact Sciences
Tel Aviv University

Department of Applied Mathematics

Algorithms for Automatic Segmentation of Trill Vocalizations in Birds

A thesis

submitted for the degree of
Master of Science
by

Hagai Barmatz

Under supervision of

Prof. Yizhar Lavner and Prof. Sivan Toledo

Tel Aviv University
March 2019

Contents

Chapter 1

Introduction

1.1 Background

Animal communication and specifically acoustic communication is the subject of ongoing ecological and biological research [1, 2, 3]. Acoustic communication was found to be a significant factor in facilitating reproductive isolation and thus selection processes in a variety of animal species [4]. In addition, vocal communication in animals is used as a research model for understanding the development and evolution of language and grammar [5] and as a model for social learning in general [6]. The study of animal acoustic communication thus enables better and thorough understanding of evolutionary, ecological and behavioral processes.

Further, with the rapid improvement of technological capabilities, digital sensors have become cheaper. Digital microphones and recording equipment is used in wildlife conservation areas to monitor animal activity and population as well as the study of complex behaviour such as migration and predation [7, 8].

However, despite this ongoing progress, automatic and reliable detection and segmentation of bioacoustic events in audio recordings remains a challenging research area.

In recent years, the fields of bioacoustics and bird acoustics have been receiving growing attention from the acoustic signal processing community. Massive accumulation of biological and ecological datasets have led to a pressing need in activities such as clustering and annotation. Manual annotation of datasets is both tedious and time consuming, but also error prone. Therefore, reliable automatic algorithms are required. To facilitate development, the biological community kindly created and shared labeled datasets, and competitions of bioacoustic data classification are held annually [9, 10, 11].

Convolutional neural networks were utilized to produce results which won several competitions, but the problem was mainly treated as an image classification problem, by transforming the audio signal with a short-time Fourier transform to produce a spectrogram. The spectrogram is a 2 dimensional signal representation, which may be either further

processed using image processing techniques or fed to the neural network classifier "as is". Other audio signal processing techniques can be used to expose important or discriminative features, or transform the data to make the detection/classification more successful.

A main disadvantage of using the spectrogram images as a feature space is the large dimension of the data. Ideally, a low dimensional manifold may be found in which most of the variance is concentrated. Compact representation of bioacoustic data should therefore increase classifier efficiency. This thesis is focused on developing reliable and efficient algorithms for detection, segmentation and parameter estimation of bird vocalizations. The estimated parameters can form a compact data representation which may be fed into a machine learning classifier such as a CNN.

The research conducted focuses on a specific animal vocalization type referred to as a *trill*. Trills are prevalent in many bird species but also in mammals and other animals. Trills are characterized by pulsatile modulations in sustained expiration, or by rapid alternations between expiration and inspiration. They consist of rapid repetitions of a basic acoustic unit, referred to as a *syllable*.

Though by definition not complex, trill physical attributes such as trill rate and bandwidth convey rather complex messages e.g. warning calls and geographic identity in bearded seals [12], beak morphology in Darwin's finches [13], and indicates male quality in several species of birds [14]. However, parameters estimated using manual segmentation performed by experts suffer from subjectivity of analysis and lack of consistency. Thus, a systematic approach to parameter estimation is also desired from the perspective of validity of the biological results obtained from vocalization analysis.

Various parameters have been used in biological research. Some common parameters include: trill duration and bandwidth, number of syllables, mean syllable rate and duration, syllable rate acceleration / deceleration and maximum amplitude time to trill duration ratio.

1.2 Definitions and Notations

Throughout the thesis, the following notations hold:

- $s(t)$ is an input signal of a continuous variable t, and $t = 0$ is its starting time.
- $s[n]$ is the discrete anti-aliased signal obtained through low pass filtering and sampling of $s(t)$, with sample rate f_s. Usually in audio applications, $f_s = 44.1$kHz, but this is not mandatory.
- $\mathcal{F}\{s(t)\}(\omega)$ is the continuous Fourier transform of $s(t)$, defined as

$$\mathcal{F}\{s\}(\omega) = \int_{-\infty}^{\infty} s(t)\ e^{-i\omega t} dt \tag{1.1}$$

ω is the angular frequency in radians. Alternatively, linear frequencies can be used:

$$\mathcal{F}\{s\}(f) = \int_{-\infty}^{\infty} s(t)\ e^{-i2\pi ft} dt \tag{1.2}$$

with f measured in Hz.

- $\mathcal{F}\{s[n]\}[k]$ is the discrete Fourier transform (DFT) of $s[n]$, defined as:

$$\mathcal{F}\{s[n]\}[k] = \sum_{n=0}^{N-1} s[n] e^{-i\frac{2\pi k}{N}n} \tag{1.3}$$

  $s[n]$ is implicitly assumed to the periodic of period N, and the integer frequencies satisfy $0 \leq k \leq N-1$.

- The short time Fourier transform (STFT) of $s(t)$ is defined as:

$$S(t_0, \omega) = \mathcal{F}\{s(t)w(t-t_0)\}(\omega) \tag{1.4}$$

  $w(t)$ is a predefined symmetric window (usually supported on a symmetric interval).

- The STFT of a discrete signal is defined similarly

$$S(n_0, k) = \mathcal{F}\{s[n]w[n-n_0]\}[k] \tag{1.5}$$

  with a discrete symmetric window $w[n]$. n_0 and k are integers representing the discrete time instance and discrete frequency, respectively.

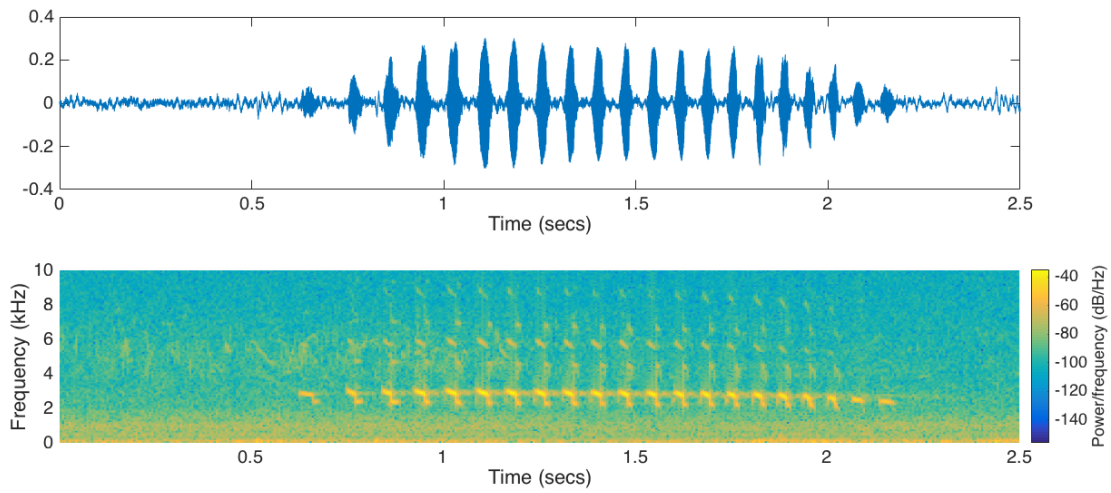
- The *fundamental frequency* f_0 of an audio signal is the inverse of T_0, the fundamental period, the smallest positive time shift under which the signal is an exact replica of itself at least locally.
  OR
  The *fundamental frequency* f_0 of a periodic audio signal is the smallest positive frequency in its Fourier series representation which is perceived as pitch.

- An *harmonic* of an audio signal is a positive integer multiple of f_0 introduced as vibrations in the system of sound production, and it is inversely related to the *fundamental period* T_0: $f_0 = \frac{1}{T_0}$
  OR
  An *harmonic* of an audio signal is any positive integer multiple of f_0 which has a corresponding term in the Fourier series representation of the signal.

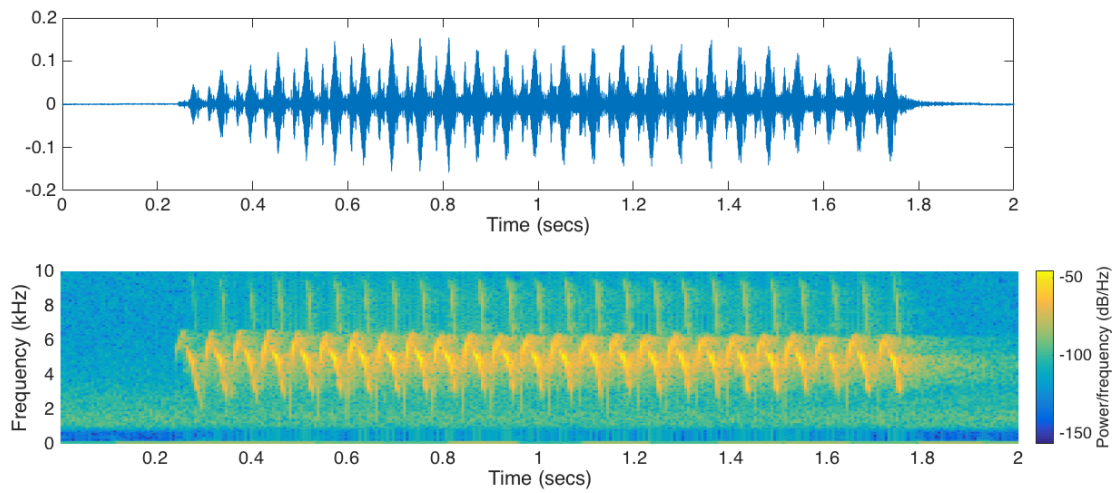
1.3 Problem Description

A relatively simple taxonomically widespread vocalization is the "trill", a rapidly repeating syllable vocalization [13]. Though by definition not complex, trill physical attributes such as trill rate and bandwidth convey rather complex messages e.g. warning calls and geographic identity in bearded seals [12], beak morphology in Darwin's finches [13] and indicates male quality in several species of birds [14]. Trill vocalizations comprise repetitions of a basic vocal unit called a syllable. The duration of a syllable ranges from as little as 6 ms to over 200 ms[1]. The syllable repetition frequency (SRF) or trill rate is the ratio between the number of syllables and the trill duration. The data we collected yielded SRF bounds of 6 and 75 syllables per second. According to a research conducted by Podos [16] on a songbird family (Emberizidae), there are at least two different modes of trill production. The first one involved both inspiration and expiration, and is used in production of trills with relatively slow SRF (under 36 according to Podos.) The second one involves only expiration, and is used in production of trills with SRF of over 36 syllables per second. The fundamental frequency can be either fast changing or approximately constant. Podos [16] points to an interesting connection between the bandwidth of the fundamental frequency and the SRF. The maximal values these two parameters can equal to are inversely related. That is, trills with very wide bandwidth have low SRF, and very fast trills (high SRF) exhibit narrow bandwidth. The explanation is rooted in the physical model which is believed to produce trills. The bird vocal tract (syrinx) has to change shape in order to produce different frequencies. For a wide bandwidth,energy has to be invested in rapidly changing the shape of the syrinx. To create syllables, the syrinx has to pulsate in the SRF frequency. High SRF necessitates fast pulsations. It is therefore argued that a tradeoff occurs between these two parameters which arises from physical constraints. These constraints are likened to the mechanism of hand-clapping, in which one cannot clap hands both fast and loud. A tradeoff exists between maximal speed and maximal loudness in this case as well. The number of syllables can be anywhere between 3 and up to 40 and more. Examples can be seen in Figs. 1.1a-1.1c.

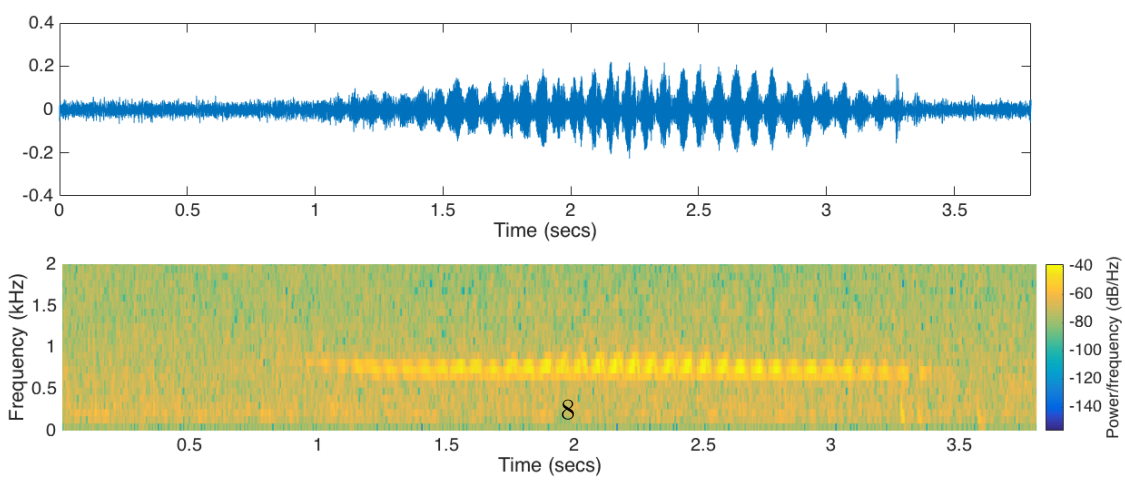
[1] based on data we collected from the Xeno-Canto[15] Website



(a) White-throated Kingfisher (*halcyon smyrnensis*) trill spectrogram and waveform. f_0 range: 1.8-3.2 kHz



(b) Dark-eyed Junco (*junco hyemalis*) trill spectrogram and waveform. f_0 range: 3-6.5 kHz



(c) Tawny Owl (*strix aluco aluco*) trill spectrogram and waveform. f_0 range: 0.6-0.8 kHz

Figure 1.1: Examples of three different trills in time and time-frequency representation. The trills differ in all major parameters: Number of syllables, duration, SRF and bandwidth.

1.3.1 Acoustic Model

The basic assumptions on trill vocalizations from an acoustic perspective are as follows. Trills comprise of discrete syllables. Depending on the method of production, intervals of silence exist between syllables, or at least content of significantly lower amplitude. The syllables are tonal, and harmonics of higher order are sometimes present, in contrast with inter-syllable intervals which are characterized by higher degrees of noise.

In order to understand trills and design algorithms for automatic analysis, the phenomena must be modelled simply yet effectively. In [17], a method for speech synthesis is presented which follows the Harmonic+Noise Model (HNM). The HNM models human speech as comprising an harmonic part and a noise part. An equation which models a voiced signal would therefore be

$$s(t) = \sum_{j=-H(t)}^{H(t)} A_j(t) e^{2\pi i j f_0(t)} + n(t) \tag{1.6}$$

$f_0(t)$ is the fundamental frequency of $s(t)$ at time t. $H(t)$ is the number of harmonics present, and $A_j(t)$ are the respective amplitudes. The noise is denoted $n(t)$, and can be modeled with filtered white noise.

This model is applicable for trills provided the amplitudes are pulse modulated.

We therefore suggest the following adjustment:

$$A_j(t) = \sum_{k=1}^{K} \mathbb{1}_k(t) \tilde{A}_j(t) \tag{1.7}$$

where K is the number of syllables in the trill. $\mathbb{1}_k(t)$ is the indicator function of the k'th syllable, i.e. it is 1 on the interval which supports the syllable and 0 outside it. The adapted model is therefore

$$s(t) = \sum_{k=1}^{K} \mathbb{1}_k(t) \sum_{j=-H(t)}^{H(t)} \tilde{A}_j(t) e^{2\pi i j f_0(t)} + n(t) \tag{1.8}$$

This model provides a conceptual framework with which segmentation algorithms can be developed.

1.3.2 Model Limitations

Some trills manifest a second type of vocalization, which we call an *inter-syllable*. We believe it is the source of alternation of expiration and inspiration which happens in trills of very quick pulsations [18]. However, without an intrusive mechanism of verification, this belief remains an untested hypothesis. In fact, little is currently known about the acoustic

structure of bird vocalizations. Members of the same species may manifest the existence or lack of inter-syllables in a single type of trill vocalization. Moreover, even a single individual may express trills with or without inter-syllables. A single species example can be seen in Figs. 1.2 and 1.3. The example shows differences between two individual white throated kingfisher males with respect to presence of inter-syllables. While inter-syllables are barely present in the first recording, they are dominant in the second recording, which makes the task of segmentation extremely difficult. Since the vocal mechanism which is responsible for the presence of inter-syllable has not been verified, nor the reasons for its participation or abstinence in the trill vocalization, it has been left out of the acoustic model of trill vocalizations.

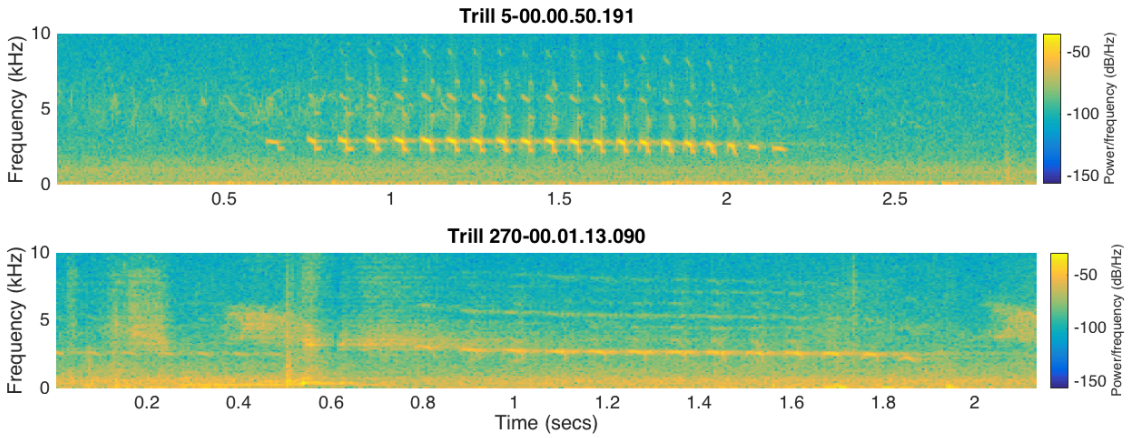


Figure 1.2: Comparison of White-throated Kingfisher trills. Top: Inter syllables are not present. Spectrogram blobs are easily separated. Bottom: Inter syllables are present. Spectrogram blobs are unseparable.

Although (1.8) describes the ideal signal desired in the recording, the model can sometimes fail in recordings taken from natural habitats. In natural habitats it is rare that a single bird is recorded in isolation or without interruptions. Other birds present in the recording may disrupt the signal and therefore limit the applicability of the model. It should be noted however that the segmentation algorithms developed and described in this thesis do not rely directly on the acoustic model of the signal, and estimation is performed using different insights regarding the existence of syllables and their nature. Therefore, the assumptions of the model can be weakened to allow presence of other birds in the recordings. Instead of assuming a single trill in the recording, it is assumed that the trill of interest is the strongest vocalization present in the signal.

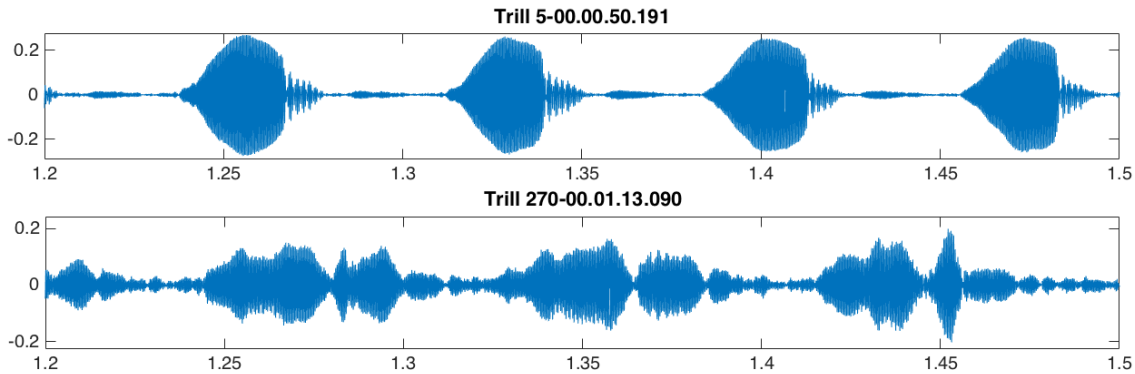


Figure 1.3: Segments of 300 ms which have been excerpted from the trills depicted in Fig. 1.2. Whereas in the top figure the syllables are clearly discernible, and inter-syllables are barely present, the syllables in the bottom sub-figure are difficult to discern.

1.3.3 Problem Formulation

The research reported in this thesis was conducted with the aim of developing an automatic system of trill parameter estimation. Initially, the system was developed to be applied in a study of White-Throated Kingfishers (*Halcyon Smyrnensis*) in the Hula Lake park, where correlation between trill rate and male quality had already been shown [19]. However, subsequently it was generalized to fit for a wide range of trills as well as other vocalizations. An automatic estimation system is beneficial both for research and conservation. Automatic parameter estimation may facilitate the discovery of other relations between acoustic parameters and biological or genetic parameters (CITATION?). The development of a working automatic system can help monitor populations and maintain ecosystem balance [20].

Syllable segmentation is the problem of estimating

1. K, the number of syllables
2. $\mathbb{1}_k(t)$, the support of each syllable

Fundamental frequency estimation involves estimating $f_0(t)$. The problem of fundamental frequency estimation can be solved independently of segmentation and amplitude estimation. Over the years, many fundamental frequency estimators were formulated and tested. In this thesis, the YIN fundamental frequency estimator [21] was selected since it was found to outperform other estimators when used for bird audio recordings [22].

Syllable segmentation is therefore the main focus of this thesis. In analysis of the problem and its solution, it is further divided into 2 parts:

1. Syllable detection, in which individual existence of syllables is determined in the trill

signal. Time instances $\{t_k\}_{k=1}^{\hat{K}}$ are extracted, in which every t_k belongs to a single unique syllable. In this way the estimate for K is also obtained.

2. Syllable segmentation, in which the indicators $\{\mathbb{1}_k(t)\}_{k=1}^{\hat{K}}$ are estimated.

The acoustic parameters of interest are either time domain or time-frequency domain based. They can be derived using computations applied to the model parameters of (1.8). Mathematically, this relation may be realized as

$$p = f(K, \mathbb{1}_k(t), H(t), A_j(t), f_0(t)) \tag{1.9}$$

where p is the acoustic parameter of interest and f is a function used to retrieve it.

However, the amplitudes $A_j(t)$ and number of harmonics $H(t)$ are usually irrelevant for estimation of acoustic parameters of interest. The main reason is the lack of information regarding proximity between recording equipment and the audio source (bird recorded.) It is rare that birds are recorded in sterile environments. Usually, recording is done outdoors, with unknown SNR levels and with recording equipment of varying quality. For this reason, amplitude information cannot be used reliably to compare between individuals, or even between two trill recordings of the same individual. For the same reason, harmonics information is also unreliable. It is impossible to tell how many harmonics were lost (if any) due to proximity from audio source. Therefore, (1.9) can be updated to

$$p = f(K, \mathbb{1}_k(t), f_0(t)) \tag{1.10}$$

It follows that all the acoustic parameters of interest can be calculated based on accurate estimation of the model parameters $K, \mathbb{1}_k(t), f_0(t)$, i.e. accurate segmentation and f_0 estimation.

The signals dealt with are assumed to contains a single trill call. It is further assumed that margins of 150 ms length are given, i.e. the first and last 150 ms of the recording does not contain a trill.

1.3.4 Course of Action

In this thesis, we propose a solution to the problem of automatic trill syllable segmentation. The solution was intended to fit only the problem of segmentation in White-throated Kingfishers, but was later generalized to fit any type of trill-like animal calls. Acoustic parameters are then estimated. A system was developed and tested

The system proposed in this thesis preforms automatic segmentation and acoustic parameter estimation for appropriate input audio signals. As a pre-processing step, it estimates the frequency band of interest in the signal using an algorithm presented in chapter 3. It then bandpass filters the signal according to the appropriate band, and performs f_0 estimation using the YIN algorithm. For the sake of completeness, the YIN algorithm is

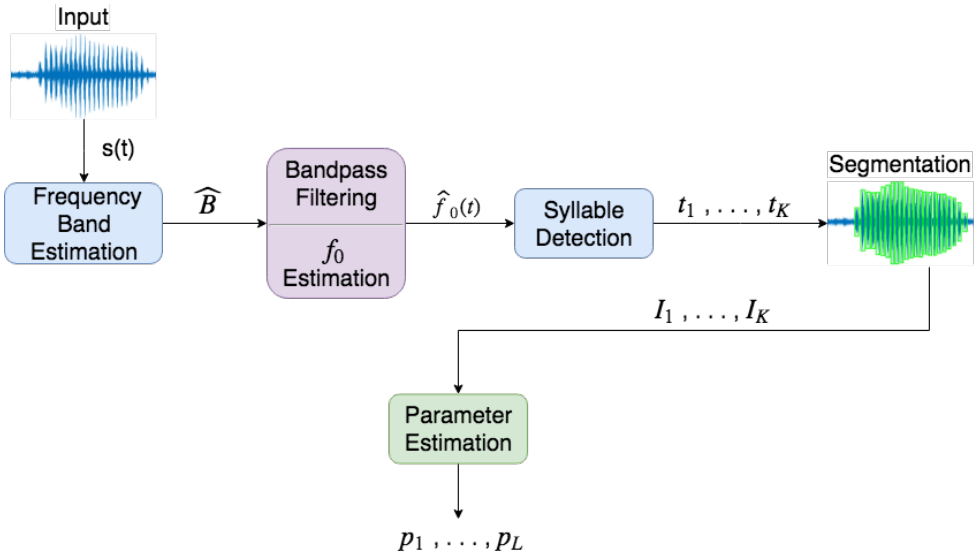


Figure 1.4: An overview of the proposed segmentation and parameter estimation framework

described in chapter 2. Then, syllable detection and segmentation is performed. Finally, acoustic parameters are estimated. This system is depicted schematically in Fig. 1.4.

The system is fully described in the following chapters. The system uses the YIN algorithm, which is an autocorrelation based fundamental frequency estimator. The system uses autocorrelation also to estimate SRF and syllable presence. Both are described in chapter 2. A major part of the generalization step involved detection of the frequency band of interest prior to band-pass filtering and segmentation. The frequency band detection algorithm is described in chapter 3. The two different segmentation algorithms are proposed in chapter 4. Testing the results obtained by the system is not trivial since an objective ground truth does not exist, and therefore had to be developed. This problem was tackled in two different ways. A comparison was held between results of the segmentation and parameter estimation algorithms and the manual work of a bird vocalization expert. In addition, two artificial data-sets were created. They comprise audio tracks which are in many cases indistinguishable from natural recordings to human ears. However, they present an objective ground truth for the estimation system. Description of the artificial ground truth data creation as well as the results obtained is presented in chapter 5. Finally, conclusions are presented in chapter 6.

Chapter 2

Autocorrelation Based Methods for Estimation of Fundamental Frequency and Trill Presence

2.1 Autocorrelation Function

Autocorrelation is a measure indicating periodicity in a time dependent signal. In audio signal processing, it is the basis for many methods for estimation of fundamental frequency. It can also be used to estimate tempo in music signals.

The widespread definition of autocorrelation of a real deterministic signal $s(t)$ defined on $\mathbb{R}$ is

$$r(\tau) = \int_{-\infty}^{\infty} s(t)s(t+\tau)dt \tag{2.1}$$

DSP implementations must deal with discretization of the integral, but also to replace the integration domain $(-\infty, \infty)$. Moreover, the infinite integration domain is typically irrelevant for most organic audio signals because they are classified as non-stationary processes due to their time-varying nature. Having said that, human speech signals exhibit stationarity in short time segments (20-40 ms) which are interpreted as steady state modes of the speech production mechanism. Therefore, in speech and similarly produced signals, it makes sense to define autocorrelation only for short segments of the signal, and for short time lags. The discrete short-time autocorrelation of a signal $s[n]$ is defined as

$$r_t(\tau) = \sum_{n=t}^{t+L-1} s[n]s[n+\tau] \tag{2.2}$$

where L is the window size in samples and the time lag τ is typically shorter than L.

The autocorrelation function (ACF) (2.2) of a periodic signal $s[n]$ is periodic as well, with peaks located at integer multiples of the period. The autocorrelation method for

f_0 estimation works by finding the highest peak in the ACF of the periodic signal in a predefined search range. The time lag corresponding to the highest peak is the estimated fundamental period, and the estimated f_0 is its reciprocal.

The ACF is one of most common methods for f_0 estimation, but its major drawback is high dependency on the time lag search range. If the search range is small, it may miss the fundamental period, and erroneously identify the period of one of the harmonics instead of the fundamental. On the other hand, if it is too large, it may erroneously choose a peak corresponding to a longer time lag, usually in integer multiples of the fundamental period. These two types of estimation errors can occur even if the search range is adequate, for example when one of the harmonics is stronger than the fundamental, when the signal is noisy or when there are numerous audio sources, yielding numerous fundamentals which may yield several fundamental frequencies simultaneously. The problem with correct search range definition is that the search range of the period is the reciprocal of the frequency search range. Therefore, the frequency band in which the fundamental resides must be either known or estimated in advance. This is clearly problematic for a wide range of applications, especially in polyphonic signals.

Despite its drawbacks, autocorrelation is used successfully (Sec. 2.3) to estimate syllable rate in a trill signal by exploiting periodicity of the signal's envelope. An initial estimate is first obtained using an over inclusive search range. This estimate is used to define an appropriate search range which is used to obtain the final estimate. Search range selection process is detailed in Sec. 2.3.5.

The YIN algorithm is an autocorellation based f_0 estimator which presumes to overcome these difficulties, and is presented next.

2.2 The YIN algorithm

The YIN algorithm [21] is based on the difference function

$$d_t(\tau) = \sum_{n=t}^{t+L-1} (s[n] - s[n+\tau])^2 \tag{2.3}$$

which is related to the ACF by the equation

$$d_t(\tau) = r_t(0) + r_{t+\tau}(0) - 2r_t(\tau) \tag{2.4}$$

This is almost an inverse relation. However, since the second term on the right hand side of (2.4) is dependent on the time lag τ, maxima of $r_t(\tau)$ and minima of $d_t(\tau)$ may fail at times to coincide. Nevertheless, the replacement of the ACF by the difference function in Eq. (2.3) reduces error rates on a selected speech dataset by 80% [21]. One way to explain this reduction is that AM signals with constant frequency but increasing amplitude manifest $r_t(\tau)$ which increases with τ at integer multiples of the fundamental

period, denoted T_0. If $k \cdot T_0$ is in the search range for $k \in \mathbb{N}$, and $r_t(k \cdot T_0) > r_t(T_0)$, the fundamental period will be erroneously estimated as $k \cdot T_0$. In contrast, $d_t(\tau)$, is an euclidean norm if treated as a function of $s[n]$, and indifferent to the direction of change in amplitude, as a consequence. Therefore, the value of $d_t(k \cdot T_0)$ will be higher than the minimum at the period, whether the amplitude increases or decreases.

A second improvement further reduces estimation errors in which the period of an harmonic or formant is detected rather than the fundamental by minimizing a normalized version of (2.3) named the cumulative mean normalized difference function (CMNDF):

$$d'_t(\tau) = \begin{cases} 1 & \text{, if } \tau = 0 \\ \frac{d_t(\tau)}{\frac{1}{\tau}\sum_{j=1}^{\tau} d_t(j)} & \text{, else} \end{cases} \tag{2.5}$$

The CMNDF normalizes $d_t(\tau)$ by its average on lower timelags. Normalization improves estimation errors caused by a strong harmonic or formant contributing to a lower dip of $d_t(\tau)$ for $\tau < T_0$. The CMNDF tends to remain large for small lags because its value is close to the average when the number of averaged timelags is small, under the assumption of bounded variation of $d_t(\tau)$. It falls below 1 only when $d_t(\tau)$ is below the average. Therefore, dips of $d_t(\tau)$ caused by harmonics are obscured by the CMNDF.

Another advantage of using the CMNDF over $d_t(\tau)$ is that an absolute threshold can be set independently of acoustic characteristics of the signal. Using the CMNDF, the algorithm estimates the fundamental period by choosing the smallest timelag τ with a dip lower than the the absolute threshold. If no dips below the threshold are present, the algorithm picks the timelag with the lowest dip in the search period. A commonly used absolute threshold is 0.1, though it can be adjusted by the amount of noise expected in the signal.

The absolute threshold has an interesting interpretation as the ratio of non-periodicity which can be tolerated in the estimation. This can be explained by the relation

$$\frac{1}{2L}\sum_{n=t}^{t+L-1} s^2[n] + s^2[n+T_0] = \frac{1}{4L}\sum_{n=t}^{t+L-1} (s[n]+s[n+T_0])^2 + \frac{1}{4L}\sum_{n=t}^{t+L-1} (s[n]-s[n+T_0])^2 \tag{2.6}$$

which can be easily proved using the identity $2(s^2[n] + s^2[n+T_0]) = (s[n]+s[n+T_0])^2 + (s[n]-s[n+T_0])^2$. Supposing $T_0 \ll L$, the left hand side term approximates twich the signal power. Both terms of the right hand side are positive, and constitute a partition of the signal power. The second term is zero if $s[n]$ is periodic with period T_0, and can thus be interpreted as the non-periodic component's contribution to the power. It is proportional to $d_t(\tau)$ for $\tau = T_0$, which is the numerator of $d'_t(T_0)$. On the other hand, the denominator of $d'_t(T_0)$ which equals the average $\frac{1}{T_0}\sum_{j=1}^{T_0} d_t(j)$ is approximately twice the signal's power. This can be explained using (2.4) and noting that on a full period, $r_t(\tau)$ goes through zero from maximum to minimum and back because as the correlated vectors move gradually out of phase to negative phase back. Summation "cancels out" these terms, leaving only

the two energy terms of (2.4). Therefore, $d'_t(T_0)$ is equal to the ratio of the non-periodic contribution to power to the total power. In light of this interpretation, an absolute threshold of the CMNDF can be viewed as maximal amount of non-periodicity which can be tolerated in the estimation.

These last steps are correct if T_0 is an exact integer multiple of the sampling period. The sampling process may obscure a dip when f_0 not is a divisor of the sample rate, because the fundamental period is not an integer multiple of the sampling period. In that case, subsample interpolation is needed to approximate the true dip value and the fractional number of samples of its timelag. YIN performs quadratic interpolation around the dip, which is motivated by the fact that the autocorrelation can be represented by a sum of cosines (As a Fourier transform of a real symmetric function) which can in turn be approximated by a Taylor series of order 2. The linear term vanishes due to symmetry. Each dip of the CMNDF uses its 2 nearest values to fit a parabola. The dip of the parabola replaces the original dip, and the signed fractional timelag is added to the integer timelag of the original dip. Finally, the estimation may be improved by replacing the estimate by the best local estimate, which is the estimate with the smallest dip value $d'_{\tilde{t}}(\tilde{T}_0)$, with $\tilde{t}$ chosen from a small neighbourhood around t: $\tilde{t} \in [t - \epsilon, t + \epsilon]$.

2.3 Short-Time Energy Autocorrelation: Trill Presence Estimator

Periodic phenomena in acoustic signals exists in a broader context than that of fundamental frequency and harmonics. In music signals, percussive instruments manifest periodicity which sets tempo and rhythm. In trill signals, periodicity is manifested in syllable rate and may be exploited for syllable detection. Syllable periodicity is manifested in the trill envelope in the same way as periods of harmonic signals. The autocorrelation method can be adapted for syllable detection in bandpass signals when applied to a short-time energy function or any smooth envelope function of the signal. The resulting method is used both for estimation of trill presence as part of the frequency band of interest detection algorithm in Sec. 3.6, and also in the syllable segmentation algorithm in Sec. 4.2.1.

2.3.1 Modified Autocorrelation Function and Correlation Coefficient

However, one adjustment has to be made to the ACF. In discrete-time periodic signals, the period length is approximately constant. In trill signals, syllable duration usually varies even if syllable repetition rate is approximately constant. This difference is accounted for by making window size L dependent on timelag τ. Since the function is intended to represent correlation between consecutive signal segments, choosing $L = \tau$ seems like a natural choice. Mathematically, suppose $E[n]$ is a short time energy function of signal $s[n]$.

The short time energy autocorrelation function (STEACF) is

$$m_t(\tau) = \sum_{n=t-\frac{\tau}{2}}^{t+\frac{\tau}{2}} E[n]E[n+\tau] \tag{2.7}$$

For convenience, the analysis point t is always in the middle of the first correlated segment. The center of the second segment is τ. The problem posed by this definition is that as τ grows, so is the number of elements in the sum, and so is the upper bound for $m_t(\tau)$. To be able to measure correlation, $m_t(\tau)$ is normalized by the L_2 norm of the correlated segments. The result is the short-time energy correlation coefficient (STECC)

$$\rho_t(\tau) = \frac{m_t(\tau)}{\sqrt{\sum_{n=t-\frac{\tau}{2}}^{t+\frac{\tau}{2}} E^2[n] \sum_{n=t-\frac{\tau}{2}}^{t+\frac{\tau}{2}} E^2[n+\tau]}} \tag{2.8}$$

$\rho(\tau)$ is bounded by the Cauchy-Schwartz inequality:

$$-1 \leq \rho(\tau) \leq 1 \tag{2.9}$$

In particular, if $E[n] \geq 0$, then also $0 \leq \rho(\tau) \leq 1$.

The first value of τ for which $\rho_t(\tau)$ has a peak is an estimate of the time difference between the syllable located at time t and the following syllable.

2.3.2 Neighbourhood Maximization

In principle, this process can repeat itself with $\tilde{t} = t + \tau$ as the new analysis point, to find the time location of the following syllable, until the end of the trill. However, errors can accumulate to a half or even full syllable offset, which creates large estimation errors.

Trill syllables are recognized by short bursts of energy. It is therefore natural to choose analysis points t for which $E[t]$ is maximal. This argument motivates the following offset correction

$$\tilde{t} = \underset{t-\epsilon\cdot f_s \leq s \leq t+\epsilon\cdot f_s}{\arg\max} E[s+\tau] \tag{2.10}$$

For bird trills encountered in this research, $\epsilon = 30$ milliseconds was chosen experimentally to yield good results. Step (2.10) is referred to as **Neighbourhood Maximization**.

2.3.3 Algorithm Steps

STECC peak finding combined with neighbourhood maximization are the basic building blocks of the syllable detection algorithm. However, trill analysis points are chosen as the time instances with highest energy levels. It makes sense to choose the first analysis point as the highest energy peak in the signal. Therefore, the algorithm does not operate

linearly in time, but rather finds a starting analysis point, and progresses both forwards and backwards in time.

To find the starting point, t_0, one would naively choose $t' = \arg\max_t E[t]$. However, if transient noises with high energy are present in the signal, they can cause the starting point to occur far away from the trill. For this reason, caution is exercised by low-pass filtering $E[t]$ with a 0.5 second long moving average filter. Denote the filter output by $\bar{E}[t]$. Since trills manifest rapid energy emission over time, $\bar{t} = \arg\max_t \bar{E}[t]$ is a rough indication of the trill time location in the signal. Choosing t_0 in proximity to $\bar{t}$ improves chances of t_0 occurring during a trill and not before or after it:

$$t_0 = \arg\max_{|t-\bar{t}|\leq T} E[t] \tag{2.11}$$

with $T = 0.3$ seconds chosen to account for the minimal expected trill duration. Choice of t_0 is depicted in Fig. 2.1, top.

After t_0 is detected, the algorithm advances forwards in time using the STECC and neighbourhood maximization to find $t_1 = t_0 + \tau_0, t_2 = t_1 + \tau_1, \ldots, t_I$, until a stopping criterion is met. This step is depicted in Fig. 2.1, middle. Then, it advances backwards in time in the same manner to find $t_{-1} = t_0 - \tilde{\tau}_0, t_{-2} = t_{-1} - \tilde{\tau}_{-1}, \ldots, t_{-J}$, as depicted in Fig. 2.1, bottom. Finally, the sequences are concatenated and relabeled to produce the sequence of syllable time estimates

$$u_1 = t_{-J}, u_2 = t_{1-J}, \ldots, u_{J+1} = t_0, u_{J+2} = t_1, \ldots, u_{I+J+1} = t_I \tag{2.12}$$

2.3.4 Stopping Criterion

The stopping condition is either of the following:

1. $\rho(t_j) < 0.7$
2. $E[t_j] < E^{NF}[t_j]$, with $E^{NF}[t_j]$ defined as the energy noisefloor. The noisefloor is calculated with a 1 second length median filter: $E^{NF}[t] = \text{MEDFILT}(E[t])$.

The first condition ensures the correlation between adjacent segments is high enough. It is based on the assumption that syllable envelopes have similar contours which yield high correlation, while correlation between syllable and noise envelopes are expected to yield lower correlation. The second condition is based on the assumption that the trill is the strongest component in the signal in terms of energy. It ensures that the energy level of the detected syllable does not go below the noisefloor.

The stopping condition is not necessarily robust and may seem arbitrary. It is prone to false alarm type of errors. However, for the purpose of estimating trill presence in the signal, exact syllable detection is unnecessary. For the use of the algorithm in the syllable detection step, outlier removal is performed prior to segmentation to remove false alarm type of errors.

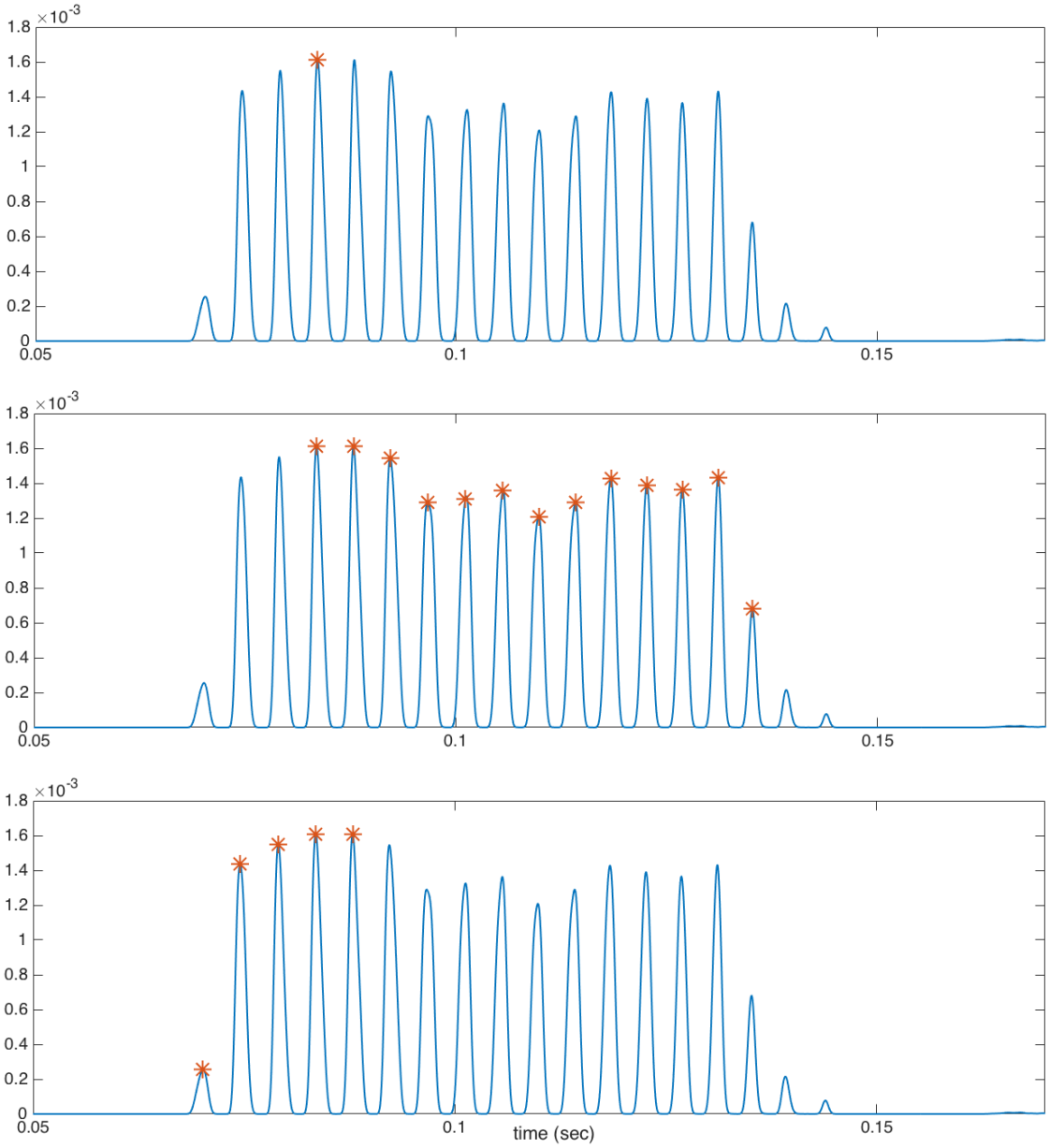


Figure 2.1: Top: at the first stage of STECC t_0 is found. Middle: the algorithm progresses forward in time and finds syllable indicators until the stopping criterion is met. Bottom: the same process is performed starting from t_0 and progressing backwards in time.

2.3.5 Timelag Search Range

Due to physical constraints, the trill syllable repetition interval (SRI) is bounded. An estimate of syllable emission frequency is no less than 6 syllables per second and no more than 75 syllables per second [1]. This translates to the bounds $m_{SRI} = \frac{1}{75} = 0.0133$ and $M_{SRI} = \frac{1}{6} = 1.667$ so that $m_{SRI} \leq SRI \leq M_{SRI}$. Therefore,

$$m_{SRI} \leq \tau \leq M_{SRI} \tag{2.13}$$

is the search range for τ calculated by the STECC.

However, this search range is too general. In trills with very fast syllable emission, the search range will include the next few syllables. If the correlation with the consecutive syllable is lower than the one after it, the syllable will be skipped. On the other hand, in slow syllable emission trills, inter syllables may be falsely recognized as syllables. Therefore, the search range of Eq. (2.13) should be updated by a more restricted range to reduce estimation errors. The solution is to run the algorithm once with a full search range to obtain an estimate for the SRI, and then run the algorithm a second time with a restricted search range calculated with the estimated SRI.

The sequence of estimates $u_1, \ldots, u_K$ returned by the first run of the algorithm is used only for the estimate:

$$\widehat{SRI} = P_{70}(u_2 - u_1, \ldots, u_K - u_{K-1}) \tag{2.14}$$

A 70th percentile is used instead of a median because, when u_i and u_{i+1} are not actually in syllables but rather in noise, $u_{i+1} - u_i$ is typically small. If the stopping criterion is too permissive (and it's usually is), false positive errors are very common at the trill's edge (i.e. at the beginning or end.) In false positive errors, the energy around an analysis point is due to noise, which have the effect of negative bias on a mean and to a lesser extent to a median. The percentile's preference of higher values is therefore introduced to counter this bias.

Once the $S\hat{R}I$ is obtained, a limited search range is obtained

$$0.4S\hat{R}I \leq \tau \leq 1.6S\hat{R}I \tag{2.15}$$

and is used for the second run of the algorithm. The new set of analysis points $u'_1, \ldots, u'_{K'}$ produces the final estimate of syllable time instances.

[1] According to a sample of over 300 trills recorded in our lab or excerpted from Xeno-canto.

Chapter 3

Frequency Band Estimation

The algorithms presented in chapter 4 (Segmentation Algorithms) rely on parameters derived from the short-time energy in a sub-band in which the fundamental frequency f_0 resides. Fundamental frequency estimation therefore must occur prior to any further segmentation. As discussed in chapter 2, f_0 estimation already requires knowledge of the minimal and maximal detectable frequencies in a standard autocorrelation method. The YIN algorithm requires knowledge of the minimal frequency only. Moreover, as discussed in [22], f_0 estimation of bird vocalizations suffers from inaccuracies and problems not usually present in speech or music signals. One of the problems is the wide frequency range in which the fundamental frequency may be expected to reside in bird vocalizations. Therefore, estimation of the frequency band in which f_0 resides can improve the accuracy of its esimation.

In this chapter, an algorithm for detection of f_0 frequency band is presented, designed especially for trill vocalizations. The algorithm uses the spectrogram to create a frequent items histogram, from which several frequency bands are extracted. An optimal band is selected from these bands, by applying regular repetitiveness criteria. The following sections describe the steps used. For presentation purposes, the discrete case is favoured over the continuous case since implementation considerations are easier to explain in the discrete domain.

3.1 Spectrogram Processing

The short-time Fourier transform of the discrete signal $s[n]$ is calculated using the definition

$$S(u, f) = \mathcal{F}\{s[n]w[n-u]\}[k] \tag{3.1}$$

with a discrete symmetric window $w[n]$. u is the discrete time index and f is the frequency of the appropriate DFT bin. Each analysis frame ($N = 256$) is multiplied by a Hann window. A zero padding factor of $Z = 4$ is introduced to improve frequency resolution.

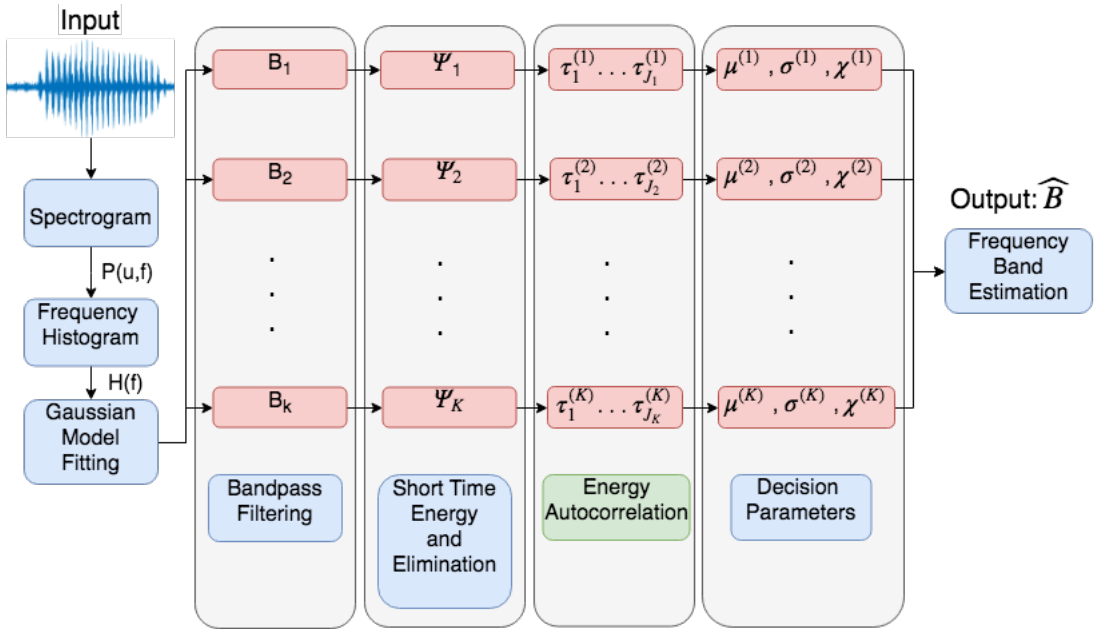


Figure 3.1: A schematic block diagram of the algorithm for fundamental frequency band estimation. For $1 \leq i \leq K$, The variables B_i represent candidate frequency bands. $\Psi_i[n]$ represent short-time energy functions. $\tau_1^{(i)}, \ldots, \tau_{J_i}^{(i)}$ is a sequence of time instances representing the time location of syllables detected at the ith frequency sub-band. $\mu^{(i)}, \sigma^{(i)}$, and $\chi^{(i)}$ are the mean, standard variation and mean energy variation, respectively, of the time instances sequence $\tau_1^{(i)}, \ldots, \tau_{J_i}^{(i)}$. For more details, please see the text.

An overlap factor of 75% is also introduced to account for time resolution loss in discrete processing. In the standard sampling rate of 44.1 kHz, each frequency bin in the DFT is about 43 Hz wide. The spectrogram (or log-spectrogram) is calculated from the STFT by:

$$P(u, f) = 10 \log(|S(u, f)|^2) \tag{3.2}$$

The spectrogram can be represented as an image (or more generally, as a 3 dimensional surface defined on a 2 dimensional grid). Therefore, image processing techniques for noise removal and object enhancement can be applied to improve data representation (see Fig. 3.2.)

1. A 5×5 median filter is applied. This is a well known technique which provides limited noise reduction while avoiding blurring contour lines of image objects.

2. Median clipping is performed. Treating $P(u, f)$ as a matrix, and denoting P_u as the row corresponding to time index u and P_f as the column corresponding to frequency

f, median clipping is defined as:

$$\text{MEDCLIP}(P)(u,f) = \begin{cases} P(u,f) & , \ P(u,f) > \max\{\text{median}(P_u), \text{median}(P_f)\} \\ -\infty & , \ else \end{cases} \tag{3.3}$$

Median clipping zeros out pixels which are smaller than either the median of the row or column of the pixel. It enhances spectrogram data by comparing each pixel to the noise profiles in the appropriate time instance and frequency bin [23, 24]. It is supported by the assumption that the noise profile is much lower than the median energy in each bin, i.e. across all time instances, and in each time instance, i.e. across all frequencies. Therefore, only pixels which stand out above the noise remain. [24] used a higher threshold for median clipping, namely $3 \times \max\{\text{median}(P_u), \text{median}(P_f)\}$. However, the main purpose of median clipping in [24] is for detection of bird presence in an audio clip data set. In contrast, in trill signals often encountered, the syllables vary in amplitude, and setting such a high threshold would wrongfully wipe out many weak syllables present in the signal. Therefore, for the purpose of this thesis, syllable detection and segmentation, (3.3) used a more permissive threshold.

3.2 Frequency Histogram

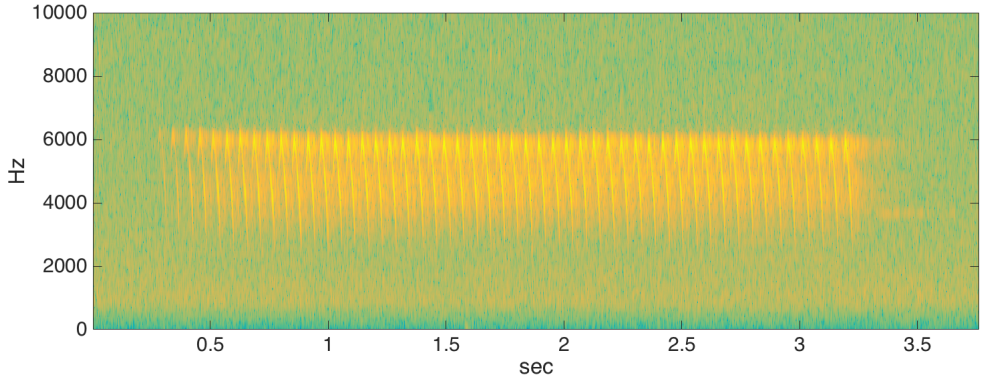
For every time index u_0, the spectrogram $P(u_0, f)$ represents energy levels in $s[n]$ of each frequency bin in a time window centered at u_0. The frequencies f for which $P(u_0, f)$ gets high values are the strongest frequency components at time u_0. This property can be used to find all prominent frequencies for all time instances u. When stacked in a histogram, it can be interpreted as a "most frequent frequencies" histogram.

For every time index u_0, the 10 highest energy bins are selected from the relevant bins only. These bins are used to populate a histogram H_0 of the most frequent frequencies in the signal. For symmetry reasons, only bins of positive frequencies are used. An example is illustrated in Fig. 3.3 (top).

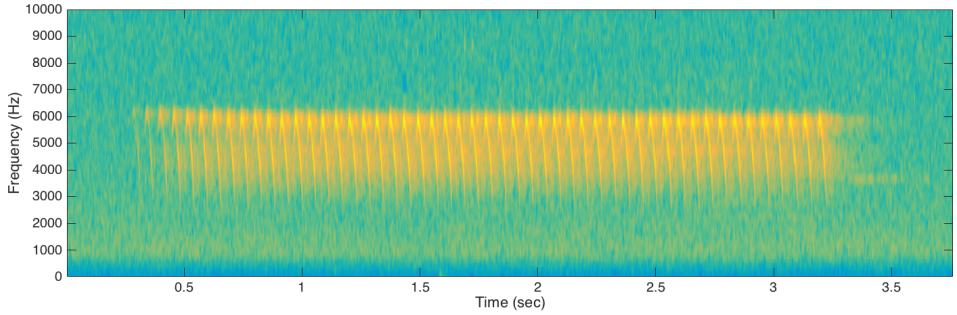
Since harmonic bird vocalizations usually do not exist in the range below 500 Hz, frequencies in this range are ignored. In the standard audio sample rate, the DFT bins corresponding to those frequencies are bins 1 - 12. Therefore, even if those bins contain high energy, they are not taken into account when the strongest frequencies are selected.

Two subsequent processing steps are performed. First, the histogram is smoothed by a median filter of length 5 to remove outliers. The underlying assumption of this step is that isolated peaks are due to noise, is contrast with the smoother "hills" which represent true frequency components of the signal. The histogram after outlier removal is denoted H_1.

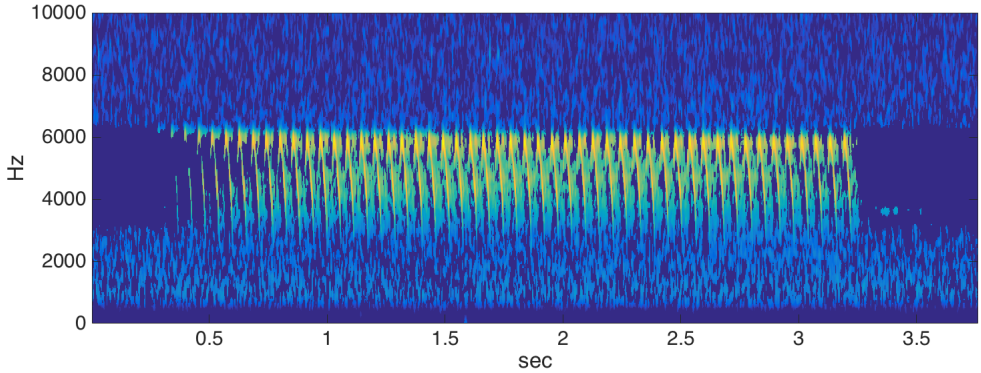
Figure 3.2: The different stages of spectrogram processing



(a) Spectrogram of a Chipping Sparrow trill



(b) Spectrogram after median filtering



(c) Spectrogram after median filtering and median clipping

Second, noise bins are removed from the histogram. noise bins can enter the histogram when audio frames which does not contain bird vocalizations are processed. To remove accumulation of noise bins, the noise floor is estimated and subtracted from the histogram.

The noise floor is estimated using a simple median filter with a wide window. The window length used is half the length of the histogram. For a window of L samples and zero padding factor Z, the histogram size is $\frac{L \times Z}{2}$ bins. Consequently, the median filter is $\frac{L \times Z}{4}$ bins long. The noise floor is denoted by $NF = \text{MEDFILT}(H_1)$.

After subtraction, the resulting histogram is

$$H = \max\{H_1 - NF, 0\} \tag{3.4}$$

which is the final histogram representation of signal frequencies. An example of the final histogram is illustrated in Fig. 3.3 (bottom).

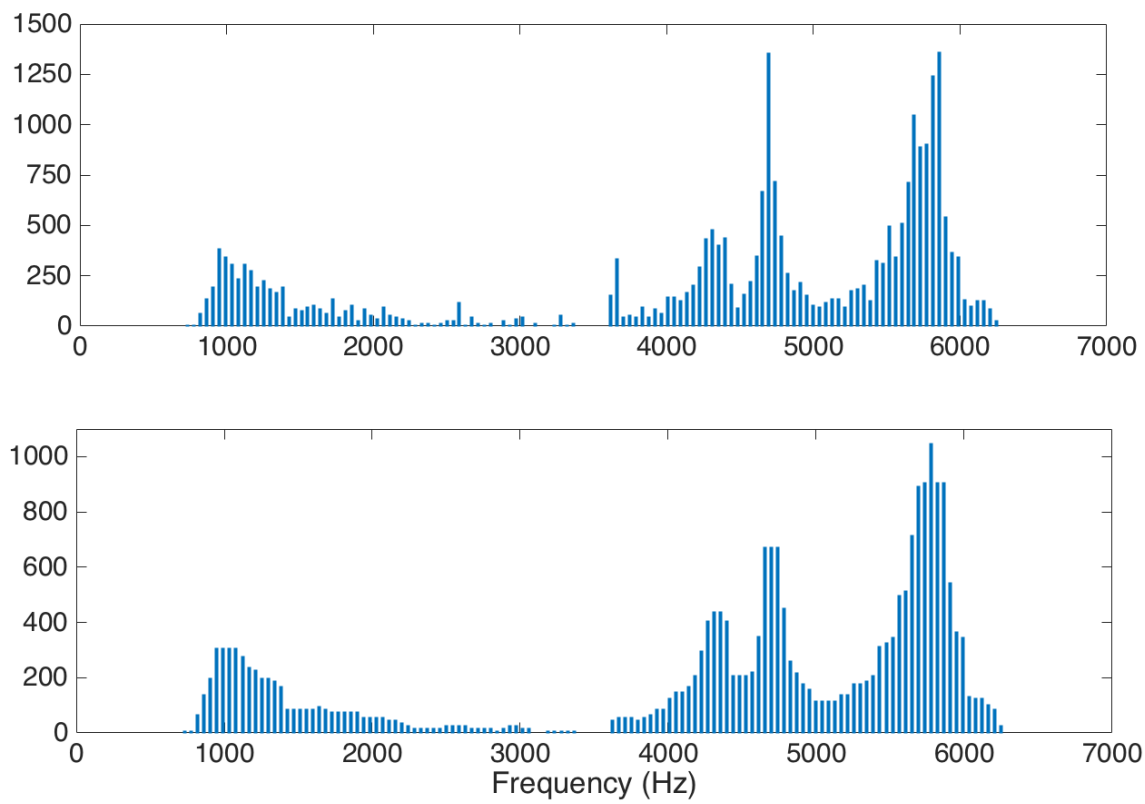


Figure 3.3: Different processing stages of the frequency histogram. Top: An example of H_0 prior to any processing. Bottom: An example H, derived from H_0 by median filtering and noise floor removal.

3.3 Frequency Band Detection

In this step of the algorithm, frequency bands are extracted from the "frequent frequencies" histogram $H(f)$ created in the previous section. Individual isolated spectral peaks are generally recognized as outliers not related to natural phenomena. Spectral lobes are favoured, and Gaussians can be used to model them. Therefore, $H(f)$ is approximated using a Gaussian mixture model. The effective support of each Gaussian is measured and used as a potential frequency band.

Formally, $H(f)$ is approximated by a function of the form:

$$M(f; \boldsymbol{a}, \boldsymbol{\mu}, \boldsymbol{\sigma}) = \sum_{i=1}^{K} a_i \exp\left(-\frac{(f-\mu_i)^2}{2\sigma_i^2}\right) \tag{3.5}$$

with K equals to the number of Gaussians in the model function and $\boldsymbol{a}, \boldsymbol{\mu}, \boldsymbol{\sigma}$ are K dimensional vectors of parameters: $\boldsymbol{a} = (a_1, \dots a_K)^T$, $\boldsymbol{\mu} = (\mu_1, \dots \mu_K)^T$, $\boldsymbol{\sigma} = (\sigma_1, \dots \sigma_K)^T$, which represent amplitude, mean and standard deviation of each Gaussian, respectively.

The approximation is carried out by solving a non-linear least squares problem:

$$\underset{\boldsymbol{a},\boldsymbol{\mu},\boldsymbol{\sigma}}{\arg\min} \|H(f) - M(f; \boldsymbol{a}, \boldsymbol{\mu}, \boldsymbol{\sigma})\|_2^2 \tag{3.6}$$

subject to constraints:

$$\mu_i > 500 \tag{3.7}$$

$$400 > \sigma_i > 100 \tag{3.8}$$

The constraint (3.7) is obvious, as frequencies below 500 Hz are assumed to be noise. The bounds in (3.8) were set because empirically it was found that a Gaussian with either very small or very large variance is inappropriate for fitting to the peaks found in trill vocalizations. As will be explained shortly, the bounds chosen for σ_i yield frequency bands of width between 500 and 2000 Hz, which are convenient to work with. Wider bands are prone to noise inclusion, and narrower bands are prone to exclusion of important information and focus on over-specific spectrum parts.

A detailed explanation of the solution method to problem (3.6) follows. The model function M can be written as:

$$M(f; \boldsymbol{a}, \boldsymbol{\mu}, \boldsymbol{\sigma}) = \boldsymbol{N}(\boldsymbol{\mu}, \boldsymbol{\sigma})^T \cdot \boldsymbol{a} \tag{3.9}$$

with $\boldsymbol{N}_i(\boldsymbol{\mu}, \boldsymbol{\sigma}) = \exp\left(-\frac{(f-\mu_i)^2}{2\sigma_i^2}\right)$ for $1 \leq i \leq K$.

$M(f; \boldsymbol{a}, \boldsymbol{\mu}, \boldsymbol{\sigma})$ is linear in the parameter $\boldsymbol{a}$, and therefore $\boldsymbol{a}$ can be estimated by solving a linear least squares problem. However, since M is not linearly separable, the linear least

squares problem has to be solved for every iteration of the nonlinear least squares solver. Therefore, (3.6) is changed to

$$\arg\min_{\boldsymbol{\mu},\boldsymbol{\sigma}} \|H(f) - M(f; \boldsymbol{a}, \boldsymbol{\mu}, \boldsymbol{\sigma})\|_2^2 \tag{3.10}$$

which is solved using the "trust region reflective" algorithm, an iterative non-linear least squares solver. Let's denote the solution approximated by the solver at iteration j by $\boldsymbol{\mu}^{(j)}, \boldsymbol{\sigma}^{(j)}$, and $\boldsymbol{a}^{(j)}$. In every iteration j, after the parameters $\boldsymbol{\mu}^{(j)}, \boldsymbol{\sigma}^{(j)}$ are updated, a subsequent update occurs for $\boldsymbol{a}^{(j)}$ by solving the linear problem:

$$\boldsymbol{a}^{(j)} = \arg\min_{\boldsymbol{a}} \|H(f) - \boldsymbol{N}(\boldsymbol{\mu}^{(j)}, \boldsymbol{\sigma}^{(j)})^T \cdot \boldsymbol{a}\|_2^2 \tag{3.11}$$

To complete the description of the optimization process, initial values are discussed. The values of $\mu_i^{(0)}$ should be set far enough apart to get coverage of the whole spectrum. Therefore, rather than choosing the K highest peaks, an iterative peak elimination process is executed. For every $1 \leq i \leq K$, denote the frequency of the highest available peak of $H(f)$ as p_i, and remove the peaks of H in the interval $[p_i - 500\,,\, p_i + 500]$. The value of $\mu_i^{(0)}$ is calculated as the mean of the frequencies of peaks in that interval which are taller than 5% of p_i. In the following iterations, the highest peak is chosen from the peaks not removed in previous iteration, ensuring that $\mu_i^{(0)}$ are chosen from disjoint intervals 1000 Hz wide. A big value of $\sigma_i^{(0)} = 1000$ Hz is set to ensure the Gaussians are as wide as the constraints allow. $\boldsymbol{a}$ doesn't need an initialization value since it is estimated using a deterministic linear least squares solver.

Once the parameters are estimated, frequency band intervals are extracted from the model using:

$$B_i = [\mu_i - 2.5\sigma_i, \mu_i + 2.5\sigma_i] \tag{3.12}$$

which contains more than 98.7% of the area under the Gaussian curve.

Therefore, the width of each B_i is between 500 Hz and 2000 Hz. There are many bird species whose trills occur over wider bandwidth. However, the applicability of the algorithm isn't necessarily diminished. Rather, the remaining frequencies in the true band should be detected by other Gaussians. The final output of the f_0 frequency band detection algorithm is a frequency band which contains several B_i intervals, as formulated in Eq. (3.37) and Eq. (3.38). The variance restriction of Eq. (3.8) is utilized only to increase analysis resolution by breaking wide frequency bands into narrower ones.

An acceptable assumption on the input signal is that it contains a discernible number of auditory phenomena. It is therefore implied that most of the content is contained in a small number of frequency intervals. Therefore, the number of Gaussians in the mixture model is initialized to $K = 3$. After the intervals B_i are estimated, a test is performed to determine whether more Gaussians should be added to the model. The area under the histogram calculated only over the estimated intervals B_i should exceed 90% of the total

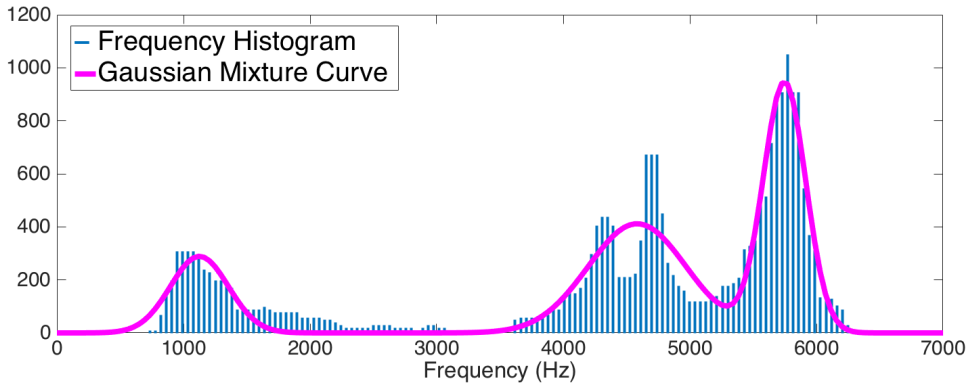


Figure 3.4: Histogram Gaussian curve fitting. 3 Gaussians are required for this example

area of the histogram. This measure of area proportionality makes sense if the histogram is interpreted as a probability density of prominent frequencies in the signal. Mathematically, the test is passed if the following holds:

$$\|H(f) \cdot \mathbb{1}_B(f)\|_1 > 0.9 \times \|H(f)\|_1 \tag{3.13}$$

where the norm used is the L_1 norm and B is the union of estimated frequency bands:

$$B = \bigcup_{i=1}^{K} B_i \tag{3.14}$$

if the test fails, K is raised to 4 and the optimization process is repeated. In principle, there is no limitation on the number K. However, under the assumptions on the frequency range and number of auditory phenomena present in the signal, it is reasonable to stop the process after $K = 4$ whether (3.13) is satisfied or not.

3.4 Filtering and Short-Time Teager Energy

In the previous section, up to 4 frequency bands were extracted, denoted B_i, with $1 \leq i \leq K$. In this section, the input signal $s[n]$ is filtered according to each frequency band B_i defined in (3.12), and a short-time energy function is calculated from each filtered signal. In the following sections, the energy functions of each candidate band will be compared in order to determine if a trill resides in them.

The filter used is a Chebyshev II filter because it is simple, yet manifests shorter transition bands than Butterworth filters. In contrast with Chebyshev I design, it is maximally flat in the pass band, thus causing less signal distortion. The filter is designed with pa-

rameters $R_p = 0.05$ db and $R_s = 40$ db, which signify maximum passband and stopband ripple, respectively. The resulting K filtered signals are denoted by $s_i(t)$, $1 \leq i \leq K$.

After filtering, short-time energy is calculated. The energy function used by the algorithm is the Teager energy operator. The definition of the discrete Teager energy operator is

$$\Psi[s[n]] = s^2[n] - s[n+1] \cdot s[n-1] \tag{3.15}$$

Teager energy operator is used for energy calculations because it encompasses both amplitude and frequency information (see [25]), and therefore can be better suited to detect signals which have declining energy but rising frequency.

The operator Ψ is applied to each of the bandpass signals $s_i[n]$. For the purpose of short-time energy computation, the operator output is filtered by a moving average low pass filter with impulse response $h[n]$ (see below). The short-time energy signal of each bandpass signal $s_i[n]$ is denoted

$$\Psi_i[n] = \Psi[s_i[n]] * h[n] \tag{3.16}$$

$\Psi[s[n]]$ is filtered using a Gaussian smoothing filter. Denote by $h_0[n]$ the impulse response of a Gaussian smoothing filter with an effective length (i.e. 5 standard deviations) of 20 ms. $h_0[n]$ is therefore a normalized Gaussian with $\sigma = L/5$, and $L = 2 \cdot 10^{-2}$ seconds. In discrete implementation $L = \lfloor 2 \cdot 10^{-2} \cdot f_s \rfloor$, and $\sigma = (L-1)/5$ samples.

To counter phase delay, and since causality is not a concern, zero phase filtering is implemented using time reversal. This process is common in non-causal IIR filtering where zero phase is desired. Symmetric FIR filters have linear phase response, and can be made zero phase by simply removing the delay which is introduced by the filter uniformly on all frequencies. Nevertheless, time reversal is favoured here for analysis. The filtering process is executed by filtering $\Psi[s_i]$ with h_0, time-reversing the output and filtering by h_0 again. Lastly, the output is time-reversed again to obtain the original time order. This process defined the output of the zero phase low pass filter $h[n]$.

Alternatively, since $h_0[n]$ is a FIR filter, $h[n]$ can be realized by the equation

$$h[n] = h_0[n] * h_0[-n] \tag{3.17}$$

For ease of presentation, analysis of $h[n]$ is performed in the continuous domain for $h(t) = h_0(t) * h_0(-t)$. Discarding normalization constants, we can write

$$h_0(t) = e^{-\frac{(t-\mu)^2}{2\sigma^2}} \tag{3.18}$$

$$h_0(-t) = e^{-\frac{(t+\mu)^2}{2\sigma^2}} \tag{3.19}$$

Since the Fourier transform of a Gaussian is another Gaussian:

$$\mathcal{F}\{e^{-\frac{t^2}{2\sigma^2}}\}(\omega) = \sqrt{2\pi}\sigma e^{-\frac{\sigma^2\omega^2}{2}} \tag{3.20}$$

and using the Fourier time shift theorem, we have

$$\mathcal{F}\{h_0(t)\}(\omega) = e^{-i\mu\omega}\sqrt{2\pi}\sigma e^{-\frac{\sigma^2\omega^2}{2}} \tag{3.21}$$

$$\mathcal{F}\{h_0(-t)\}(\omega) = e^{i\mu\omega}\sqrt{2\pi}\sigma e^{-\frac{\sigma^2\omega^2}{2}} \tag{3.22}$$

Using the Fourier convolution theorem, it follows that the frequency response of the concatenated filter is a Gaussian with zero phase and standard deviation of $\frac{1}{\sqrt{2}\sigma}$ radians/sec:

$$H(\omega) = \mathcal{F}\{h_0(t)\}(\omega) \cdot \mathcal{F}\{h_0(-t)\}(\omega) = 2\pi\sigma^2 e^{-\frac{\omega^2}{2\left(\frac{1}{\sqrt{2}\sigma}\right)^2}} \tag{3.23}$$

Neglecting the multiplicative constant, and since $H(\omega)$ is Gaussian, it follows from Eq. (3.20) that $h(t)$ is also Gaussian, with standard deviation of $\sqrt{2}\sigma$ secs. This analysis will be useful in resampling the energy signal in section 3.6. Note that the standard deviation of (3.23) in ordinary frequencies is $\sigma' = \frac{1}{2\sqrt{2}\pi\sigma}$ Hz.

3.5 Low Energy Band Elimination

Since the signal is assumed to contain a trill, low energy bands B_i are assumed to contain mostly noise and can be discarded. This section presents a simple method for comparison between the total band energy of several frequency bands. First, the highest energy band is found. A threshold is calculated using the energy of this band, and the energies of the other bands are compared to it. Bands which fail to meet this requirement are eliminated.

Energy is calculated here with median subtraction for noise floor normalization, to counter white noise effects. For comparison purposes, the energy of band B_i is defined as:

$$E_i = \sum_n \max\{\Psi_i[n] - M_i\,,\, 0\} \tag{3.24}$$

with M_i denoting the median of $\Psi_i[n]$, and the sum going over all sample indices of $\Psi_i[n]$. The maximal energy band is the band B_ι for which $\iota = \arg\max_i E_i$. Energy difference relative to the maximal energy E_ι is calculated in DB:

$$\Delta E_i = 10\log\left(\frac{E_i}{E_\iota}\right) \tag{3.25}$$

A threshold of -13db is set so that if $\Delta E_i < -13$db, the frequency band B_i is discarded as noise. The value of -13db was chosen experimentally. It was found to be low enough to retain bands which contain signal information, yet high enough to exclude most noise bands.

In some cases, the bands B_i obtained in (3.12) are overlapping. If most of the information of one band is contained in another, it should be removed to avoid redundancy. Overlapping bands can be eliminated by calculating overlap percentage, defined for $1 \leq i, j \leq K$ as:

$$O_{i.j} = \frac{|B_i \cap B_j|}{|B_j|} \tag{3.26}$$

A threshold T_O can be set such that if $O_{i,j} > T_O$, and $E_i > E_j$, B_j can be discarded. A value of $T_O = 80\%$ was experimentally found sufficient for determining redundancy of overlapping bands.

3.6 Short-Time Energy Autocorrelation Method

In order to determine if a frequency band contains a trill, a discernible phenomena of pulsatile and often quasi-periodic nature should be detected in $\Psi_i[n]$, the short-time energy of each of the bandpass signals. The short-time energy autocorrelation algorithm for syllable presence estimation, which is presented in section 2.3, can be used to detect trill presence in each subband. The algorithm outputs a sequence of time instances which represent the presence of estimated syllables. Low variance in the inter-time interval of these syllables can be a measure for pulsatility, which is a good indicator for trill presence.

However, this process is computationally intensive, as the autocorrelation is calculated with $O(n^2)$ operations. Using the Fourier transform, this can be reduced to only $O(n \log n)$ operations, however, due to the unusual dependency of the window size on the timelag τ in Eq. (2.7), a more sophisticated approach should be devised. Alternatively, recalling that the energy bands are lowpass filtered in Sec. 3.4, $\Psi_i[n]$ can be downsampled first to reduce the computational overhead.

Recall from (3.23) that the Gaussian smoother frequency response $H(f) = H(\frac{\omega}{2\pi})$ has standard deviation equal to $\sigma' = \frac{5}{2\sqrt{2}\pi L}$ Hz, with $L = 2 * 10^{-2}$ seconds. In principle, setting the Nyquist frequency at 5 standard deviations should be enough to retain all the energy signal information. However, since time resolution is still important for autocorrelation result accuracy, this factor is further multiplied by 8. Therefore, each $\Psi_i[n]$ can be downsampled to:

$$F_s = 2 \cdot 5 \cdot 8 \cdot \sigma' \approx 2.25 \text{ kHz} \tag{3.27}$$

which is a 20-fold improvement compared to the standard sample rate of 44.1 kHz.

After resampling, each $\Psi_i[n]$ is passed to the short-time energy autocorrelation algorithm, to produce a sequence of time instances $\tau_0^{(i)}, \ldots, \tau_{J_i}^{(i)}$, which will be further analyzed to determine the level of pulsatility in the subband.

3.7 Decision Parameters

Various parameters can be extracted from the K sequences of time instances $\tau_0^{(i)}, \ldots, \tau_{J_i}^{(i)}$ obtained by passing Ψ_i to the short-time energy autocorrelation algorithm. These parameters can be used to determine trill presence in the corresponding subband. The parameters calculated are the period mean, period standard deviation and the energy mean variation, defined below.

For every time instance sequence $\{\tau_j^{(i)}\}$, define the difference sequence $\{\delta_j^{(i)}\}$ by

$$\delta_j^{(i)} = \tau_j^{(i)} - \tau_{j-1}^{(i)} \ , \ \ 1 \leq j \leq J_i \ , \ \ 1 \leq i \leq K \tag{3.28}$$

Since outliers may be present in $\delta_j^{(i)}$, especially in high noise signals due to false positives, pseudo interquartile sequences are calculated

$$\Delta^{(i)} = \{\delta_j^{(i)} | \delta_j^{(i)} > P_{40}^{(i)} \wedge \delta_j^{(i)} < P_{95}^{(i)}\} \tag{3.29}$$

$$T^{(i)} = \{\tau_j^{(i)} | \delta_j^{(i)} \in \Delta^{(i)}\} \tag{3.30}$$

P_{40} and P_{95} are the 40th and 95th percentiles of $\delta^{(i)}$, respectively. The asymmetry is introduced to remove bias caused by noise effects on the energy normalized correlation, as explained in 2.3.

$\Delta^{(i)}$ and $T^{(i)}$ can be viewed as subsequences of $\delta^{(i)}$ and $\tau^{(i)}$, respectively, of length Λ_i, so that $1 \leq \Lambda_i \leq J_i$. The differences between $T^{(i)}$ and $\tau^{(i)}$ are illustrated in Fig. 3.5.

The following parameters are extracted:

1. Period mean,

$$\mu^{(i)} = \frac{1}{\Lambda_i} \sum_{j=1}^{\Lambda_i} \Delta_j^{(i)}$$

2. Period standard deviation,

$$\sigma^{(i)} = \sqrt{\frac{1}{\Lambda_i - 1} \sum_{j=1}^{\Lambda_i} (\Delta_j^{(i)} - \mu^{(i)})^2}$$

3. Energy mean variation, $\chi^{(i)}$, which is the mean of the energy difference between every point $T_j^{(i)} \in T^{(i)}$ to the absolute minimum between it and every nearest neighbour it has in the sequence, $T_{j-1}^{(i)}$ and $T_{j+1}^{(i)}$. Denoting for every $1 \leq i \leq K$ and every $1 \leq j \leq \Lambda_i$:

$$m_j^{(i)} = \min_{T_{j-1}^{(i)} < n < T_j^{(i)}} \Psi_i[n] \tag{3.31}$$

The Energy mean variation can therefore be defined as:

$$\chi^{(i)} = \frac{1}{2\Lambda_i} \left(\sum_{j=1}^{\Lambda_i} |\Psi_i[T_j^{(i)}] - m_j^{(i)}| + \sum_{j=1}^{\Lambda_i} |\Psi_i[T_{j-1}^{(i)}] - m_j^{(i)}| \right) \tag{3.32}$$

An illustration is provided in Fig. 3.5.

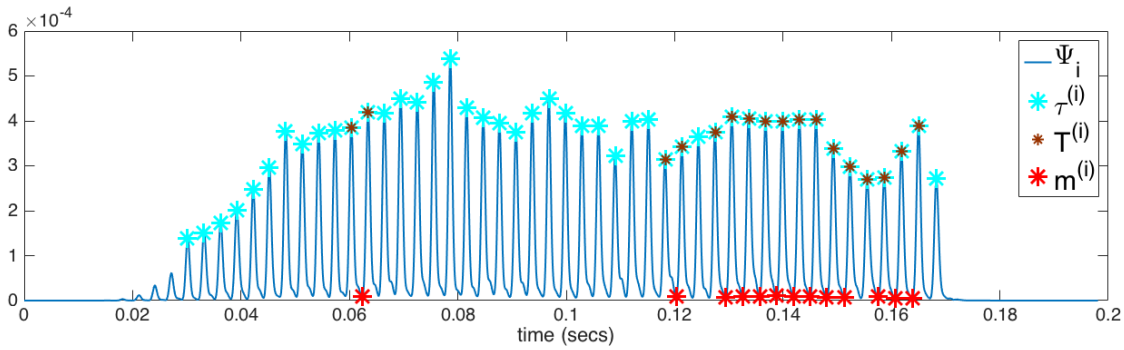


Figure 3.5: A short-time energy obtained from a sub-band signal of a Chipping Sparrow trill. The cyan asterisks represent the peaks $\tau_j^{(i)}$ obtained by the STEAC method. The brown asterisks represent only those peaks whose intervals to their consecutive peaks manifest $P_{40} < \Delta_j^{(i)} < P_{95}$. They are denoted as $T_j^{(i)}$. The red asterisks represent the absolute minimum points $m_j^{(i)}$, computed only between adjacent peaks $T_j^{(i)}$.

3.8 Frequency Band Estimation

The parameters calculated in the previous subsection are used to estimate a frequency band for the trill. Band estimation and elimination is performed by identifying the *best band* and comparing it to the rest of the bands.

The *best band* is set as the band with the maximum energy mean variation:

$$\tilde{B} = B_I \text{ s.t. } I = \arg\max_i \chi^{(i)} \tag{3.33}$$

This parameter was found to be the most decisive for trill presence estimation in the sub band. The STEAC process performed in Sec. 3.6 ensures that the time instance series

$T^{(i)}$ already manifest some sort of periodicity. In order to differentiate between trill bands and noise bands, the amount of variation between loud and quiet signal parts, which is quantified in $\chi^{(i)}$, should be examined. The parameters of the *best band* are denoted by $\tilde{\mu}, \tilde{\sigma}, \tilde{\chi}$.

The other bands are accepted or rejected depending on the measure of resemblance of their parameters to those of $\tilde{B}$. Thresholds are set for acceptable deviation of each of the 3 parameters from the corresponding parameters of the *best band*, and are denoted by T_μ, T_σ, T_χ. A band B_i is labeled as a *good band* if all of the following conditions are met:

$$|\mu^{(i)} - \tilde{\mu}| < T_\mu \tag{3.34}$$

$$\frac{\sigma^{(i)}}{\tilde{\sigma}} < T_\sigma \tag{3.35}$$

$$\frac{\chi^{(i)}}{\tilde{\chi}} < T_\chi \tag{3.36}$$

The threshold values were heuristically set to: $T_\mu = 10^{-2}, T_\sigma = 15, T_\chi = 10^{-1}$.

The estimated band $\hat{B}$ is set as the shortest interval containing all *good bands*. If the frequency bands are written explicitly as $B_i = [f_1^{(i)}, f_2^{(i)}]$, then the estimated band is

$$\hat{B} = [\min_i f_1^{(i)} \ , \ \max_i f_2^{(i)}] \tag{3.37}$$

with index i going over all *good bands* only.

Finally, the margins of the estimated band are widened by a fixed number of $f_M =$ 300Hz, yielding a final estimate

$$\hat{B} = [\min_i f_1^{(i)} - f_M \ , \ \max_i f_2^{(i)} + f_M] \tag{3.38}$$

where $\hat{B}$ is an estimate of the frequency band of the fundamental frequency of the signal. The signal is filtered to accordingly, using a bandpass filter with cutoff frequencies equal to the upper and lower bounds of $\hat{B}$. The lower bound is used as an input parameter for fundamental frequency estimation with the YIN algorithm.

3.9 Algorithm Evaluation

For evaluation, a dataset of 48 trill recordings belonging to 21 different bird species was assembled. The recordings were either downloaded from the Xeno-canto website, or recorded in the Hula lake. All tracks were sampled at 44.1kHz. The only editing done was cutting a single trill from each track. Other than that, the tracks are unedited and are fair representatives of recordings encountered in practice. The frequency band of interest in each track was identified manually through spectrogram inspection, and subsequently set as benchmark for evaluation. Spectrograms were computed using analysis frame of 1024 samples,

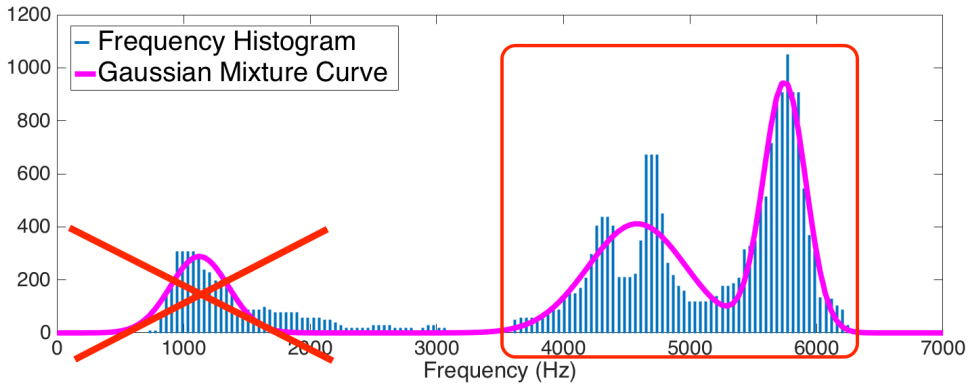


Figure 3.6: Frequency band estimation. The lower band is discarded because of low energy compared to the best band. The higher bands are both accepted because the corresponding parameters $\mu^{(i)}, \sigma^{(i)}, \chi^{(i)}$ are close.

multiplied by a Hamming window with 50% overlap. Center frequency and bandwidth were computed as the middle frequency and the difference between maximal and minimal frequency of each benchmark frequency interval, respectively. The mean bandwidth is 1.71 kHz and the mean center frequency is 3.34 kHz.

The "ground truth" values were compared to the frequency bands obtained by the algorithm described in this section. For each track, the algorithm output was recorded, and evaluation errors were calculated. Suppose the ground truth interval of a specific track is $B = [f_1, f_2]$ and the corresponding estimated interval is $\hat{B} = [\hat{f}_1, \hat{f}_2]$. Evaluation errors were calculated as:

$$E_T = f_2 - \hat{f}_2 \tag{3.39}$$

$$E_B = \hat{f}_1 - f_1 \tag{3.40}$$

where E_T is the error at the top edge and E_B is the error at the bottom edge of the frequency interval. Calculating the errors in this was has the benefit of uniform interpretation: positive errors are interpreted as inclusion of the relevant ground truth edge in the estimated interval, both at the top or bottom edge.

For each track, errors are presented at both upper and lower edge of the frequency interval. Positive top edge errors mean the estimated value is greater than the ground truth value, while positive bottom edge errors mean the estimated value is lower than the ground truth value. For the purposes of our research, positive estimation errors of under 500 Hz are not considered errors. Positive errors under 1 kHz are still tolerable. Negative errors cause information loss after bandpass filtering. However, in most cases errors above -500 Hz do not cause large information loss. We therefore use the term *positive miss* for positive errors above 1 kHz and *negative miss* for negative errors below -500 Hz. As a

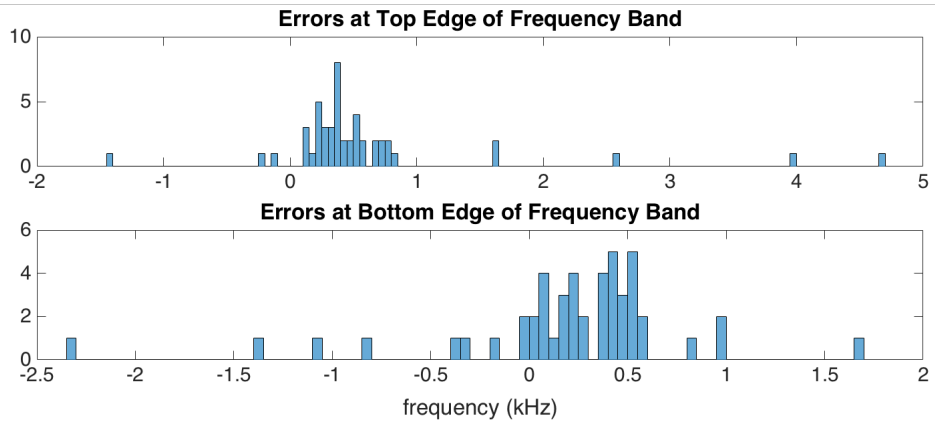


Figure 3.7: Frequency band of interest evaluation. Both histograms present nominal evaluation errors of the edges of the ground truth frequency interval. Positive errors mean the estimated interval contains the ground truth interval near the appropriate edge. They are generally preferred over negative errors, in which the estimated interval is contained in the ground truth interval, since subsequent band pass filtering results in information loss. Top: a histogram of the upper edge evaluation errors. Most errors reside in the error interval $[0, 1000]$ Hz. Bottom: a histogram of the lower edge evaluation errors. Most errors reside in the error interval $[0, 600]$ Hz.

result, for bottom edge estimation, there are 4 negative misses and 1 positive miss, which stand for 8.3% and 2.1% of the estimates. For the top edge, there is 1 negative miss and 4 positive misses, which stand for 2.1% and 8.3% of the estimates (see Fig. 3.7). We conclude that for most trills, the algorithm finds the range in which f_0 resides quite accurately.

Chapter 4

Segmentation Algorithms

In this chapter, two segmentation algorithms are presented. Both algorithms assume that the frequency band of the trill fundamental frequency is either known or estimated with the algorithm presented in chapter 3. The algorithms operate in three general steps: in the first step, the trill signal is bandpass filtered according to the fundamental frequency band. In the second step, the trill syllables are detected. In the last step, the segments containing each syllable are found.

The first algorithm uses the YIN algorithm for fundamental frequency estimation. In contrast, the second algorithm doesn't use fundamental frequency. Instead, it exploits the self correlation section of the frequency band estimator described in chapter 3 and apply it to syllable detection.

4.1 Algorithm FVDNEM

The algorithm's syllable detection mechanism uses the signal's frequency data as well energy data. Syllables are initially detected using dips in a frequency stability measure (described below). They are subsequently verified using comparison to the envelope of the signal's energy. A schematic block diagram of the method is depicted in Figure 4.1. Each of the stages is described and referred below. The algorithm's acronym FVDNEM refers to its key steps: the Fundamental Frequency Variance Derivative, used as a stability measure for syllable recognition and the Neighbourhood Energy Maximization, used to further differentiate syllables and inter-syllables.

4.1.1 Short-Time Variance of f_0 Derivative

In most harmonic trills the fundamental frequency f_0 within syllables changes relatively smoothly, whereas inter-syllables may exhibit less stable f_0 since the energy invested in sound production temporarily decreases, producing no (or a weak) f_0 component. Thus,

the estimated f_0 in inter-syllables or in segments that precede or succeed the trill is characterized by high variations and with relatively abrupt changes. Therefore, for initial detection of syllables, we used the short-time variance of the derivative of the fundamental frequency which is defined as follows:

$$f_{vd}(\bar{t}) = Var\left[\frac{d}{dt}\hat{f}_0(t)|_{t\in S_{\bar{t}}}\right] \tag{4.1}$$

where $\hat{f}_0(t)$ is the estimated fundamental frequency of the original signal, and the set $S_{\bar{t}} = \{t||t-\bar{t}| < \frac{L}{2}\}$ is the short-time frame. The values of $f_{vd}(\bar{t})$ are calculated for each frame of L samples centered at t, using the following steps:

1. The fundamental frequency estimation $\hat{f}_0(t)$ is computed using the YIN algorithm [21], at discrete time instances $\{t_n\}_{n=1}^{K}$, usually with a fixed step size S. A version of YIN, adapted and tuned for $f_0(t)$ of birds was implemented, since it was found to be most adequate for determination of f_0 in birds [22]. The algorithm (see section 2.2) follows a time domain approach using a normalized form of the difference function $d(\tau) = \sum_{n=1}^{W}(s[n]-s[n+\tau])^2$ where $s[n]$ designates the discrete signal. For quasi-periodic signals, a local minimum in $d(\tau)$ is expected, with a lag value of the fundamental period, i.e., the abscissa τ of that minimum is considered as an estimate of the period (see [21]). The local dip is then quadratically interpolated to achieve sub-sampling resolution.

2. A forward difference scheme is used for derivative approximation:

$$\frac{d}{dt}\hat{f}_0(t_n) \approx \frac{1}{S}(\hat{f}_0(t_{n+1}) - \hat{f}_0(t_n)) \tag{4.2}$$

3. Suppose $\tilde{K}$ is the number of time instances inside a frame of length L centered around t_n, i.e., it is the size of the set $S_k = \{t_k||t_k - t_n| < \frac{L}{2}\}$. A sample variance is used to calculate the variance of $\frac{d}{dt}\hat{f}_0(t)$:

$$\frac{1}{\tilde{K}-1}\sum_{k\in S_k}(\frac{d}{dt}\hat{f}_0(t_k) - \mu)^2 \tag{4.3}$$

   where $\mu = \frac{1}{\tilde{K}}\sum_{k\in S_k}\frac{d}{dt}\hat{f}_0(t_k)$ is the non-causal moving average of $\frac{d}{dt}\hat{f}_0(t)$.

The frame length L is chosen adaptively as a certain percentage of the Syllable Repetition Interval (SRI) of the trill. The SRI is defined here as the average time interval between consecutive trill syllables, and is estimated using the Discrete Fourier Transform (DFT) of the signal's energy $E(t)$. The first non-trivial (non DC) peak of the DFT is used for estimation:

$$\text{SRI} \approx \left(\frac{\omega_1}{2\pi}f_s\right)^{-1} \tag{4.4}$$

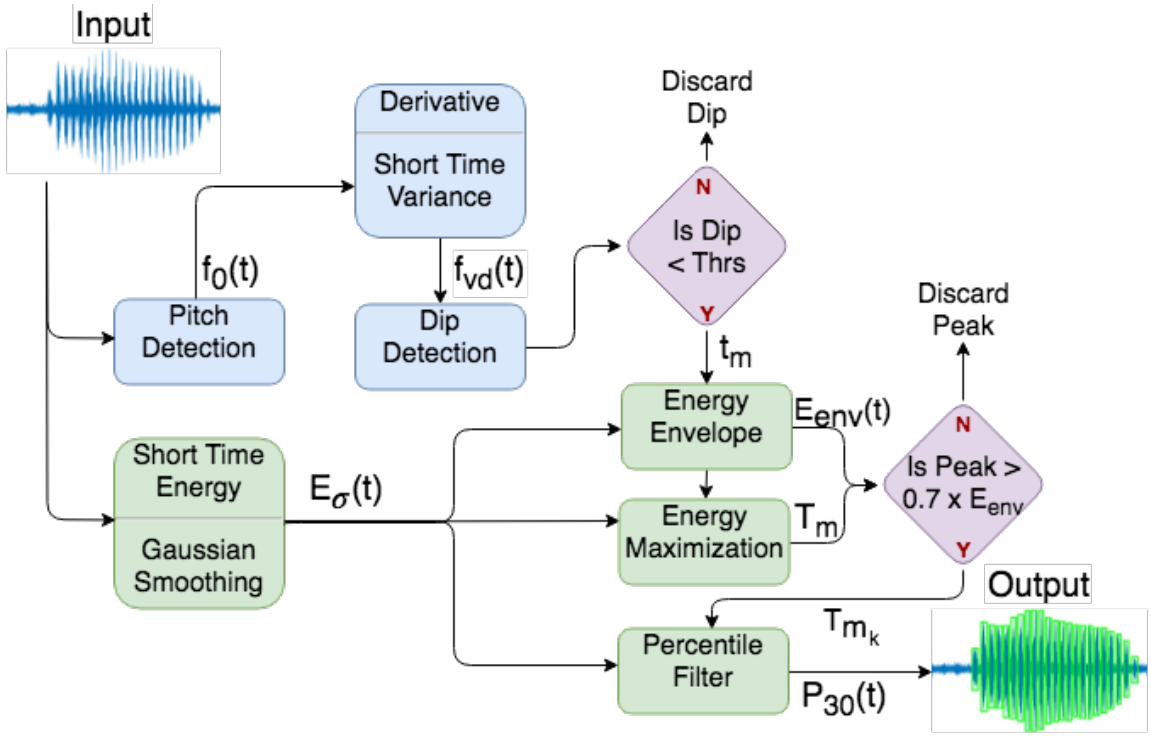


Figure 4.1: A schematic block diagram of the FVDNEM algorithm

where ω_1 is the angular frequency corresponding to the peak of the DFT. Finally, $L = 0.2 \times \text{SRI}$ is set to ensure it does not exceed syllable duration. Therefore, local minima of $f_{vd}(t)$ are expected to be found in segments within syllables, and thus $f_{vd}(t)$ can be used as an indicator for the presence of syllables. However, in order to exclude local $f_{vd}(t)$ minima of high values (i.e., high variance points), an adaptive threshold is computed, based on f_{vd} in background noise frames adjacent to the trill. As can be seen in Fig. 4.2 (top part), dips with variance higher than the threshold are discarded.

4.1.2 Accuracy Improvement with Energy Functions

In most cases dips in $f_{vd}(t)$ indicate the presence of a syllable. However, since dips do not generally coincide with peaks in the short-time energy function of the trill signal, and may lie outside the syllable due to time offset of $\hat{f}_0(t)$, it is important to align the dips with the corresponding energy peaks. Moreover, even if a dip indeed lies inside a syllable, but fails to correspond to an energy peak, the bulk of syllable energy may be overlooked in segmentation step, thereby contributing to a partial syllable detection only.

Accuracy improvement involves replacing each syllable indicator (i.e., a dip of $f_{vd}(t)$)

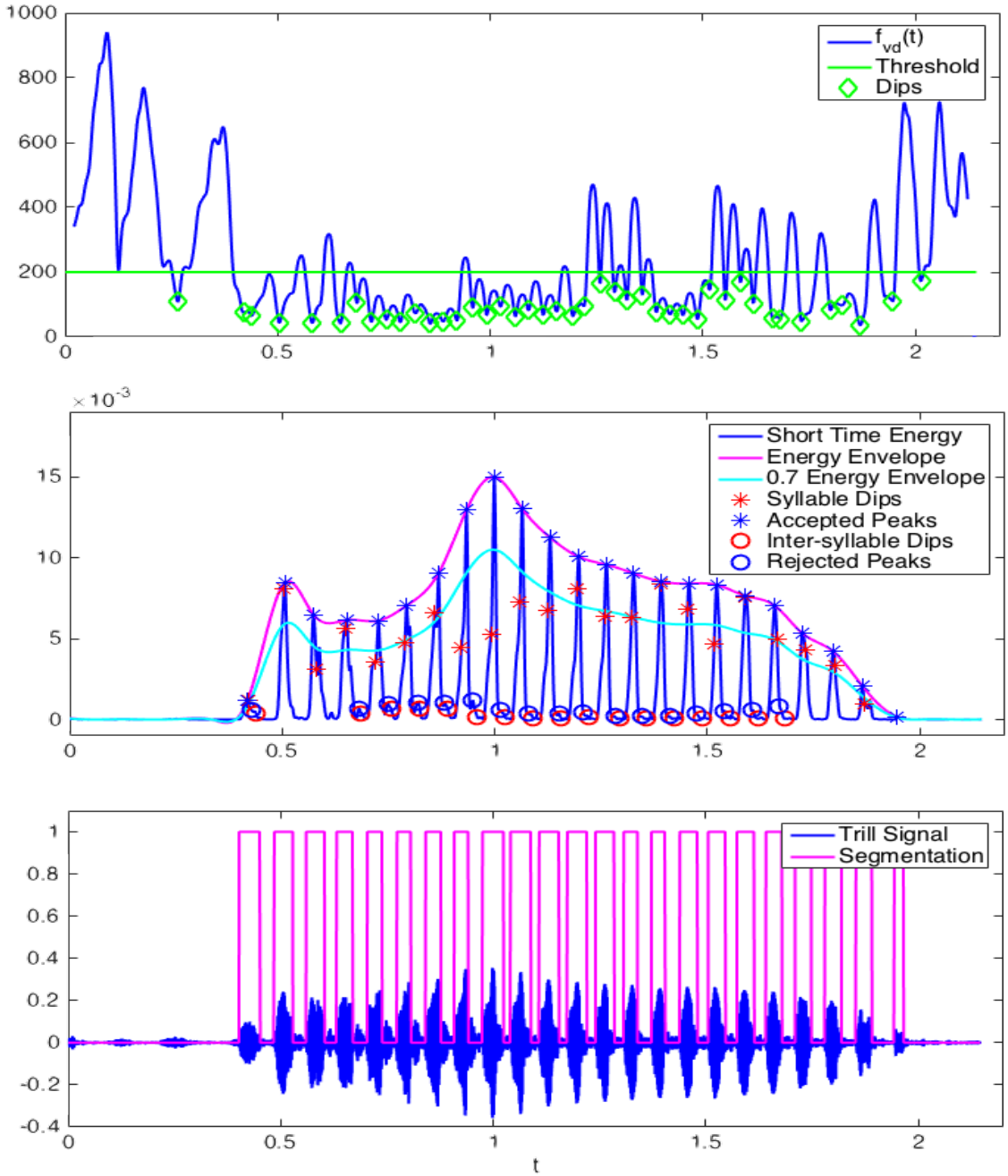


Figure 4.2: Algorithm stages. Top: The short-time variance of f_0 derivative, $f_{vd}(t)$ (blue line), with its dips (green circles). Middle: Fine-tuning of peaks by energy maximization. Bottom: Final Segmentation (Output of Alg. 1)

by the point with maximum energy in a small neighbourhood around it. Formally, set parameter $\varepsilon > 0$, which represents the search range for an energy peak in the proximity of the $f_{vd}(t)$ dip. Suppose $\{t_l\}_{l=1}^{L}$ is a sequence of dips obtained from $f_{vd}(t)$. Construct a new sequence $\{T_l\}_{l=1}^{L}$ using the rule:

$$T_l = \underset{t \in [t_l - \varepsilon, t_l + \varepsilon]}{\arg\max} \; E_\sigma(t) \tag{4.5}$$

where $E_\sigma(t)$ is a Gaussian smoothed TEO function of $s(t)$, which is calculated using the discrete definition:

$$E[n] = s[n]^2 - s[n-1] \cdot s[n+1] \tag{4.6}$$

where $s[n]$ denotes the sampled version of the signal $s(t)$.

4.1.3 Inter-Syllables Exclusion

Occasionally, dips in f_{vd} may also indicate the presence of inter-syllables with harmonic components. Normally, these inter-syllables should be eliminated prior to further analysis. An adaptive threshold, based on the energy envelope $E_{env}(t)$ is used for their exclusion. The energy envelope is generated using spline interpolation between the global maxima of $E(t)$ in each syllable. To ensure that only one maximum from each syllable is used, the detected maxima are separated by at least $0.8 \times SRI$ apart, using equation (4.4). A down-scaled version of E_{env} is used to create an adaptive threshold based on E_σ, which is used to discard syllable indicators T_n which were falsely detected in previous steps. Let p be a scalar parameter, $0.5 < p < 1$, that represents the rate of acceptable deviation of $E_\sigma(T_n)$ from energy envelope E_{env}. A new subsequence of indicators can be extracted from $\{T_n\}_{n=1}^{N}$ by selecting only the members T_{n_k} so that

$$E_\sigma(T_{n_k}) > p \times E_{env}(T_{n_k}) \tag{4.7}$$

Empirically, it was found that $p = 0.7$ should be used for most trill signals. Following this step the final sequence of trill indicators $\{T_{n_k}\}_{n_k=1}^{K}$ is obtained.

4.1.4 Interval Segmentation

Each point in the time series T_{n_k} is assumed to be within an interval which contains a syllable. To complete the segmentation, the boundaries of each syllable should be found. To this end, a simple adaptive lower threshold is calculated by percentile filtering of E_σ. Define $P_p(t)$ as the p'th percentile of $E_\sigma(t)$ in the domain $[t - L/2, t + L/2]$. It was found empirically that $p = 30$ and $L = 5$ms yield good results.

Interval syllable segmentation is carried out by finding for each T_{n_k} the unique interval I_k for which $T_{n_k} \in I_k$ and

$$\forall t \in I_k \quad E_\sigma(t) > P_{30}(t) \tag{4.8}$$

Consequently, the final interval segmentation is achieved using a simple descent algorithm as described in Alg. 1.

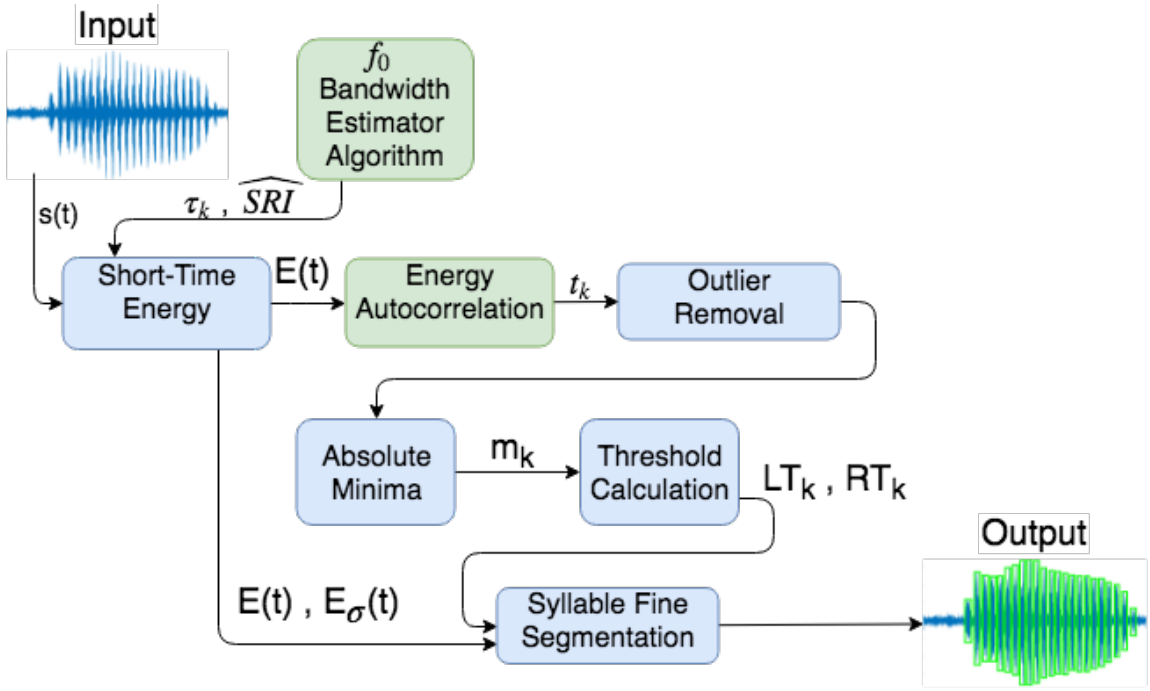


Figure 4.3: A schematic block diagram of the SACATS algorithm

4.2 Algorithm SACATS

The algorithm presented in 2.3 finds periodic energy bursts in signals where such bursts are present. The criterion used for burst detection is periodicity in the signal short-time energy. Midpoints of the correlated segments are treated as syllable centers. Neighbourhood maximization contributes to this interpretation by disallowing accumulation of small errors which creates offsets. It is therefore reasonable to use this algorithm for syllable detection as well. The algorithm acronym is derived from two of its key components: the Short-time energy Auto-Correlation and the Adaptive Thresholds used in the Segmentation phase.

4.2.1 Using Energy Autocorrelation for Syllable Detection

The energy autocorrelation algorithm of section 2.3 requires a short time energy function of a signal as input. The time duration window of the short time energy should be set large enough to filter out high frequency noise, but small enough to avoid filtering out syllable information in the signal. Window duration of L=SRI/2 (see section 4.1.1) was found to fit these conditions, since it assures that the syllable structure of the trill is kept intact, while scaling adaptively for any trill used as input.

To find an estimate for the trill SRI, the output of the algorithm for frequency band

estimation presented in chapter 3 is used. The algorithm returns a frequency band B and a sequence of approximate syllable time instances $\{\tau_k\}_{k=1}^{\tilde{K}}$. Since SRI can be loosely defined as the mean or median interval between consecutive syllables in the trill, it can be estimated using

$$\widehat{SRI} = \text{MEDIAN}\{\tau_{k+1} - \tau_k\}_{k=1}^{\tilde{K}-1} \tag{4.9}$$

A short time energy function is calculated by averaging the squared signal $s^2(t)$ with a normalized Hamming window $w(t)$ of length $L = 0.5 \times \widehat{SRI}$.

$$E(t) = s^2(t) * w(t) \tag{4.10}$$

Finally, The STEACF algorithm from Sec. 2.3 is applied to $E(t)$. the output sequence of time instances $\{t_k\}_{k=1}^{K}$ is an estimate of the maximal energy time instances of unique syllables present in the trill signal. Uniqueness is understood in the sense that if t_m and t_n belong to the same syllable then necessarily $m = n$.

In the discrete implementation, $E[n]$, which is obtained by sampling $E(t)$, is subsampled to reduce computational overhead. Considerations identical to those of section 3.6 yield a Nyquist rate of $\frac{30}{L}$ Hz (or $30\frac{2\pi}{L}$ rads/sec.)

Notice that the first side lobe of the Hamming window occurs at $4\frac{2\pi}{L}$ rads/sec and it is at -42.9 db compared to the peak of the window transform. Since the spectral rolloff is of -6 db/octave, the excess is of $\log_2(\frac{30}{4}) \approx 3$ octaves, the window attenuation is -61db at the Nyquist rate. The subsampled energy signal is therefore not distorted by aliasing.

4.2.2 ZCR and Energy Outlier Removal

Since the autocorrelation algorithm in section 2.3 uses a permissive threshold for detection, time instances at the trill's edge may be erroneously detected. Zero crossing rate comparison, combined with energy comparison is a quick and efficient way to eliminate such falsely detected syllables. Since ZCR reflects the dominant frequency in the signal, the comparison is carried out to check whether the events or bursts at the margin are similar to those at the more typical parts of the trill. The zero crossing rate is computed at each time instance t_k in a neighbourhood of $\frac{5}{f_L}$ seconds, where f_L represents the lowest harmonic frequency present in the signal. Therefore, the t_k neighbourhood duration equals 5 periods of that frequency. Denote the sequence of ZCR values as $z_1, \ldots, z_K$, and its 25th and 75th percentiles as Q_1 and Q_3, respectively.

Since energy should take a part in determining outliers as well, denote by M_E the median of the energy of all detected peaks, i.e.

$$M_E = \text{MEDIAN}\{E(t_1), E(t_2), \ldots E(t_K)\} \tag{4.11}$$

Denote by z_L the ZCR at the left margin, i.e. the first 0.1 seconds of the signal. For $k = 1 \ldots K$, the syllable at t_k is recognized as an outlier and removed if the following

condition holds:

$$[(z_L < Q_1 \text{ AND } z_k < z_L) \text{ OR } (z_L > Q_3 \text{ AND } z_k > z_L)] \text{ AND } (E(t_k) < 0.1M_E) \quad (4.12)$$

In other words, if z_L is outside the interquarile range, and z_k is bounded by it in the appropriate direction, and in addition an energy threshold is met, then t_k is recognized as an outlier and removed. However, if t_k is not an outlier, the process stops and all time instances of index $k+1, k+2, \ldots, K$ are not tested.

The same test is applied from the other side, i.e. z_R is calculated at the right margin, and the condition

$$[(z_R < Q_1 \text{ AND } z_k < z_R) \text{ OR } (z_R > Q_3 \text{ AND } z_k > z_R)] \text{ AND } (E(t_k) < 0.1M_E) \quad (4.13)$$

is tested for $k = K, K-1, \ldots, 1$. Again, the process stops after the first time t_k is recognized as a legitimate syllable.

4.2.3 Segmentation

For an accurate detection of the boundaries of the syllable S_k of each time instance t_k, an adaptive threshold is used, based on local energy levels.

Time instances $\{t_k\}$ represent energy peaks by virtue of neighbourhood maximization. The energy dips are therefore computed as

$$m_k = \underset{t \in [t_k, t_{k+1}]}{\arg\min}\ E(t) \quad (4.14)$$

Each m_k represents the absolute minimum of $E(t)$ between t_k and t_{k+1}. The purpose of this calculation is to "trap" each maximum of $E(t)$ between two minima (see an illustation in Fig. 4.4), but it produces only $K-1$ minima points. The missing points are set at the edges as

$$m_0 = \text{MEDIAN } \{E(t) | t < t_1\} \quad (4.15)$$

$$m_K = \text{MEDIAN } \{E(t) | t > t_K\} \quad (4.16)$$

where t_1 and t_K are time instances of the first and last detected syllables, respectively.

The median is used for a noise floor estimate instead of the perhaps more obvious minimum to prevent errors caused by smooth fade ins or outs, which is often present in some recordings.

Two pairs of threshold values are set for each maximum point t_k. At each side, one permissive and one restrictive thresholds are determined as follows:

for all $1 \leq k \leq K$, set:

$$RT_H^k = p_1 \times E(t_k)_{dB} + (1 - p_1) \times E(m_k)_{dB} \quad (4.17)$$

$$RT_L^k = p_2 \times E(t_k)_{dB} + (1 - p_2) \times E(m_k)_{dB} \quad (4.18)$$

$$LT_H^k = p_1 \times E(t_k)_{dB} + (1 - p_1) \times E(m_{k-1})_{dB} \quad (4.19)$$

$$LT_L^k = p_2 \times E(t_k)_{dB} + (1 - p_2) \times E(m_{k-1})_{dB} \quad (4.20)$$

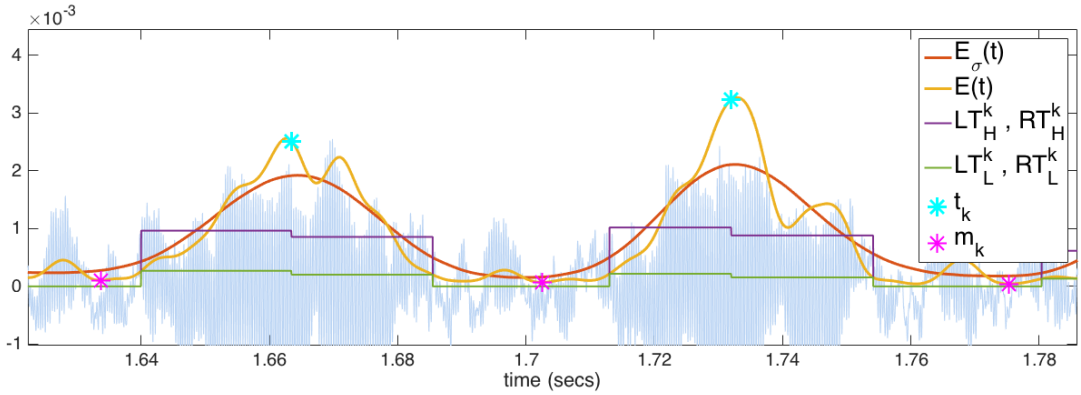


Figure 4.4: Segmentation step of the SACATS algorithm. The yellow and red lines represent the energy functions $E(t)$ and $E_\sigma(t)$, respectively. Notice the difference between them which arises from low pass filtering. m_k (cyan) is the absolute minimum of $E(t)$ in the segment $[t_k, t_{k+1}]$. The thresholds $L(R)T_H^k$ (purple) and $L(R)T_L^k$ (green) use interpolation parameters $p_1 = 0.7$ and $p_2 = 0.3$, respectively. Segmentation is depicted as the support of the purple or green lines in each syllable.

where p_1 and p_2 are set empirically , $0 < p_2 < p_1 < 1$, and LT_H^k, LT_L^k, RT_H^k, RT_L^k are the left and right thresholds, each with high and low values, respectively. An illustration is available in Fig. 4.4.

The low thresholds, LT_L^k and RT_L^k are used for a short time energy function $E_\sigma(t)$, calculated in the same manner as in equation 4.10 but with window size twice as short. The high thresholds, LT_L^k and RT_L^k are used for both $E(t)$ and $E_\sigma(t)$.

Final segmentation is performed by a simple algorithm which compares the energy to the respective threshold. In the first phase, moving from each local maximum point $E(t_k)$ downwards in each direction, the first point for which both $E(t)$ and $E_\sigma(t)$ are lower than the corresponding high threshold (i.e. either RT_H^k or LT_H^k) is considered as a first candidate for a syllable boundary. In a second phase the search continues, and the first points at which $E_\sigma(t) < LT_L^k$ or $E_\sigma(t) < RT_L^k$ from left and right, respectively, are considered as the final boundaries of the syllable. A discrete implementation of the segmentation procedure is implemented in pseudocode in Alg. 2. The output of the algorithm is the interval sequence $I_1, \ldots, I_K$ which are an estimate of syllable bounds in the trill signal.

Algorithm 1 Interval Segmentation

```
for k = 1 to K do
    a_k ← T_{n_k},   b_k ← T_{n_k}
    while E_σ(a_k) > P_30(a_k) do
        a_k ← a_k − 1/f_s
    end while
    while E_σ(b_k) > P_30(b_k) do
        a_b ← a_k + 1/f_s
    end while
    I_k ← [a_k, b_k]
end for
return I_1, I_2, ..., I_K
```

Algorithm 2 Syllable Fine Segmentation

```
for k = 1 to N do
    n ← t_k
    while E_σ[n] > RT_H^k or E[n] > RT_H^k {1st phase} do
        n ← n + 1
    end while
    while E_σ[n] > RT_L^k {2nd phase} do
        n ← n + 1
    end while
    b_k ← n/f_s
    n ← t_k
    while E_σ[n] > LT_H^k or E[n] > LT_H^k {1st phase} do
        n ← n − 1
    end while
    while E[n] > LT_L^k {2nd phase} do
        n ← n − 1
    end while
    a_k ← n/f_s
    I_k ← [a_k , b_k]
end for
return I_1, ..., I_K
```

Chapter 5

Segmentation Algorithm Evaluation

The automatic segmentation algorithms proposed here were evaluated using two different benchmarks. The first benchmark is a group of synthetic signals with additive noise, effectively setting a ground truth for segmentation and parameter estimation. Signals were analyzed according to the harmonic+noise model (HNM) [17] and the harmonic part was synthesized. The ground truth is set de facto by the model parameters. Two data sets were used for synthesis. One is a set of 20 Kingfisher recordings, similar to the ones used for the expert benchmark (see below). The other is a set of 48 trills of various birds, mostly from the Xeno-Canto Database [15].

The second benchmark is an expert's manual segmentation. It comprises a group of over 300 recordings of White-throated kingfishers (*Halcyon Smyrnensis*), recorded in Agmon Hahula over the years 2016-2017.

5.1 Synthetic Ground Truth (SGT)

The following section summarizes the method of preparation of synthetic data.

5.1.1 SGT dataset Selection and Preparation

The SGT dataset comprises two datasets. The first is KFSD, which is produced using 20 recordings of Kingfisher long trills from the Hula lake, with 368 syllables in total.

The second dataset is named VTSD. It is produced using 48 recordings of trills belonging to 21 different species of birds. The recordings were extracted from the Xeno-Canto database [15]. Since recordings often contain many trill calls, they were edited manually so that every recording contains one call only. The tracks at this initial step are referred to as the **organic tracks**.

In many of the tracks, strong transient and stationary noises were still present. Therefore, resulting tracks were edited manually using sound editing software (Audacity [26]) for additional noise removal. Noise with frequency lower than the fundamental frequency of the signal is a major hindrance to correct fundamental frequency estimation. In the case of kingfishers, the fundamental frequency resides within a frequency band of 1.8 - 3.5 kHz, so a simple high pass Butterworth filter was used with cutoff (3 db) frequency of 1.6 kHz. Recordings at this step are referred to as the **clean tracks**. They comprise 1 trill call each. No other manual processing was performed, so signal volume may vary between tracks.

5.1.2 Modelling and Synthesis using HNM

A description of the steps which were carried out for efficient signal modelling using the harmonic+noise model follows.

Fundamental frequency (f_0) estimation was performed at this step. A variant of the YIN algorithm was used, as it was found to outperform other pitch detection algorithms for bird f_0 estimation [22]. Values were saved as the basis for ground truth calculations for parameters involving f_0.

Following [17], The number of harmonics $H(t)$ was determined by spectral peak detection. The Fourier transform absolute value $|\mathcal{F}\{s(t)w(t-t_k)\}(f)|$ is first calculated. The window $w(t)$ used is the Hann window. At each analysis point t_k, lobes are searched for in intervals of the form $R_l = [l \times f_0 - \frac{f_0}{2}, l \times f_0 - \frac{f_0}{2}]$ for an integer l. Lobes are defined here as a frequency interval which contains a local maximum. The lobe bounds are the two local minima closest to the maximum, one on each side. Suppose L is a lobe of maximal area detected in range R_l. Suppose f_l is the frequency at the peak, and lobe area is A_l. For every l, if a lobe was detected in the relevant interval, the following conditions are tested:

$$A_l > 1.5\% \text{ of } \int_{-\infty}^{\infty} |\mathcal{F}\{s(t)w(t-t_k)\}(f)|df \tag{5.1}$$

$$A_l > 2 \times \int_{R_l} |\mathcal{F}\{s(t)w(t-t_k)\}(f)|df \tag{5.2}$$

$$\frac{|f_l - l \times f_0|}{l \times f_0} < 10\% \tag{5.3}$$

Condition (5.1) compares lobe area to the total spectral energy. Condition (5.2) compares lobe area to the spectral energy in the search range only. Condition (5.3) compares detected frequency with the ideal harmonic frequency. If all conditions are met, the l'th harmonic is considered detected, and the test is repeated for the $l+1$ harmonic.

After all harmonics and their frequencies were detected, HNM parameters are estimated and the signal is synthesized. The HNM model assumes the signal can be represented using

the equation [17]

$$HNM(t) = \sum_{j=-H(t)}^{H(t)} A_j(t) e^{2\pi i j f_0(t) t} + n(t) \tag{5.4}$$

$A_j(t)$ are estimated at analysis points t_k by using windowing and least squares approximation. Denoting the harmonic part of (5.4) by $M(t) = \sum_{j=-H(t)}^{H(t)} A_j(t) e^{2\pi i j f_0(t) t}$, the estimation problem takes the form of

$$\arg\min_{A_j} \|w(t - t_k)\,[s(t) - M(t)]\,\|_2^2 \tag{5.5}$$

i.e., the residual to minimize is the noise part. Hann window of 1 ms duration with 50% overlap was used for $w(t)$. In discrete-time processing, (5.5) is formulated using linear algebra. First of all, M(t) can be formulated using matrix multiplication:

$$M(t) = Bx \tag{5.6}$$

with $x = (A_{-H(t_k)}, \ldots, A_{H(t_k)})^T$ signifying the vector of unknowns of length $2H(t_k) + 1$. The matrix B is of size $L + 1 \times 2H(t_k) + 1$, with L+1 equals the length of $w(t)$ in samples. The j'th column of B takes the form

$$E = (e^{2\pi i j f_0(t_k) \frac{0}{f_s}}, e^{2\pi i j f_0(t_k) \frac{1}{f_s}}, \ldots, e^{2\pi i j f_0(t_k) \frac{L}{f_s}})^T \tag{5.7}$$

Multiplication by the window $w(t - t_k)$ can be realized as multiplication by an $L+1 \times L+1$ diagonal matrix W whose diagonal entries equal the L+1 samples of the window function. Using these definitions, the least squares problem (5.5) can be reformulated as

$$\arg\min_{x} \|WBx - Ws\|_2^2 \tag{5.8}$$

with s equals $(s[t_k - \frac{L}{2} + 1], \ldots s[t_k + \frac{L}{2}])^T$, the L+1 length vector whose entries equal the signal samples centered at analysis point t_k.

At this point all model parameters have been estimated, and the harmonic part of (5.4) can be synthesized. Synthesis is implemented with overlap and add (OLA) using the same window used for analysis, Hann window with 50% overlap.

It should be noted that errors in f_0 or harmonic estimation may in turn result in (5.5) having an ill conditioned matrix, which will result in failure of the optimization problem. To fix this, intervals of silence are detected and extrapolation is performed using a method presented in [27]. The gist of this method is to extrapolate the missing data using information from both sides of the missing interval, and then combine both extrapolations using appropriate weighting. In this method, an LPC filter of order 20 is used to extrapolate the signal from both left and right. Afterwards, each extrapolation version is multiplied by

one half of a window function (Hann can be used here as well), to provide smooth decay to zero over the extrapolated interval. The values extrapolated from the left are multiplied by the right half of the window, and the values extrapolated from the right are multiplied by the left half of the window. Thus, weighting the extrapolated data from the left yields smooth decay to zero over the missing interval. The weighted data is exactly zero at the rightmost point of the missing interval. Similarly, the weighted data extrapolated from the right decays smoothly to zero, and reaches zero on the leftmost point of the missing interval. Extrapolation is performed on intervals of maximal length of 5 ms. The two sequences are added up to produced the final extrapolation. Ground truth f_0 over the extrapolated interval is updated to the mean of f_0 estimate of the adjacent analysis points. The resulting tracks are referred to as **clean synthesized tracks**.

5.1.3 Manual Syllable Demarcation

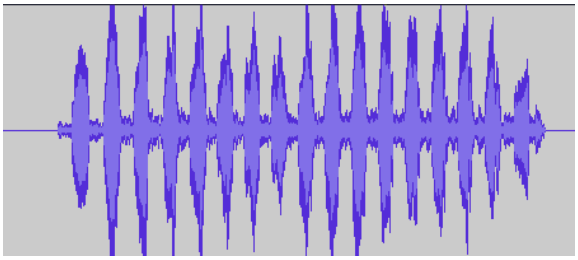
The analysis and synthesis method of section 5.1.2 does not employ any voice activity detection (VAD) mechanism. This is intentional and in fact establishing a syllable VAD algorithm has been the main focus of this thesis. Ambiguity in syllable demarcation was discussed in 1.3 and unique solution wasn't determined. In light of this, an artificial syllable ground-truth indicator is created by manually removing all non-syllable content from the clean synthesized tracks and thus demarcating the syllables' boundaries. The editing was done in sound editing software (Audacity). Resulting tracks are synthesized trills with gaps of silence between syllables. This step produces tracks referred to as **synthetic cut tracks**. Figure (5.1a) presents an original synthesized long trill, and figure (5.1b) presents the same trill after editing.

5.1.4 Smooth Transitions

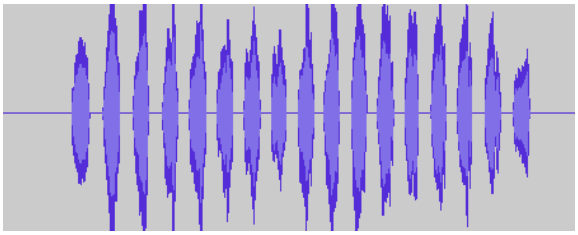
The output of the editing process is usually a discontinuous signal, with unnatural audible clicks. To remove the clicks, and produce natural sounding ground truth synthetic data, a smooth fade in or out transition is required. Therefore, syllables were continued at syllable edges for a short duration for smooth transition effect. This continuation was done both at the beginning and at the end of every syllable.

Only intervals of silence longer than 2 ms were considered as legitimate silence intervals. After silences are detected, syllable presence is deduced by elimination, and ground truth for segmentation is recorded.

New data points were generated using a 10-order linear prediction model, for a total duration of 2% of the original syllable length. Generated data was further multiplied by a half of a Hanning window. First half was used at the beginning of a syllable, where the desired effect was smooth fade in transition. The second half was used for fade out transitions, at the end of each syllable.



(a) Synthesized Long Trill



(b) Edited Synthesized Long Trill

Figure 5.1: Example of a single ground truth track. The original synthetic file is presented above. Below it is the same trill but with the non-syllable part edited out.

This method resembles the method of signal reconstruction through double sided extrapolation used in [27]. However, since the gaps are long enough, only one sided extrapolation is used at each edge of the silence interval. Segmentation ground truth is updated accordingly.

Tracks at this step are referred to as **"fixed synthetic tracks"**. Tracks should sound smooth and resemble their organic counterparts. All ground truth data has been recorded, and finally noise is added to the signals.

5.1.5 Additive Noise

Noise in varying amplitudes was added to the fixed synthesized tracks. In order to simulate natural ambient noise, colored noise was used in addition to white noise. Different tracks with uncorrelated noise were generated with SNR values in the set $\{20, 15, 10, 5, 0, -5, -10, -15\}$ for each noise type, which resulted in 8×2 different tracks for each fixed synthetic track. Evaluation for the two types of noise was performed separately, unless otherwise noted. In the following subsections, considerations involving the implementation of noise addition are detailed.

5.1.5.1 Additive white Gaussian noise

White noise was generated with zero mean and σ^2 variance, according to the formula:

$$SNR_{DB} = 10\log_{10}(\frac{E}{\sigma^2}) \tag{5.9}$$

where SNR_{DB} is the required SNR level in decibels and E is the mean energy of the clean synthesized signal, which was calculated according to the formula:

$$E = \frac{1}{|\mathbb{K}|}\sum_{n\in\mathbb{K}} s^2[n] \tag{5.10}$$

and $\mathbb{K} = \{k \in \mathbb{Z} | s[k] \neq 0\}$ is the discrete support of $s[n]$.

5.1.5.2 Additive Agmon Ha'Hula Gaussian noise

Stationary noise in the Hula lake can come from natural audio sources such as wind, or artificial sources such as airplanes or agricultural machinery. These noise sources produce uncorrelated noise with non-white power spectrum. In order to produce evaluations which are more reflective of natural environments, their power spectrum is estimated and an ambient noise emulating FIR filter is created.

An organic track of pure noise of 0.4 seconds length was used. Its power spectral density was estimated using Welch's method [28]: The track was divided into 50 equally sized segments of 352 samples each. Zero padding is introduced for spectral interpolation, obtaining FFT size of 512 samples. PSD was calculated for each segment, and then averaging was carried out. The result is an estimate of the PSD of the stationary stochastic component in the entire interval:

$$\widehat{PSD} = \frac{1}{50}\sum_{M=0}^{49} |\mathcal{F}\{s_M\}[k]|^2 \tag{5.11}$$

The respective filter is created by using the frequency sampling method of FIR filters. The steps to create the filter are:

1. The desired frequency response is obtained by taking a square root of the PSD: $\tilde{H}(k) = \sqrt{PSD}$
2. An inverse Fourier transform is performed. $\tilde{h} = \mathcal{F}^{-1}\{H\}[k]$
3. The result is truncated by a factor of 2 and multiplied by an interpolating window, chosen here to be a Hamming window of 257 samples length.

$$h[1:257] = \tilde{h}[1:257] * w \tag{5.12}$$

4. An impulse response is obtained by multiplying the result by a gain factor. The gain factor is introduced to normalize the filter's overall power. The output filtered noise should have the same power as a white gaussian noise with variance 1. Therefore, the gain term should be calculated as:

$$g = \sqrt{\frac{N}{\sum |\mathcal{F}\{h[n]\}[k]|^2}} \tag{5.13}$$

Where N is the length of the filter, equals 257 here, which is also the power of a white gaussian random process of the same length.

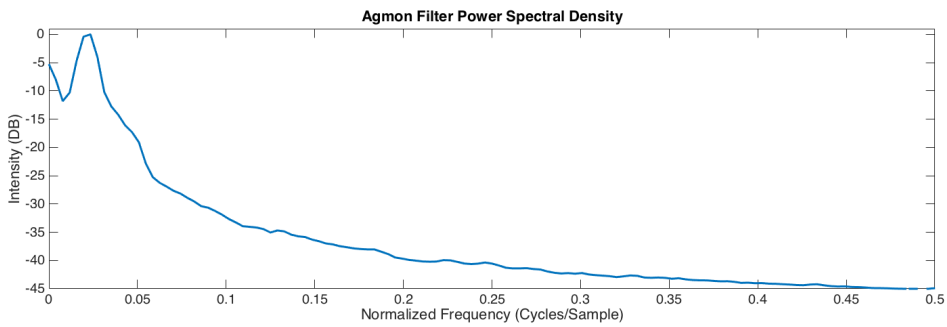


Figure 5.2: Agmon noise filter power spectral density. The peak at 0.02 cycles per sample corresponds to about 880 Hz with $f_s = 44.1$kHz

Once the filter is created, it can be used to produce synthetic noise which exhibits the same PSD as a typical Agmon noisy environment. The colored noise is produced by generating uncorrelated white noise with power according to desired SNR level. The colored noise has the same power as the input noise because of the normalization of (5.13).

5.1.6 Listening Tests

Listening assessments were conducted in order to assess the quality of the synthesis method described in this section. A set of synthesized recordings was played back to four bird experts. Their opinion of the quality of the playback was recorded. It is emphasized that while results were positive, they only show the similarity between organic and synthetic tracks as conceived by human listeners. While this can not be regarded as a sufficient condition to the validity of the method, it can nevertheless be regarded as a necessary one.

The optimal test to asses the naturalness of the synthetic trills would be to present the synthetic trills via playback to birds of the appropriate species, and to use the bird's reaction to evaluate the synthesis quality. This kind of experiment was not feasible when the research was conducted. Instead, it was necessary to settle on humans for quality assessment.

Two listening tests were composed and conducted for the four bird experts who participated in the test. Tests were conducted individually. Audio playback was carried out using a set of home studio monitors (ROKIT RP5 G2 from KRK Systems).

The group of tracks set for the experiment was the long-trill songs of white-throated kingfishers. The stimuli for the tests were prepared using the fixed synthetic tracks (section 5.1.4) and the clean organic tracks (section 5.1.1). Colored Agmon noise was added to each signal to provide authentic ambient noise for the listener. Constant average SNR of 10 db was maintained. For each participant, tracks were randomly shuffled and relabeled. Dictionary was saved for score calculation.

5.1.6.1 Qualitative Test

In this test, 47 audio tracks were played, consisting of 2 warmup tracks, 20 synthetic tracks, 20 organic tracks and 5 unrelated tracks downloaded from Xeno-Canto database [15]. All tracks were randomly shuffled, except for the warmup tracks which always appeared first. Tracks were played sequentially to the experts. After each playback, the participants were asked to rate each track on a 1-5 scale in terms of naturalness, where 1 represents a low quality score, meaning the audio track clearly did not originate from a real bird, and 5 represents a high quality score, meaning the expert was certain a real bird produced the audio signal.

The results are summarized in figure 5.3. The average score of all participants for synthetic trills (3.7 ± 1) is similar to that of natural trills (3.3 ± 1.1). It is evident that the average quality rating given for organic and synthetic tracks is not correlated with the group these tracks belongs to. Experts 1 and 4 rated the organic tracks slightly better, in contrast with experts 2 and 3 who ranked the synthetic tracks better. In summary, this test show that the synthetic trills are indistinguishable from natural trills, at least by bird experts.

5.1.6.2 ABX Test

In this test, the tracks are paired so that each pair contains an organic track and its synthetic counterpart. The order of pairs is shuffled, with a probability of 0.5 that the organic trill would appear before the synthetic one. For each pair, experts are asked to determine which trill is the organic and which is the synthetic. On average, only 40% of the pairs were recognized correctly by the experts. These results and those of the former test provide sufficient evidence that the synthesized signals are indistinguishable from their natural counterparts.

5.1.7 Results

Two synthetic datasets are produced using the steps describes above (sections 5.1.1-5.1.5) for the creation of the "fixed synthetic tracks" with additive noise. The first dataset, la-

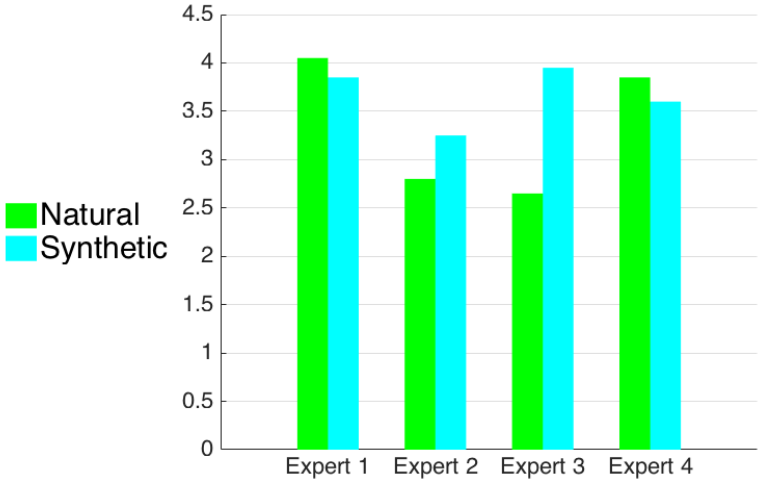


Figure 5.3: Results of acoustic test for evaluating the perceptual differences between natural and synthetic trills by expert listeners. Average rating for organic and synthetic tracks, according to each expert.

beled KFSD (kingfisher synthetic dataset), is produced from the 20 natural white-throated kingfisher recordings collected in the Hula lake. The second dataset, labeled VTSD (various trills synthetic dataset), is produced using the various trills data which contains 48 trills collected mainly from the Xeno-Canto database. For each trill in the synthetic datasets, eight tracks were produced with varying SNR values: 20 , 15 , 10 , 5 , 0 , -5 , -10 and -15 db. Furthermore, there are two types of noise. Therefore, KFSD contains $20 \times 8 \times 2 = 320$ tracks and VTSD contains $48 \times 8 \times 2 = 768$ tracks in total.

Syllable detection results are evaluated using both KFSD and VTSD datasets. Parameter estimation is evaluated using VTSD only.

5.1.7.1 Syllable Detection

Syllable detection rates obtained by each of the automatic algorithms presented in chapter 4 are hereby presented. PD is the probability of detection or true positive rate, and PFA is the probability of false alarm or false positive rate. Both values are plotted as a function of SNR. Eight SNR levels were tested: 20 , 15 , 10 , 5 , 0 , -5 , -10 , -15 db. Both types of noise were tested for the two segmentation algorithms proposed. A syllable is considered detected if there is at least $p\%$ overlap between its benchmark support and estimated support. Detection rates are calculated per syllable, so that PD is the rate of correctly detected syllables, and PFA is the rate of wrongly detected syllables, i.e of detection of a syllable while there is no one present.

For the first set, which contains the 20 White-throated Kingfisher trills, $p = 80$ was used. Results for this dataset are presented in Fig. 5.4 and Fig. 5.5. PD for both algorithms and noise types are higher than 95% for high SNR values (i.e. ≥ 5 db.) For synthetic trills with white noise, even for 0dB SNR, the PD is still higher than 90For tracks with white noise, the PFA is less than 5% for SACATS and less than 15% for FVDNEM for SNR > 0. In fact, for SNR > 5, PFA falls below 1% for SACATS. SACATS manifests better PFA because its detection criterion (correlation) is evidently more restrictive than the f_{vd} criterion of the FVDNEM. Stability of the f_0 estimator is a good indicator of syllable presence. However, its shortcoming is that stable f_0 estimation can also occur in parts of the signal where syllables are not present. The correlation criterion of SACATS, on the other hand, uses the assumption of the rhythmical nature of syllable emission. Therefore, it will ignore signal parts which does not exhibit the assumed rhythmical relation to the other syllables in the trill. For this reason, FVDNEM tends to detect calls of other birds as syllables, while SACATS does not. The choice of algorithm ultimately depends on the user's preferences.

Fig. 5.6 and Fig. 5.7 show the graphs of the second dataset, which contains the 48 trills of various bird species. This set uses $p = 20$. Results for this dataset are interpreted similarly to the first dataset. Detection rates are higher than 90% for high SNR values (> 0) for all noise levels and algorithms. False alarm rates are under 7% for FVDNEM and under 2% for SACATS and the same explanation as the above holds here as well. The only anomaly is a decline in detection rates of SACATS (82%) in Agmon noise, which is slightly lower than the performance of FVDNEM or even SACATS in white noise.

5.1.7.2 Parameter Estimation

Four parameters were evaluated and compared to the synthetic benchmark. These parameters are: **over-detection** and **under-detection** of syllable boundaries, **mean fundamental frequency** (f_0) and the **bandwidth** of f_0. The parameters were evaluated for the various trills dataset. They were calculated only for detected syllables. Undetected syllables are not taken into account in parameter error computation. For this reason, mean errors presented in this section are represent the data better for noise levels with high PD (i.e. high SNR) than for those with low PD.

Over detection quantifies the amount by which the estimated syllable support exceeds the ground truth support. Formally, suppose I_k is the ground truth support of the k'th syllable in a trill, and $\hat{I}_k$ is the estimated support. The over-detection parameter for the syllable is computed as $m(\hat{I}_k - I_k)$ where $m(A)$ is the Borel measure of set A, and the difference is understood as set difference. The relative over-detection parameter is defined as $\frac{m(\hat{I}_k - I_k)}{m(I_k)}$. The relative parameter expresses the error relative to the benchmark syllable duration.

Similarly, **under detection** is defined as the amount by which the ground truth support exceed the estimated support. Mathematically, absolute under-detection is defined as

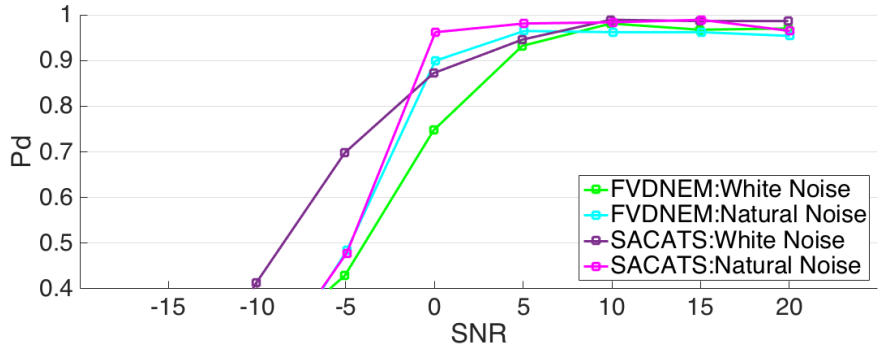


Figure 5.4: Detection rate (PD) vs. SNR of the two proposed algorithms (FVDNEM and SACATS) for the KFSD dataset.

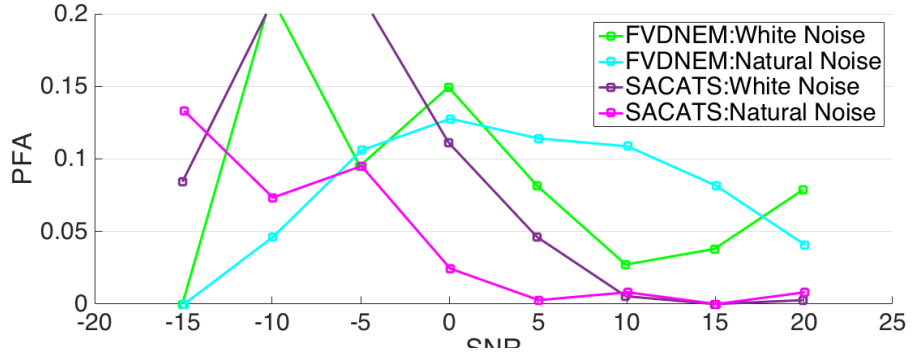


Figure 5.5: PFA vs. SNR for the KFSD dataset

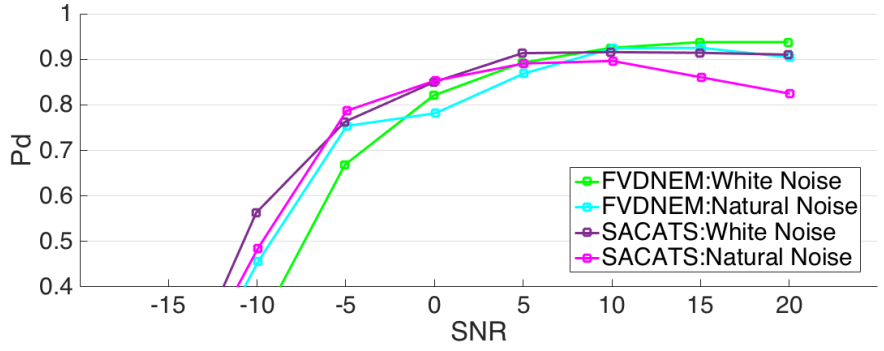


Figure 5.6: PD vs. SNR for the VTSD dataset

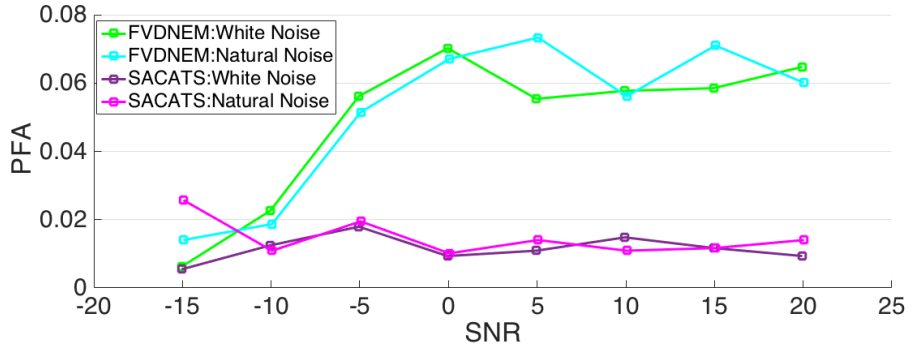


Figure 5.7: PFA vs. SNR for the VTSD dataset

$m(I_k - \hat{I}_k)$ and relative under-detection is defined as $\frac{m(I_k - \hat{I}_k)}{m(I_k)}$.

Over-detection mean relative errors per noise level for both white and Agmon noise are summarized in table 5.1. It is evident that both algorithms perform similarly for both noise types. For positive SNR values, a positive bias of $10\% - 30\%$ is found for white noise, and $20\% - 40\%$ for Agmon noise. The bias can be controlled by the parameters of the segmentation part in both algorithms. It should be noted that results for SNR values with low detection rates (below -5db) should not be considered as measuring parameter estimation reliably (for reasons discussed at the beginning of the section.) In contrast with the remaining parameters, whose estimation errors tend to increase with noise, over-detection estimation errors behaves conversely (gradually decrease from 20db down to -5db where they start to increase again.) The peculiar behaviour of the over-detection parameter can be justified by noticing that syllables which are prone to over-detection are typically the weaker syllables in the signal, because they are hard to distinguish from the noisefloor. In strong noise, they can get covered up completely. In these situations they are not detected by the algorithms and are therefore left out of the parameter estimation error computation, thereby reducing the errors as noise increases.

The equivalent information for the **under-detection** parameter is summarized in table 5.2. No error bias is evident, as both algorithms scored mean errors lower than 10% for positive SNR values.

The definitions for **bandwidth** and **mean** f_0 are similarly defined per syllable. Suppose the minimal and maximal estimated f_0 in the k'th syllable are $\hat{f}_k^m$ and $\hat{f}_k^M$, respectively. The estimated bandwidth is $\hat{B} = \hat{f}_k^M - \hat{f}_k^m$ and the mean fundamental frequency is $\hat{\bar{f}}_k = \frac{1}{2}\left(\hat{f}_k^M + \hat{f}_k^m\right)$.

Mean error values for bandwidth and syllable mean f_0 are presented in tables 5.3 and 5.4, respectively. As in the case of over and under detection measures, the behaviour of both algorithms is quite similar. However, while for mean f_0 there is no notable difference in estimation errors for signals generated with both noise types, bandwidth estimation manifested high sensitivity to the different noise types. In fact, the performance of estimation in white noise was two times better than that of Agmon noise estimation for 20 db SNR. The gap gradually diminishes going down to 0 db SNR. This behaviour however is in correlation with the errors of the over-detection parameter, and can thus be explained in the following way: Over detection of syllable boundaries causes wrongful inclusion of signal parts with no bird vocalizations in syllables. In many cases, these parts do not contain any harmonic component at all. f_0 estimation of the YIN algorithm will provide inaccurate results which in turn increase the error of bandwidth estimation. By contrast, the mean f_0 estimator does not suffer from over-detection as much as the bandwidth estimator, which can be taken as an empirical proof of its noise resilience compared to the bandwidth estimator. We can also deduce that the excess error caused by the difference in noise type can be either positive or negative, and in fact its expectation is close to zero since the mean f_0

values are almost identical in both cases.

White Noise	**20 db**	**15 db**	**10 db**	**5 db**	**0 db**	**-5 db**	**-10 db**	**-15 db**
FVDNEM	31%	23.5%	18.5%	11.7%	8%	8.6%	24%	37.3%
SACATS	35%	27.7%	20.3%	12.6%	9.7%	10.1%	17.9%	18%
Agmon Noise	**20 db**	**15 db**	**10 db**	**5 db**	**0 db**	**-5 db**	**-10 db**	**-15 db**
FVDNEM	49.5%	42.3%	30.6%	27.8%	22.2%	17.1%	21.5%	48.9%
SACATS	54.3%	43.4%	40%	28.5%	24.3%	18.3%	17.9%	74.1%

Table 5.1: Over detection of syllable boundaries: mean relative error

White Noise	**20 db**	**15 db**	**10 db**	**5 db**	**0 db**	**-5 db**	**-10 db**	**-15 db**
FVDNEM	4%	4.5%	6.1%	8.9%	14.2%	19.1%	20.5%	32%
SACATS	1.4%	1.9%	3.9%	6.1%	10.4%	15.8%	17.5%	28.7%
Agmon Noise	**20 db**	**15 db**	**10 db**	**5 db**	**0 db**	**-5 db**	**-10 db**	**-15 db**
FVDNEM	2.8%	2.9%	3.2%	4.2%	4.6%	6.9%	9.6%	13.5%
SACATS	0.7%	0.9%	1.4%	2.3%	3.8%	4.8%	7%	42.5%

Table 5.2: Under detection of syllable boundaries: mean relative error

White Noise	**20 db**	**15 db**	**10 db**	**5 db**	**0 db**	**-5 db**	**-10 db**	**-15 db**
FVDNEM	34.8%	36.6%	40.7%	45.7%	57.5%	79.9%	96.2%	21.6%
SACATS	34.3%	36.2%	41.4%	45.1%	56.3%	77.8%	95.6%	97.2%
Agmon Noise	**20 db**	**15 db**	**10 db**	**5 db**	**0 db**	**-5 db**	**-10 db**	**-15 db**
FVDNEM	56.4%	52.6%	46.8%	46%	47.6%	48.7%	55.1%	47.5%
SACATS	61.2%	51.2%	49.1%	50.1%	50.7%	48.3%	60.5%	77.8%

Table 5.3: Bandwidth: mean relative error

White Noise	20 db	15 db	10 db	5 db	0 db	-5 db	-10 db	-15 db
FVDNEM	18.8%	18.6%	18.6%	19.3%	20.4%	25.3%	45.9%	50.6%
SACATS	19.7%	19.3%	19.3%	19.8%	20.3%	24.4%	40.3%	35.4%
Agmon Noise	**20 db**	**15 db**	**10 db**	**5 db**	**0 db**	**-5 db**	**-10 db**	**-15 db**
FVDNEM	17.7%	17.6%	17.2%	18.8%	22%	23.2%	36.9%	9.3%
SACATS	20.3%	19.1%	19%	19.6%	25.1%	23.4%	50%	68%

Table 5.4: Mean f_0: mean relative error

5.2 Expert Benchmark

5.2.1 Dataset DK1

In her MSc thesis [19], Dana Klein analyzed white-throated kingfisher trills which had been recorded over two consecutive nesting periods (2016-2017) and found statistical correlation between syllable rate and reproduction potency. A group of 312 manually segmented trills from the 2017 nesting period was used for evaluation of segmentation and parameter estimation. The trills are unedited and taken from the corpus of data collected by Klein as part of her research.

The performance of the automatic algorithms was evaluated compared to the human expert annotations and parameters estimation. Results are presented in table 5.5. We note that SACATS attains higher detection rates and lower false alarm rates, and is thus in higher correlation with the expert's performance.

	PD	PFA
FVDNEM	88.2%	26.8%
SACATS	91.8%	7.2%

Table 5.5: Detection rates: comparison to 312 trills segmented by human expert (DK1)

	Over-detection	Under-detection	Bandwidth	Mean f_0
FVDNEM	24.6%	2.5%	1.3%	-42.5%
SACATS	24.75%	2.2%	1.5%	-40%

Table 5.6: Parameter estimation: average relative errors on DK1

Table 5.6 presents mean relative errors for the 4 parameters defined in Sec. 5.1.7.2. Similar to previous results, a larger bias in estimation of the over-detection parameter

is present, in comparison with the under-detection parameter. However, the bias is not passed on to the bandwidth parameter, which obtained lower estimation errors ($1.3-1.5\%$, compared to 24.7% obtained by the over-detection parameter.) Interestingly, the mean f_0 parameter obtained mean errors much higher than those of the SGT dataset. However, they could originate in a qualitative preference imposed by Klein in her manual work which had not yet been formulated during data collection. No visible differences exist between the performance of the two algorithms.

5.2.2 Dataset DK2

In addition, Klein composed 4 groups of 5 trills each. The trills are grouped according to the track noise level, as subjectively experienced by Klein. The groups are labeled by the integers: 1-4, in decreasing degrees of noise. Group number 1 is composed of the noisiest trills, and group number 4 is composed of the cleanest trills. We stress that classification was not performed by objective standards. It reflects Klein's experience with the data, and may be biased towards her subjective preferences and inclinations.

Fig. 5.8 and Fig. 5.9 show plots of PD and PFA vs noise level, respectively. Pd is the probability of detection per syllable, and is defined as the number of true positives divided by the total number of syllables. A *true positive* is a detected syllable, and is considered as detection if its estimated support shares at least 80% with the real syllable support (as segmented by Klein). PFA is the probability of false alarm and is calculated as the number of false positives, i.e. estimated syllables which does not conform to the detection criterion , divided by the total number of non-syllable detections, i.e. false positives + true negatives. As can be seen in the plots, detection rates are high for both algorithms ($\geq 94\%$) for noise levels 3 and 4, while PFA is around 30% for algorithm FVDNEM and less than 5% for SACATS.

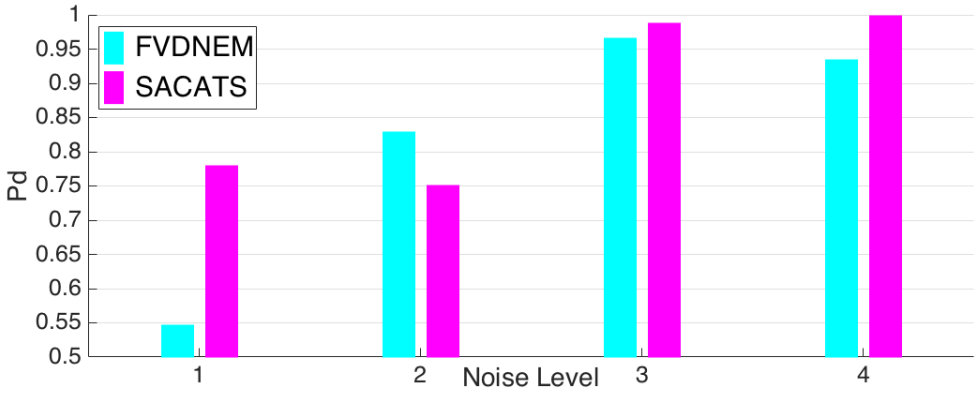


Figure 5.8: PD (probability of detection) of segmentation algorithms, with expert segmentation used as benchmark. Trills are grouped in ascending order according to noise and recording quality, with 1 being the noisiest and 4 being the most coherent. As can be seen, outstanding results are obtained for groups 3-4. See the text for more details.

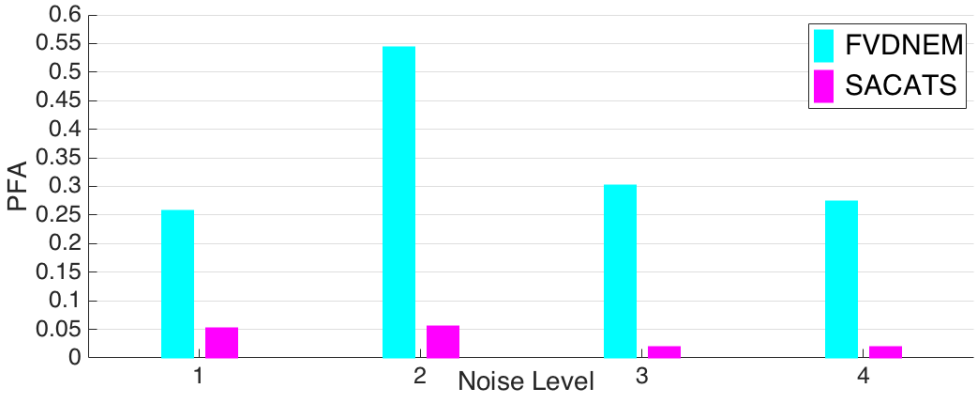


Figure 5.9: PFA (probability of false alarm) vs. Noise Level. FVDNEM obtains around 30% PFA, while SACATS obtains a very low PFA of under 5%.

Chapter 6

Summary and Conclusions

This thesis presents a study that attempts to design an automated system for trill syllable segmentation and parameter estimation. The aim of the conducted research was to determine whether an automatic system can reliably replace manual segmentation performed by human experts, which is both tedious and time consuming. The results obtained show strong correlation between the performance of the two segmentation algorithms proposed here with both synthetic trill benchmark and human expert annotation. SACATS performed slightly better in both human and synthetic benchmarks for detection rate and false positive rate. In contrast, FVDNEM presented slightly smaller average error rates for parameter evaluation. Error rates were found to be sufficiently small for both algorithms to conclude that the algorithms can replace human manual work.

The system was developed and implemented using Matlab. It was packaged into a software called the "TrillOmatic", which will hopefully be used by ecologists specializing in conservation and research of bird populations. An illustration of the software interface is presented in Fig. 6.1. The TrillOmatic's interface allows the user to perform automatic segmentation using both of the algorithms developed in this research. The user can optionally manually adjust the algorithm output, or perform automatic parameter estimation. Spreadsheet export is also available.

Further research would include automatic detection and classification of trills and other birdsong signals in long and noisy recordings, using the estimated parameters as feature vectors for machine learning algorithms. It will be interesting to compare the performance of machine learning algorithms between this feature space and a more traditional feature space which uses time-frequency features. Aggregating features from both spaces could potentially provide better results for detection and classification. Also, further generalization of the segmentation and parameter estimation algorithms to trills of animals other than birds was assumed but never tested. Creating a dataset similar to those composed for the present research and testing the algorithms on the new data can confirm and substantiate the results obtained thus far.

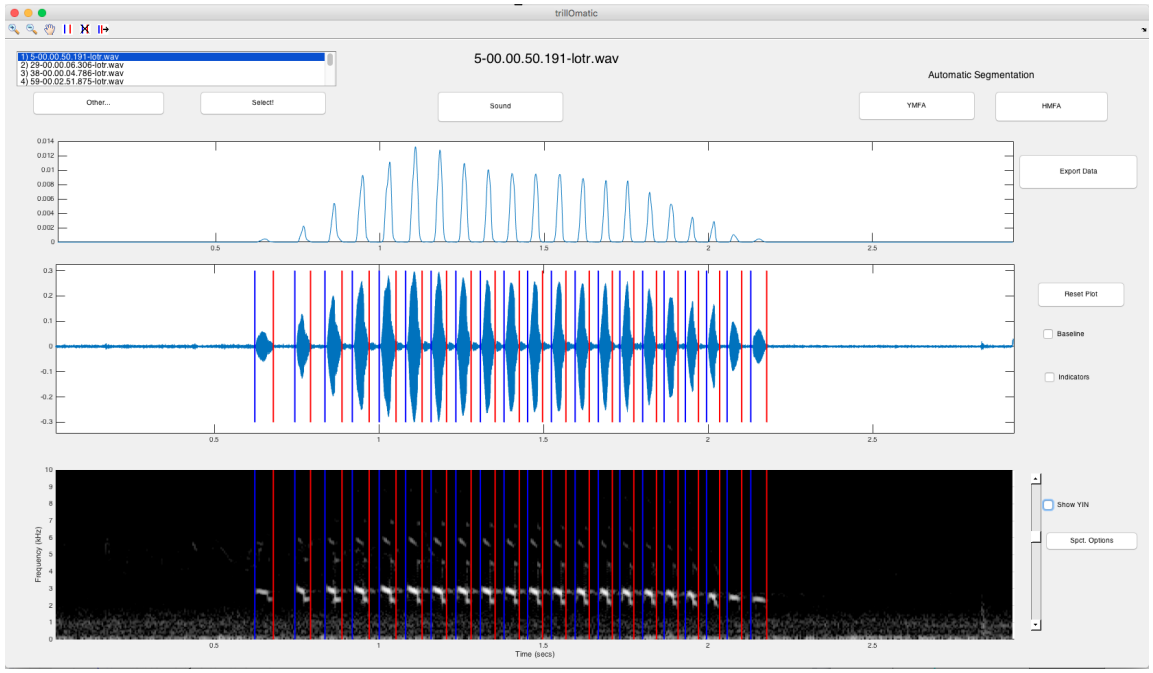


Figure 6.1: TrillOmatic: automatic trill syllable segmentation research utility

Chapter 7

Bibliography

[1] K. J. Odom, M. L. Hall, K. Riebel, K. E. Omland, and N. E. Langmore, "Female song is widespread and ancestral in songbirds," *Nature Communications*, vol. 5, p. 3379, 2014.

[2] M. E. Tom, S. E. Campbell, M. D. Beecher, *et al.*, "Song type matching is an honest early threat signal in a hierarchical animal communication system," *Proceedings of the Royal Society of London B: Biological Sciences*, vol. 280, no. 1756, p. 20122517, 2013.

[3] J. W. Bradbury and S. L. Vehrencamp, *Principles of animal communication.* Sinauer Associates, Incorporated, 1998.

[4] K. E. Boul, W. C. Funk, C. R. Darst, D. C. Cannatella, and M. J. Ryan, "Sexual selection drives speciation in an amazonian frog," *Proceedings of the Royal Society of London B: Biological Sciences*, vol. 274, no. 1608, pp. 399–406, 2007.

[5] C. ten Cate and K. Okanoya, "Revisiting the syntactic abilities of non-human animals: natural vocalizations and artificial grammar learning," *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, vol. 367, no. 1598, pp. 1984–1994, 2012.

[6] Y. Prat, M. Taub, and Y. Yovel, "Vocal learning in a social mammal: Demonstrated by isolation and playback experiments in bats," *Science Advances*, vol. 1, no. 2, p. e1500019, 2015.

[7] 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 Recognition Letters*, vol. 31, no. 12, pp. 1524–1534, 2010.

[8] J. Cai, D. Ee, B. Pham, and P. Roe, "Sensor network for the monitoring of ecosystem: Bird species recognition," in *3rd Int. Conf. on Intelligent Sensors, Sensor Networks and Information*, IEEE, 2007.

[9] D. Stowell, M. Wood, Y. Stylianou, and H. Glotin, "Bird detection in audio: a survey and a challenge," *arXiv preprint arXiv:1608.03417*, 2016.

[10] D. Stowell, Y. Stylianou, M. Wood, H. Pamuła, and H. Glotin, "Automatic acoustic detection of birds through deep learning: the first bird audio detection challenge," *Methods in Ecology and Evolution*, 2018.

[11] H. Goëau, S. Kahl, H. Glotin, B. Planqué, W. Vellinga, and A. Joly, "Overview of birdclef 2018: Monospecies vs. sundscape bird identification," in *CLEF (Working Notes)*, vol. 2125 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2018.

[12] I. Charrier, N. Mathevon, and T. Aubin, "Bearded seal males perceive geographic variation in their trills," *Behavioral Ecology and Sociobiology*, vol. 67, no. 10, pp. 1679–1689, 2013.

[13] J. Podos, "Correlated evolution of morphology and vocal signal structure in darwin's finches," *Nature*, vol. 409, no. 6817, p. 185, 2001.

[14] B. Ballentine, "The ability to perform physically challenging songs predicts age and size in male swamp sparrows, melospiza georgiana," *Animal Behaviour*, vol. 77, no. 4, pp. 973–978, 2009.

[15] Xeno-Canto Foundation, "Xeno-canto." `https://www.xeno-canto.org/`. Accessed: 2018-10-30.

[16] J. Podos, "A performance constraint on the evolution of trilled vocalizations in a songbird family (passeriformes: Emberizidae)," *Evolution*, vol. 51, no. 2, pp. 537–551, 1997.

[17] Y. Stylianou, "Applying the harmonic plus noise model in concatenative speech synthesis," *IEEE Transactions on speech and audio processing*, vol. 9, no. 1, pp. 21–29, 2001.

[18] C. Brown and T. Riede, *Comparative Bioacoustics: An Overview*. Bentham Science Publishers, 2017.

[19] D. Klein, "Song and plumage coloration in white-throated kingfishers (halcyon smyrnensis): correlation between signal expression and reproductive success," Master's thesis, Tel-Hai College, Israel, 2018.

[20] T. M. Aide, C. Corrada-Bravo, M. Campos-Cerqueira, C. Milan, G. Vega, and R. Alvarez, "Real-time bioacoustics monitoring and automated species identification," *PeerJ*, vol. 1, p. e103, 2013.

[21] A. De Cheveigné and H. Kawahara, "Yin, a fundamental frequency estimator for speech and music," *The Journal of the Acoustical Society of America*, vol. 111, no. 4, pp. 1917–1930, 2002.

[22] C. O'Reilly, N. M. Marples, D. J. Kelly, and N. Harte, "Yin-bird: Improved pitch tracking for bird vocalisations.," in *INTERSPEECH*, pp. 2641–2645, 2016.

[23] M. Lasseck, "Large-scale identification of birds in audio recordings.," in *CLEF (Working Notes)*, pp. 643–653, 2014.

[24] Y. Fukuzawa, S. Marsland, M. Pawley, and A. Gilman, "Segmentation of harmonic syllables in noisy recordings of bird vocalisations," in *Int. Conf. on Image and Vision Computing New Zealand (IVCNZ), 2016*, pp. 1–6, IEEE, 2016.

[25] P. Maragos, J. F. Kaiser, and T. F. Quatieri, "Energy separation in signal modulations with application to speech analysis," *IEEE transactions on signal processing*, vol. 41, no. 10, pp. 3024–3051, 1993.

[26] Audacity® software is copyright © 1999-2019 Audacity Team. Web site: `https://audacityteam.org/`. It is free software distributed under the terms of the GNU General Public License. The name Audacity® is a registered trademark of Dominic Mazzoni.

[27] I. Kauppinen and J. Kauppinen, "Reconstruction method for missing or damaged long portions in audio signal," *Journal of the Audio Engineering Society.*, vol. 50, 01 2012.

[28] P. D. Welch, "The use of fast fourier transform for the estimation of power spectra: A method based on time averaging over short, modified periodograms," *Audio and Electroacoustics, IEEE Transactions on*, vol. 15, pp. 70 – 73, 07 1967.