HELSINKI UNIVERSITY OF TECHNOLOGY
Department of Electrical and Communications Engineering
Laboratory of Acoustics and Audio Signal Processing

Seppo Fagerlund

Automatic Recognition of Bird Species by Their Sounds

Master's Thesis submitted in partial fulfillment of the requirements for the degree of Master of Science in Technology.

Espoo, Nov 8, 2004

Supervisor: Professor Unto K. Laine
Instructor: Dr. Tech. Aki Härmä

HELSINKI UNIVERSITY OF TECHNOLOGY

ABSTRACT OF THE MASTER'S THESIS

Author:	Seppo Fagerlund
Name of the thesis:	Automatic Recognition of Bird Species by Their Sounds
Date:	Nov 8, 2004 **Number of pages:** 56
Department:	Electrical and Communications Engineering
Professorship:	S-89
Supervisor:	Prof. Unto K. Laine
Instructor:	Dr. Tech. Aki Härmä

Bird sounds are divided by their function into songs and calls which are further divided into hierarchical levels of phrases, syllables and elements. It is shown that syllable is suitable unit for recognition of bird species. Diversity within different types of syllables birds are able to produce is large. In this thesis main focus is sounds that are defined inharmonic.

Automatic recognition system for bird species used in this thesis consist of segmentation of syllables, feature generation, classifier design and classifier evaluation phases. Recognition experinments are based on parametric representation of syllables using a total of 19 low level acoustical signal parameters. Simulation experinments were executed with six species that regularly produce inharmonic sounds. Results shows that features related to the frequency band and content of the sound provide good discrimination ability within these sounds.

Keywords: bird sounds, species recognition, audio classification, pattern recognition, feature extraction

TEKNILLINEN KORKEAKOULU DIPLOMITYÖN TIIVISTELMÄ

Tekijä:	Seppo Fagerlund	
Työn nimi:	Lintulajien automaattinen tunnistaminen äänien avulla	
Päivämäärä:	8.11.2004	**Sivuja:** 56
Osasto:	Sähkö- ja tietoliikennetekniikka	
Professuuri:	S-89	
Työn valvoja:	Prof. Unto K. Laine	
Työn ohjaaja:	TkT Aki Härmä	

Lintujen äänet jaetaan niiden tehtävän perusteella lauluihin ja kutsuääniin, jotka edelleen jaetaan hierarkisen tason perusteella virkkeisiin, tavuihin ja elementteihin. Näistä tavu on sopiva yksikkö lajitunnistukseen. Erityyppisten äänten kirjo linnuilla on laaja. Tässä työssä keskitytään ääniin, jotka määritellään epäharmonisiksi.

Tässä työssä käytettävä lintulajien automaattinen tunnistusjärjestelmä sisältää seuraavat vaiheet: tavujen segmentointi, piirteiden irrotus sekä luokittelijan opetus ja -arviointi. Kaikki lajitunnistuskokeilut perustuvat tavujen parametriseen esitykseen käyttäen 19:ta matalan tason äänisignaalin parametria. Tunnistuskokeet toteutettiin kuudella lajilla, jotka tuottavat usein epäharmonisia ääniä. Tulosten perusteella piirteet, jotka liittyvät äänten taajuuskaistaan ja -sisältöön luokittelevat hyvin nämä äänet.

Avainsanat: lintujen äänet, lajitunnistus, äänimateriaalien luokittelu, hahmontunnistus, piirreirrotus

Acknowledgements

This Master's thesis has been done in the Laboratory of Acoustics and Audio Signal Processing at the Helsinki University of Technology. The reseach has been done within the Avesound project and it is funded by the Academy on Finland.

I want to thank my thesis supervisor Professor Unto K. Laine and instructor Dr. Tech. Aki Härmä for their valuable ideas and comments throughout this work. I would like to thank also the Avesound consortium for the momentous meetings during this project. On behalf of the project I want also to thank Mr. Ilkka Heiskanen and other nature recording experts, whose large high quality bird sound recording archives have made this project possible.

Furthermore, I would like to thank my family and friends, whose presence and support have helped me during my studies and especially during my thesis. Finally, I would like to thank my fiancée Anu Tanninen for her love and support in all possible ways.

Otaniemi, November 8, 2004

Seppo Fagerlund

Contents

Abbreviations

ABS/OLA	Analysis-By-Synthesis/Overlap-Add
BW	Signal Bandwidth
DSM	Delta Spectrum Magnitude (Spectral Flux)
FFT	Fast Fourier Transform
kNN	k-Nearest Neighbour classifier
LDA	Linear Discriminant Analysis
LL	Lateral Labia
ML	Medial Labia
MSE	Mean square error
MTM	Medial Tympaniform Membrane
NN	Nearest Neighbour classifier
SC	Spectral Centroid
SF	Spectral Flatness
SRF	Spectral Rolloff Frequency
STFT	Short Time Fourier Transform
XML	eXtensible Markup Language

Symbols

Models of bird sounds production

c	dissipation constant
f	frequency in Hz
g	size of the beak tip gape
h	length of the MTM
k	wave number
l	length of the tube
l_B	length of the beak
m	efective mass of the membrane
p_0	Air pressure on bronchial side of the syrinx
p_1	Air pressure on tracheal side of the syrinx
r	radius of the bronchus
u	displacement of the membrane/labia
u_a	displacement of the upper edge of the labia
u_b	displacement of the lower edge of the labia
u_0	position of the membrane/labia at rest
v	speed of the sounds
A	cross sectional area of the tube
A_B	cross sectional area of the beak
C	constant term
F	force against membrane/labia
K	input impedance of the beak
L	input impedance of the larynx
M	input impedance of the mouth
T	input impedance of the trachea
U	Air flow through syrinx
Z	input impedance

α	attenuation coefficient
δ	correction term of the beak
ϵ	small constant term
κ	damping coefficient
ρ	Air dencity
τ	phenomenological constant

System for bird species recognition

a_m	Amplitude of the sinusoid
d_E	Euclidean distance measure
d_M	Mahalanobis distance measure
f_m	Frequency of the sinusoid
$s(n)$	Synthesized signal vector
x, y	Feature vector
$\widehat{x}$	Normalized feature vector
$\overline{x}_k$	Mean of the feature k
$x(n)$	Original signal vector
$x_i[n]$	Frame of the signal $x(n)$
$w(n)$	Window sequence
A_m	Arithmetic mean of $X(n)$
$E(m)$	Energy envelope of the signal $x(n)$
$E_{\mathcal{H}}(n)$	Envelope of the signal $x(n)$
F_s	Sampling frequency
G_C	Modelling gain
G_m	Geometric mean of $X(n)$
H_C	Class likelihood
J	Discriminative power
P	*A-priori* probability
S_b	Between-class scatter matrix
S_w	Within-class scatter matrix
$X(n)$	Fourier transformed signal $x(n)$
μ	Mean vector of x
σ_k^2	Variance of the feature k
ϕ_m	Phase of the sinusoid
ω_m	Angular frequency of the sinusoid
Σ	Covariance matrix
$\mathcal{H}$	Hilbert transform

Chapter 1

Introduction

Birds and especially sounds of birds are important for humans and to our culture. For many people sound of birds is the sign for starting of the spring. Bird-watching is also popular hobby in many countries. Birds can be heard even in big cities and there they are one of the few reminders of the surrounding nature. Most of the people are able to recognize at least few most common species by their sound and experts can recognize hundreds of species only by their sound. Goals in this work is to develop methodology for the system that could automatically recognize bird species or even individual birds by their sounds.

Birds produce their sounds mainly by syrinx, witch is unique organ for birds (King & McLelland 1989). The organ is complex in structure and function but also diversity of the organ within the bird species is large. Therefore spectrum of the different sounds birds are able to produce is also large, which sets great challenges to the development of an automatic sound classification system. Bird sounds can be divided into songs and calls, which can be further divided into hierarchical levels of phrases, syllables and elements or notes (Catchpole & Slater 1995). Syllables are constructed of one or more elements, but it can be seen as suitable unit for recognition of bird species because they can be more accurately detected from continuous recordings than elements. Phrases and songs include more regional and individual variability than syllables .

Recognition of bird species by their sounds is a typical pattern recognition problem. Patterns are sound events produced by birds which are represented with few acoustical parameters (features) of the sound. Recognition is done based on this parametric representation of the sound events by comparing those with the models of sounds produced by species in recognition experinment. Features should be selected so that they are able to maximally distinguish sounds that are produced by different bird species (classes).

Relatively little have been done previously on automatic recognition on bird species. However this problem is related to the other audio data classification problems like classifi-

cation of general audio content (Li, Sethi, Dimitrova & McGee 2001, Wold, Blum, Keislar & Wheaton 1996), auditory scene recognition (Eronen, Tuomi, Klapuri, Fagerlund, Sorsa, Lorho & Huopaniemi 2003), music genre classification (McKinney & Breebaart 2003) and also to the speech recognition, that have been studied relatively extensively during last few years. In these problems classification is done based on parametric representation of congruent segments of the raw recordings.

1.1 Previous work

Only few studies have been done on automatic recognition of bird species and efficient parametrization of bird sounds. In (Anderson, Dave & Margoliash 1996, Kogan & Margoliash 1998) dynamic time warping and hidden Markov models were used for automatic recognition of songs of Zebra Finches (*Taeniopygia guttata*) and Indigo Puntings (*Passerina cyanea*). In these studies syllables were represented by spectrograms. Comparison of spectrograms is computationally demanding and, in the case of field recordings, they often include also environmental information that is not relevant for recognition of bird species.

In (McIlraith & Card 1997) were tested recognition of songs of six species common in Manitoba, Canada. In this work songs were represented with spectral and temporal parameters of the song. Dimensionality of the feature space were reduced by selecting features for classification by means of their discriminative ability. Neural networks were used for classification of the songs. Training of the neural networks is computationally demanding, but classification with the network is relatively fast. Also backpropagation algorithm weights automatically different features.

Nelson (Nelson 1989) studied discriminative ability of different features in Field Sparrow (*Spizella pusilla*) and Chipping Sparrow (*Spizella passerina*) against 11 other bird species. He noted that features have different classification ability in context of different species. Nelson uses canonical discriminant analysis to determine and select features that maximize the recognition result.

This Thesis has been done within the Avesound project (Fagerlund 2004) at Helsinki University of the Technology. Previous work in this project have been related to the sinusoidal modelling of syllables. In (Härmä 2003) syllables were parametrized with simple one sinusoidal model. Modeled sinusoid were represented by time-varying frequency and amplitude trajectories. Albeit this model may be oversimplified in many cases, it gave encouraging recognition results within species that produce regularly tonal and harmonic sounds.

In (Härmä & Somervuo 2004) bird sounds were labeled into four classes based on their harmonic structure. In the classes each harmonic component was modeled with one time-varying sinusoid. The first class was for pure tonal sounds and the syllables in this class

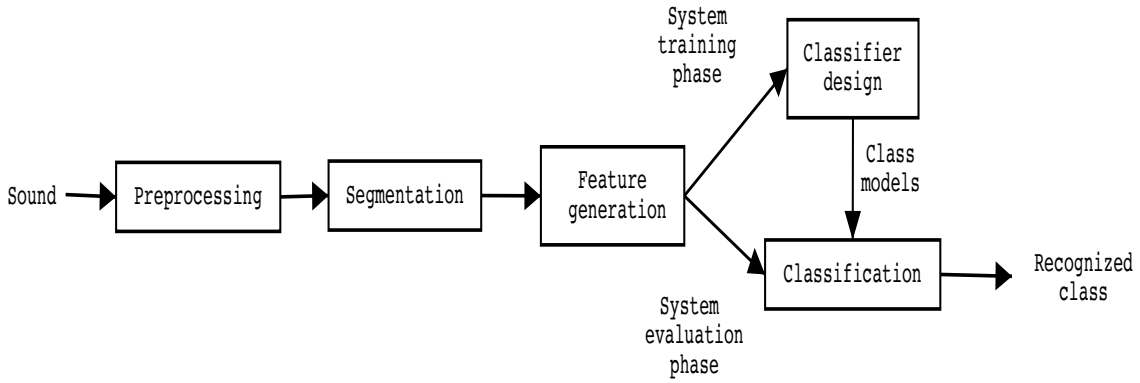

Figure 1.1: Elements and structure of general audio classification system.

were modeled with the model described in (Härmä 2003). Syllables in the class II are pure harmonic in structure, whose strongest component is fundamental frequency. Classes III and IV are for sounds whose strongest component is first and second harmonic of the fundamental frequency.

1.2 Sound classification system

Automatic identification of audio patterns have become a popular research topic in the field of audio and signal processing in recent years. General structure of such system is quite similar in different audio classification problems. Stages of a basic audio classification system is presented in figure 1.1. Building audio classification system involves two phases; system training and system evaluation phase. In the first phase models of classes are trained with training data set, and in system evaluation phase the system performance is evaluated with test data set, which is usually different from the training data set. Incoming sound needs often some preprocessing before it can be classified. Preprocessing phase may include for example noise reduction and transformation of the data into a desired format.

In segmentation phase the data is divided into concurrent segments or they are extracted from the raw data. In this work the syllables of bird sounds are extracted from raw recordings. Feature generation is often called also data reduction phase because in this phase segments are represented with a number of parameters or features. Features are selected so that they include the information that can discriminate different classes. Data reduction can also be included into the feature generation, but then the set of features should be selected for classification from larger number of available features. In classifier design stage classifier is trained with the training data. In this phase decision boundaries of the classes are created. Once the classifier is designed and trained its performance for given classification

task is evaluated in the system evaluation stage.

1.3 Objective and outline of the thesis

Long term objective in the Avesound project is to develop methodology for the system that could automatically recognize bird species or even individual birds by their sound in field conditions. Portable system for automatic recognition of birds would probably have significant impact to the methodology in many areas in biology. Also such system would probably have high commercial interest because bird watching is popular hobby in many countries. Main interest in this project is to study birds that are common in Finland.

Sinusoidal model of the syllables is a feasible representation for tonal and harmonic sounds. However some birds produce regularly sounds that have a more complex spectrum. An example of such species in Hooded Crow (*Corvus corone cornix*). Main goal in this work is to develop feasible representation for inharmonic bird sounds for automatic recognition. In this work sounds which do not fit to the sinusoidal model are called inharmonic.

This thesis in divided into 6 chapters. Chapter 2 focuses to the sound production in birds and to the organization of bird sounds. It also describes shortly the models of the sound production mechanism in birds. Chapter 3 describes structure of the database of the bird sounds developed within Avesound project and methods for segmentation of the syllables from continuous recordings. Chapter 4 presents methods for feature extraction and describes the features of the syllables used in this work. The methods for classification and recognition of the syllables are described in the Chapter 5. Recognition results with different configurations of the recognition system proposed for bird song recognizer are also summarized in Chapter 5. Chapter 6 concludes this work and discusses future directions of this research.

Chapter 2

Bird Sounds

2.1 Sound production mechanism

Main parts of sound production mechanism in birds are lungs, bronchi, syrinx, trachea, larynx, mouth and beak. Airflow from lungs propagates through the bronchi to the syrinx, which is the main source of sound. Sound from syrinx is then modulated by vocal tract, which consist of the trachea, larynx, mouth and beak. In figure 2.1 is presented schematic view of the mechanism. Dimensions of mechanism and parts of it varies considerably among different species, but organization is rather uniform.

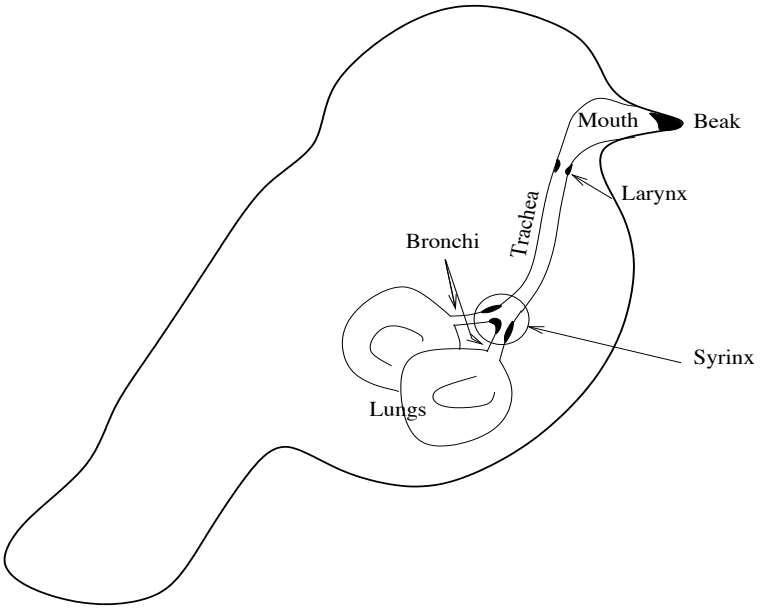

Figure 2.1: Parts and organization of the avian sound producing mechanism.

2.1.1 Syrinx

Syrinx (see Figure 2.2) is the most important and most extensively studied organ in the bird sound production mechanism. Besides that the organ is important in sound production, it has also provided valuable information about the taxonomy of birds because of the differencies in anatomy of the organ in different species. The German anatomist Müller classified bird species by their syringeal anatomy already in 1878 (Müller 1878). Müller limited his studies to Passeriformes, but Beddard (Beddard 1898) took a wider range in his studies. Many later studies have confirmed the classification by Müller and Beddard.

Three different types of syrinx, namely tracheobronchial, tracheal and bronchial, can be found according to distinction between tracheal and bronchial elements of syrinx and topographical position of the main sound producing mechanism. When main sound production mechanism is located in the bronchi it can be in different position in the two bronchi. Tracheal elements are cartillage rings (see Figure 2.2), typically complete, in direct continuation with trachea. Bronchial elements are paired incomplete C-shaped cartilage rings with open ends against each other. Classification into these three classes is however difficult task because intermediate forms are common.

Songbirds (order Passeriform suborder Passeri) are the largest group of the birds, they cover about 4000 out of 9000 total number of birdspecies (Catchpole & Slater 1995). Songbirds and the syrinx of songbirds are most extensively studied among all birds. The syrinx of songbirds is complex in structure but relatively uniform in this group (King 1989) and it can be regarded as the prototype syrinx (Figure 2.2). The syrinx is located in the junction of the trachea and two bronchi and therefore it belongs to the group of tracheobronchial syrinx. When a bird is singing, airflow from lungs makes syringeal medial tympaniform membrane (MTM) in each bronchi to vibrate through the Bernoulli effect (Fletcher 1992). The membrane vibrates nonlinearly opposite to the cartilage wall. Voice and motion of the membrane is controlled by a symmetrical pair of muscles surrounding the syrinx. Membranes can vibrate independently to each other with different fundamental frequencies and modes. Membranes are pressure controlled like a reed in woodwind instruments, but membranes are blown open while the reed in the woodwind instruments is blown closed.

In contrast to the MTM theory recent studies with endoscopic imaging have shown that MTM would not be the main source of sound (Goller & Larsen 1997*b*). Goller suggests that sound is produced by two soft tissues, medial and lateral labia (ML and LL in Figure 2.2), similar to human vocal folds. Sound is produced by airflow passing through two vibrating tissues. Further evidence to this comes from a study where MTM's were surgically removed (Goller & Larsen 2002). After removal birds were able to phonate and sing almost normally. Small changes in song structure however were found, which indicates that MTM's have a function in sound production. However it is possible that birds may be able to compensate

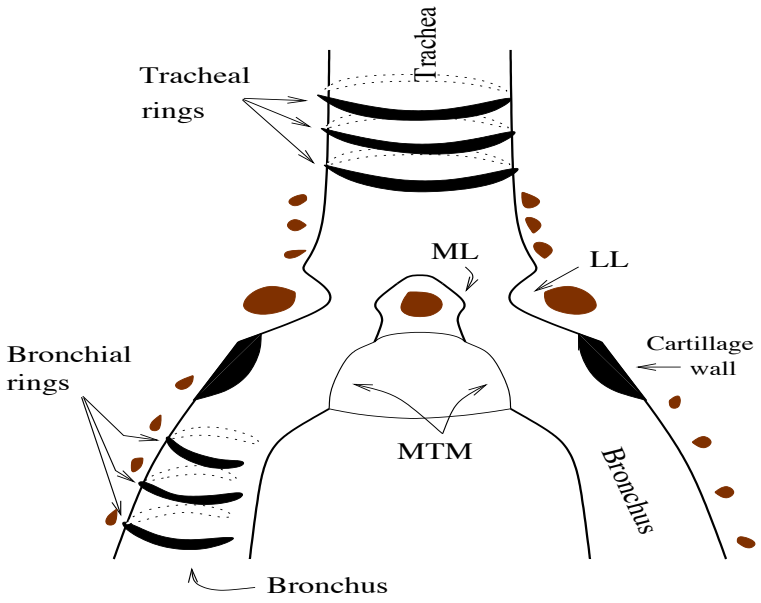

Figure 2.2: Schematic view of songbird syrinx. According to the classical theory sound is produced by the vibrations of the medial tympaniform membrane (MTM) against the cartilage wall. Recent studies suggests that sound is produced in similar way than in the human vocal folds by the medial labia (ML) and the lateral labia (LL).

the loss of MTM.

Also, because of the large diversity in structure of avian syrinx and also in sounds, it is possible that the MTM theory is correct for some species. For example Goller and Larsen limited their study only to cardinals (*Cardinalis cardinalis*) and zebra finches (*Taeniopygia guttata*). In contrast in (Gaunt, Gaunt & Casey 1982) ring doves (*Streptoperia risoria*) were studied as evidence for the MTM theory. Furthermore in (Goller & Larsen 1997*a*) it was found that the main source of sound in pigeons and doves is the tympaniform membrane. However this membrane is located in the trachea and not in the bronchi.

2.1.2 Trachea

The trachea in the birds is a tube between the syrinx and the larynx which acts as a resonator to the sound produced by the syrinx. Elements of the tube are cartilage rings, which are typically complete (McLelland 1989). The number of the tracheal cartilage rings depends on the length of the neck and it ranges from about 30 in small passerines to about 350 in long necked flamingos and cranes. However in the number of species the trachea is arranged in loops or coils so that the length of the trachea is much longer than the length of the neck. It have been argued that the tracheal loops improve transfer function so that the

trachea can have many different vibration modes (Gaunt, Gaunt, Prange & Wasser 1987). In some species the trachea is joined with air sacs or bulbous expansions. In some penguins (*Spheniscidae*) and petrels (*Procellariidae*) the trachea is fragmented into two channels. These species-specific features are responsible for some characteristic sounds in these species.

2.1.3 Larynx, mouth and beak

The larynx in the birds does not include vocal folds like in humans. Only few studies have examined the the function of the larynx in birds. Its function to the sound production is still controversial. The larynx seems to play only little or no role in sound production.

The mouth operates in birds as a cavity resonator like in humans, but it is less flexible. Birds can control the cross-sectional area of the mouth (Fletcher & Tarnopolsky 1999) with the tongue, but only few species, mainly parrots, can use the tongue for sound production like humans (Patterson & Pepperberg 1994) because in most of the birds the tongue is rather stiff.

Analysis of the acoustical behaviour of the beak is a difficult because the form of the beak is rather complex (Fletcher & Tarnopolsky 1999). The analysis cannot be reduced into one or two dimensions without losing vital information. Another difficulty with beak acoustics is its highly dynamic nature. Beak opening and closing change acoustical properties of the beak by changing dimensions of the gape. Recent studies suggest even bigger role for the beak in sound production (Hoese, Podos, Boetticher & Nowicki 2000). Hoese et al. shows that beak opening and closing changes the effective length of the vocal tract, but effect to the vocal tract resonances is nonlinear. Birds can also change dimensions of the vocal tract by movements of the head (Westneat, Long, Hoese & Nowicki 1993).

2.2 Organization of bird sounds

Bird sounds are typically divided into two categories: songs and calls (Krebs & Kroodsma 1980) . Singing is limited to songbirds, but they cover only about half of the birds. Non-songbirds use also sounds to communicate and it is not less important than for songbirds (Beckers, Suthers & ten Cate 2003). Generally songbird sounds are more complex and they have a larger repertoire than non-songbirds, because the ability to control sound production is better (Gaunt 1983).

Diversity of different sounds birds can produce is large. Characteristics of simple voiced sounds are a fundamental frequency and its harmonics. Voiced sounds in birds are closely related to the human vowel sounds in both structure and in a way they are produced. However control of the vocal tract in birds is less complex than in humans. In voiced sounds

in birds fundamental frequency lies between $100Hz$ and $1kHz$ in different species. Birds can emphasize intensities of different harmonics with filtering properties of the vocal tract. Birds can also produce pure tonal or whistled sounds that does not include any harmonics. Both voiced and whistled cases sounds can be modulated in both frequency and amplitude. Amplitude modulations of the fundamental element are mostly generated by the syrinx (Beckers et al. 2003) but intensity differences between harmonics are based on the properties of the vocal tract. Frequency modulation can be divided into two categories: continuous frequency modulations and abrupt frequency jumps. Both frequency modulations are source-generated (Beckers et al. 2003) (Fee, Shraiman, Pesaran & Mitra 1998). In addition, bird sounds can be also noisy, broadband, or chaotic in structure (Fletcher 2000). Characteristic of chaotic behaviour is unpredictability in future waveform even though source and filter conditions are rather well-known. Figure 2.3 shows examples from songs and calls from different species and it illustrates diversity of sounds birds can produce.

2.2.1 Two-Voice Theory

With two independently vibrating membranes in the syrinx, birds can in theory produce two totally independent carrier waves. It have been suggested that this makes possible to sing "internal duet". Different species use two sound sources in sound production in different manner. For example Canaries (*Serinus canarius*) use only one syringeal source to sound production whereas Black-capped chickadees (*Parus atricapillus*) produce complex call notes by using both sources (Nowicki 1997). Three different methods can be found: sound can be produced by either membrane alone, by both acting together or by switching sound source from one membrane to other (Suthers 1990). When both membranes are active together they may generate same or different sound. It is also common for some species that they use all three methods in sound production. First syllable from call sound of great tit (*Parus major*) in the lower center panel in figure 2.3 represented in figure 2.4 is a example from sound whose generation two sound sources were used.

2.2.2 Bird Songs

Generally songs are long and complex vocalization produced spontaneously by males. In few species females also sing and some species sing even duets. Female songs tend to be simpler than song produced by males. Most species sing in a certain time of the year but birds have also a particular daily schedule when they sing. Best time for observing bird singing is in the breeding season at spring. Some birds do not sing at all during the rest of the year. Especially during the breeding season male bird song has two main functions. One is to attract females and the other one is to repeal rival males. In some species songs used

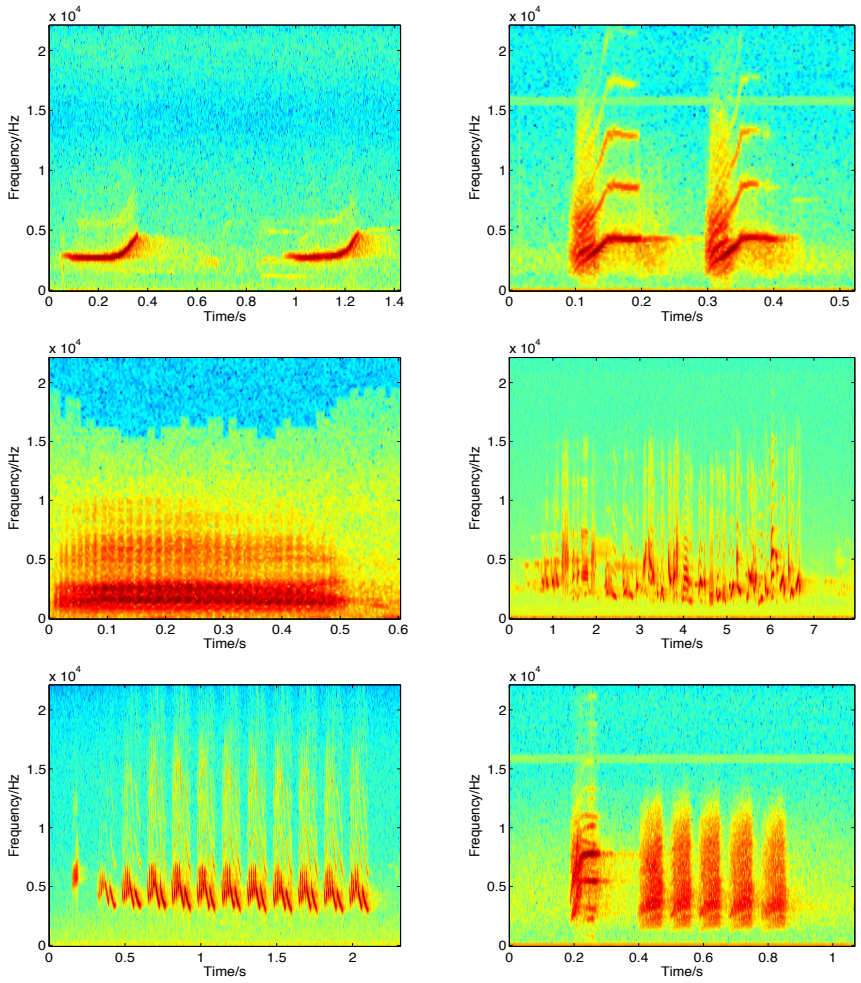

Figure 2.3: Examples of bird sounds from different species. In the upper row are Willow Warbler (*Phylloscopus trochilus*), Common Chaffinch (*Fringilla coelebs*), in the centre Hooded Crow (*Corvus corone cornix*) and Garden Warbler (*Sylvia borin*) and in the lower row Arctic Warbler (*Phylloscopus borealis*) and Great Tit (*Parus major*). The x and y-axis in panels represent time in seconds and frequency in Hz, respectively.

to attract females tends to be longer and more complex than song for territorial defence. Similar features and functions in female song can be found than in male song.

During the day time birds have largest activity in singing at dawn. Several explanations to this have been proposed. Feeding conditions are better after dawn and therefore birds have more time to sing at dawn. It is also best time to take over vacant territories. Female birds are most fertile at dawn and it is best time to copulate. For example it has been observed that great tit (*Parus major*) males sing at dawn until female wakes and then copulates with

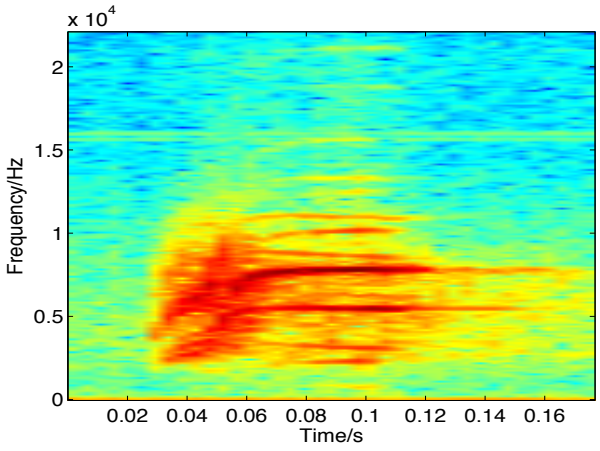

Figure 2.4: Call sound of Great Tit (*Parus major*). Some components of the sound are not in harmonic ratio.

her (Mace 1987). One practical reason is also that conditions for sound transmission are favourable at dawn, because wind and air turbulence are reduced.

Transmission conditions are important otherwise also and those set limitations to the structure of sound. Two fundamental phenomena that affect the transmission of sound are attenuation and degredation, which is a problem especially in the dense environments. Sound propagation properties are different in different environments and also at different heights in a specific environment. Bird sounds adapt to environmental conditions so that sound is transmitted to receiver optimally. Optimality condition depends on the function of the sound and it does not always mean maximal distance.

Hierarchical levels of bird song are phrases, syllables and elements or notes. Elementary building unit of bird song is called element, which is the smallest separable element in spectrogram (see Figure 2.5). Elements are building blocks of syllables, that may be produced by one or more elements or notes. The structure on syllables varies a lot and therefore also the number of elements in syllables. Series of syllables that occur together in a particular pattern is called a phrase. Syllables in a phrase are typically similar to each other, but they can also be different like in the last phrase (end phrase) in figure 2.5. A song is constructed of a series of phrases. When a bird changes the order or types of the phrases in the songs the bird is said to have different types of songs and a repertoire of song types. Repertoire size varies typically from few to several hundred song types in different species.

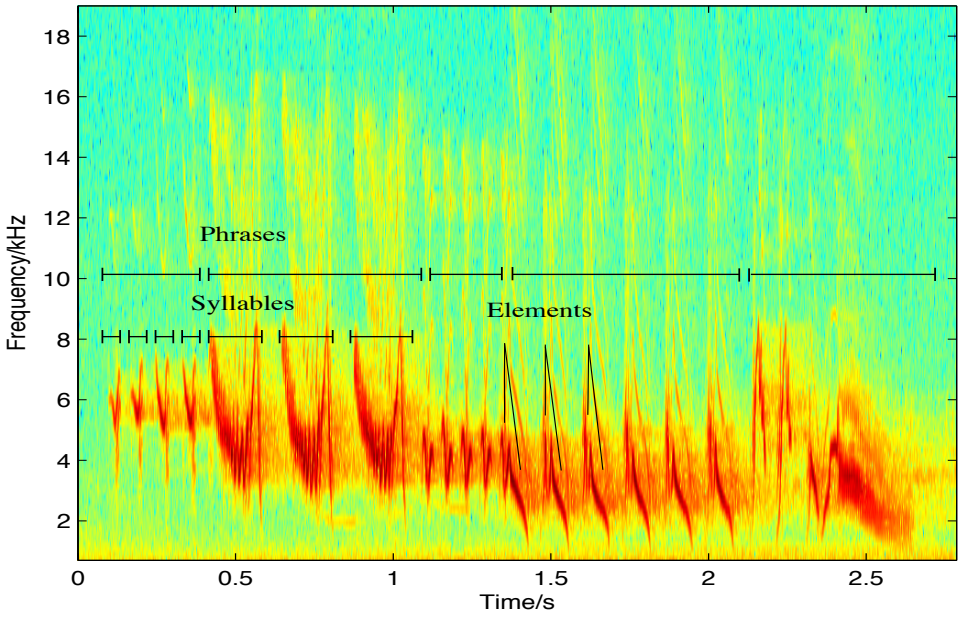

Figure 2.5: Hierarchical levels of common chaffinch (*Fringilla coelebs*) song. The y-axis represents frequency in Hz and x-axis time in seconds.

2.2.3 Bird Calls

Bird calls are usually short and simply, but they can also be complex and can sometimes be confused with simple songs, especially when series of call sounds are connected. Calls typically occur in specific context and carry some function and they are produced by both sexes through the year. Calls have a large functionality and at least 10 different call categories (e.g. alarm, flight and feeding call etc.) can be found. Furthermore some birds have more than one call for one category and some use very similar calls for different meaning. Call sounds are important for songbirds also and generally they have greater repertoire of call sounds than non-songbirds.

2.3 Models of sound production

Sound production models in birds were first studied by Greenewalt (Greenewalt 1968), whose work has constituted the basis to the studies on the classical model of bird sounds. A similar source-filter model can be used to model avian sounds than is used in speech or wind instrument modeling. A difference to speech production is that birds may have two independent sound sources in the syrinx. Sound from the syringeal source is then filtered

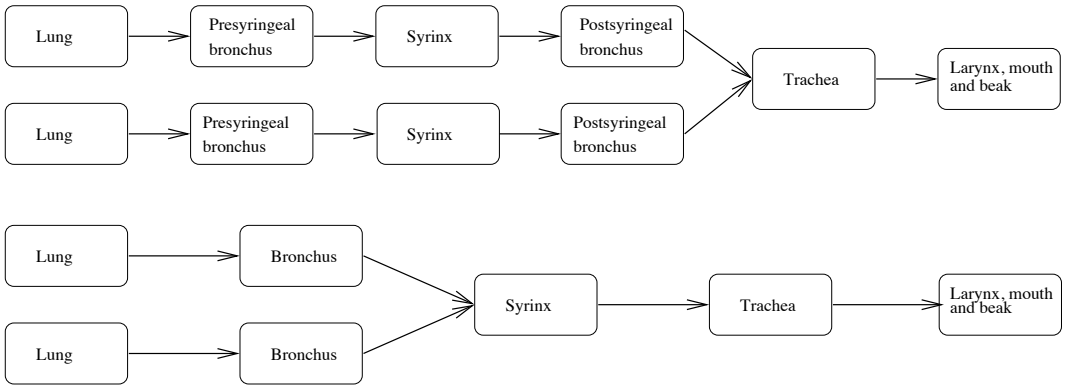

Figure 2.6: Two models of avian vocal tract.

by the vocal tract. Two different models of avian sound production system are presented in figure 2.6.

2.3.1 Models of sound source

In the classical model of the syrinx tympaniform membrane is assumed to be the main source of oscillation. Vibration is driven by membrane's natural frequency and air column resonance in bronchi. Membrane motion against the cartilage wall changes the cross-section area of bronchi causing a nonlinear change in pressure and air flow. The pressure at tracheal side of the syrinx p_1 depends on pressure at the bronchial side of the syrinx p_0, air density ρ, displacement of the membrane u, radius of the bronchus r and air flow through syrinx U and is given in (Fletcher 1988) as

$$p_1 = p_0 + \frac{\rho}{2}\left[\left(\frac{U}{2ru}\right)^2 + \frac{1}{\sqrt{2ru}}\frac{dU}{dt}\right] \tag{2.1}$$

The airflow U from lungs varies in time as a bird is breathing. Membrane can be modeled as a simple taut membrane, whose displacement can be calculated as function of driving force F against the membrane as:

$$m\left[\frac{d^2u}{dt^2} + 2\kappa\frac{du}{dt} + f^2(u - u_0)\right] = \epsilon F \tag{2.2}$$

where f is the mode frequency in Hz, κ is the damping coefficient, m is the effective mass of the membrane associated to the mode and u_0 is position of the membrane at rest. Coefficient ϵ is small constant term, which is referred to the coupling between F and the mode. For the driving force in (2.2) Fletcher gives

$$F \approx 2Crh\left(\frac{p_0 + p_1}{2} - \frac{\rho U^2}{\sqrt{rw^3}}\right) \tag{2.3}$$

where C is constant term order of unity and h is length of the membrane.

The model is successful with voiced sounds but incapable to produce whistled or tonal sounds. A one string model to model also tonal sounds was suggested in (Casey & Gaunt 1985). In this model the membranes in former model are stretched into the shape of a string. The string model is capable to produce sound with one fundamental and its harmonics. In (Doya & Sejnowski 1995) these two models were mixed so that the string model produces tonal harmonic sounds and a mixture of sound from both models produce noisy components.

As mentioned earlier, recent studies have suggested that the sound of birds is produced by the tissue folds similar to the human vocal folds (Larsen & Goller 1999). Also recent models of bird sound production have been build on models of the human vocal folds (Gardner, Gecchi & Magnasco 2001). Gardner's two-mass model is a simplification to geometrical dimensions of the folds. In the model it is assumed that the folds are controlled by bronchial pressure p_0. Averaged pressure at tracheal side of the folds can be calculated as a function of bronchial pressure and position of the folds:

$$p_1 = p_0\left(1 - \frac{u_a}{u_b}\right) \tag{2.4}$$

where u_a and u_b are calculated in terms of phenomenological constant τ and position of the center of the folds u as:

$$u_a = u_{a0} + u + \tau\frac{du}{dt} \tag{2.5}$$

$$u_b = u_{b0} + u - \tau\frac{du}{dt} \tag{2.6}$$

Position of u can be calculated as given in (Laje, Gardner & Mindlin 2002):

$$\frac{d^2u}{dt^2} - \left(cu^2 - p_0\right)\frac{du}{dt} - ku - F = 0 \tag{2.7}$$

where k is the restitution constant, c is the dissipation constant, p_0 is the driving pressure and F is force term against the vibrating labia.

2.3.2 Models of vocal tract

Relatively little has been done on modeling of the bird vocal tract although its essential role in sound production has been discovered for example in (Nowicki 1987) and (Brittan-Powell, Dooling, Larsen & Heaton 1997). In (Fletcher & Tarnopolsky 1999) the acoustics

of the vocal tract of Oscine birds has been studied. Although Fletcher studies limits only to song birds, models can be easily modified to correspond to many other birds. In model both syringeal sound sources are first connected to the bronchial tube that leads to the trachea. Both bronchi and trachea are modeled with an acoustical impedance matrix whose coefficients can be calculated by

$$Z_{11} = Z_{22} = -j\frac{\rho v}{A}\tan kl \tag{2.8}$$

$$Z_{12} = Z_{21} = -j\frac{\rho v}{A}\csc kl \tag{2.9}$$

where ρ is the air density, v is the speed of sound, A is the cross-sectional area of the tube, l is the length of the tube and $k = \omega/v + j\alpha, \omega = 2\pi f$ is the wavenumber. f is the frequency in Hertz and α is the attenuation coefficient for sound propagating in tube. The input impedance for the system that includes two bronchi and the trachea is given by

$$Z_{in} = B_{11} - \frac{B_{12}^2(B_{11}^{'}T_{22} + T_{11}T_{22} - T_{12}^2)}{B_{11}^{'}(B_{22}T_{22} + T_{11}T_{22} - T_{12}^2) + B_{11}(T_{11}T_{22} - T_{12}^2)} \tag{2.10}$$

where B and $B^{'}$ refer to the two bronchi and T refers to the trachea.

Fletcher presents also a models for the larynx, mouth and beak. The larynx is modeled by a simple series impedance $L = j\omega\rho l/A$, where l is length and A is cross-sectional area of the larynx. The mouth can be modeled in a similar way that is used for models of the human mouth. Fletcher considers the mouth as a short tube, with varying cross-sectional area controlled by raising and lowering of the tongue.

For the beak Fletcher provides a simplified conical model. The main motivation in this model is that it can be solved analytically

$$K(f,g) = j\frac{\rho c}{A_B}\left[\frac{\csc^2(k\delta/2)}{\cot(k\delta/2) - k\delta/2} - \cot\left(\frac{k\delta}{2}\right)\right] \tag{2.11}$$

where $k = 2\pi f/c$ and A_B is cross-sectional area of peak base. δ is end correction based on measurements with a light sheet-metal beak model and it is given by terms of length of the peak l_B, frequency f and tip gape g as

$$\delta \approx 0.05 l_B + 10^{-5} f l_B^2/g \tag{2.12}$$

Mixing all elements to a network Fletcher gives the final input impedance as

$$Z_{in} = T_{11} - \frac{T_{12}^2(M_{22} + K)}{(T_{22} + M_{11} + L)(M_{22} + K) - M_{12}^2} \tag{2.13}$$

where T, L, M and K refer to input impedances of the trachea, larynx, mouth and the beak, respectively.

Chapter 3

Material and Segmentation Methods

3.1 Bird sound database

Material used in this work is stored in XML based database, which is a collection of bird sounds from many different species and sources. The database system has been developed at HUT/Acoustics since 2002 in the context of the Avesound project (Fagerlund 2004) and it has been partially funded by *Jenny and Antti Wihuri foundation*. Most of the recordings have been made in Finland in field conditions. In addition, some of recordings were taken from Finnish and foreign commercial CD-collections. Original audio files include often environmental sounds and other unwanted sound events. Therefore some preprocessing is needed to extract true bird sounds from raw recordings.

Altogether database includes approximately 2400 audio files in total from 289 different bird species. Each file contains a song, call sound or series of call sounds isolated from raw recordings. Files are in WAVE format with sampling frequency of $44.1kHz$. Audio files are located in a folder accordant with the species and further species are in a folder accordant with the order. Widely used abbreviation derived from the Latin name of species is used as name for folder of species. For example abbreviation for Sedge Warbler (*Acrocephalus Schoenobaenus*) is ACRSCH. Names of audio files are in format 'r#$_i$s#$_j$.wav', where 'r#$_i$' refers to the i:th recording and 's#$_j$' refers to the j:th sound event (song, call, series of calls) extracted from the recording i. Each species has an data description entry based on the eXtensible Markup Language (XML), which describes the material and has reference to the audio files.

3.1.1 XML-entry

Description of recordings from species are stored in the XML-file. Father element of the file is **recordings**, who has child elements **species** and **recording**. Element species is simply

the name of the species (same Latin name derived abbreviation is used as for the name of the folder and XML-file). Recording element is created from each individual raw audio file and it includes information related to that particular recording. Recording element has attributes **author** and **editor**, who refers to the author of the recording and person who has added recording to the database. Each recording has elements **date**, **location**, **wording** and **files**. Date is the date when original recording have been made and location refers to the place. Wording includes description of the sound event and it can also include some notable additional information, for example information on the recording equipment used. Element files include elements **raw** and **songs**, which further include element **song** with attributes **id** and **time**, which refers to the names of audio files related to the recording and to the date when files have been added to the database respectively. Element raw is optional and it refers to the original audio file.

In figure 3.1 is a XML-file of two recordings of species Common Chaffinch (*Fringilla Coelebs*). Recordings have taken place in different locations and dates. Both were recorded by Mr. Ilkka Heiskanen and added to the database by Mr. Aki Härmä. Three and five sound files respectively have been extracted from original audio files PPO22488.wav and PPO27488.wav.

3.1.2 Increasing the database

Though total number of data in the database is quite large only for few species number of recordings and bird individuals is large enough for reliable classification and recognition experiments. Therefore it is necessary to increase the number of recordings in the database. This is however very time-consuming work. Currently addition of new entries to the database is done semi-automatically. First candidates for bird sounds are extracted from raw audio files automatically using the 'Birdcapture'-tool developed in the project. This is done by tracking the maximum peak of the spectrogram. When peak exceeds the threshold new audio file is written and XML-file is updated with new entry. After raw datafile is examined candidates are looked through by hand and they are either accepted or rejected to the database.

3.2 Segmentation of syllables

Segmentation is preceding phase for the analysis and classification of audio data. In segmentation raw data is divided into distinct objects. Segmentation is always based on some change in the characteristic features of the data. There are several reasons why segmentation is performed. It is easier to build analysis and classification systems for segmented objects than for raw data. It reduces data space and calculations in analysis phase if it is

```
<?xml version="1.0" encoding="ISO–8859–1"?>
<?xml–stylesheet href="BDBstyle.xsl" type="text/xsl"?>

<recordings>
 <species> fricoe </species>
 <recording author="Ilkka Heiskanen" editor="Aki Härmä">
  <date> 22.4.1988 </date>
  <location>Nastola Ruuhijärvi </location>
  <wording>  Tint–sound, song etc. convey of migrating birds 22.4.1988 Nastola Ruuhijärvi.
On the background Common Crossbill, buzz etc. Speech.</wording>
  <files>
    <raw>  PPO22488.WAV </raw>
    <songs>
      <song id="r12s1" time="Oct 31 2003"> </song>
      <song id="r12s2" time="Oct 31 2003"> </song>
      <song id="r12s3" time="Oct 31 2003"> </song>
     </songs>
  </files>
 </recording>

 <recording author="Ilkka Heiskanen" editor="Aki Härmä">
  <date> 27.4.1988 </date>
  <location> Iitti Vuolenkoski </location>
  <wording> Tint–sound and song 27.4.1988 frosty morning after 7 o'clock Iitti
Vuolenkoski. On the background Hooded Crow, Black Grouse etc. </wording>
  <files>
    <raw> PPO27488.WAV </raw>
    <songs>
      <song id="r13s1" time="Oct 31 2003"> </song>
      <song id="r13s2" time="Oct 31 2003"> </song>
      <song id="r13s3" time="Oct 31 2003"> </song>
      <song id="r13s4" time="Oct 31 2003"> </song>
      <song id="r13s5" time="Oct 31 2003"> </song>
    </songs>
  </files>
 </recording>

</recordings>
```

Figure 3.1: XML-entry of species Common Chaffinch (*Fringilla coelebs*)

done after segmentation.

Depending on the application segmentation can be performed in different phases of analysis or classification. In general audio classification and annotation segmentation is typically done after feature calculation. In these applications goal is to index long periods of audio data so that they can be easily and fast searched and retrieved by information retrieval programs (Wold et al. 1996). Calculating features for long periods of the data can attain better performance of segmentation phase if features of classes are sufficiently stationary and criterion for the detection of borders is good. However in this case more calculations are needed compared to the simple temporal segmentation. Segmentation can be also calculated in parts. For example in (Li et al. 2001) first segmentation phase is for distribution of raw data into silent and signal parts. In second phase signal part is further segmented into distinct objects by means of higher level features than used for silence/signal segmentation.

In this work in segmentation phase syllables of bird sounds are extracted from audio files in the database. Segmentation is performed before actual feature calculation, because only a parametric representation of syllables is needed, and there is no need to calculate features for silent or no syllable parts. Also temporal and spectral changes of syllables are rather diverse, which would cause errors in feature-based segmentation. On the other hand, detection of boundaries of syllables is difficult because adjacent syllables can overlap in time and frequency, and because the onset and offset of the sound is often slow and actual onset and offset may occur below the background noise level.

Segmentation is a critical part for the subsequent steps of classification. Unsuccessful segmentation would result unsuitable syllable candidate because parametric representation would be different for syllables similar to each other in real world. This increases noise in the whole classification system. Another requirement for the successful parametrization of syllables is accurate and unambiguous detection of onsets and offsets.

In this work we tested two different approaches to the segmentation of syllables. Methods are described below. In the first method segmentation is done purely in the time domain and based on the energy envelope of the audio data. In the second approach short-time spectral representation is first applied. Segmentation is done based on the magnitude of the maximum bin of the spectral representation.

3.2.1 Segmentation based on the short-time signal energy

Segmentation method is based on the energy of audio file and the noise level estimate. Method is straightforward and fast, because calculation is done in temporal domain. Segmentation is performed in three phases. In the first phase onset and offset of the syllable candidates are calculated. Near onset and offset areas of syllable candidates energy envelope curve commonly fluctuates around the threshold, which causes short erroneous can-

didates. This phenomena is called *border effect* (Li et al. 2001) and it is well-known in audio segmentation and classification. Also syllables of bird sounds can be constructed of temporally distinct elements that are detected separate syllable candidates. In second phase syllable candidates enough close to each other in temporal domain are connected to a single syllable. Last phase include true syllable detection where too short candidates are omitted. Omitted candidates are often caused by random fluctuation of the noise level above and below the threshold.

Onset/offset detection

Energy envelope is calculated for the input audio file. The file is first divided into the overlapping frames. In this work the frame size was 128 samples corresponding $3ms$ and adjacent frames overlap 50% (stepsize 64 samples). Distance between adjacent syllables can be as short as $20ms$ the and distance between the pulses of a syllable can be even shorter. This causes a limit to the frame length that can be used. Efficient discrimination of syllables requires at least few samples between them. By overlapping frames, a better time resolution is obtained with same frame length.

The signal frames used for the computation of the signal energy are windowed using the hanning window. The maximum energy of the envelope is normalized to $0dB$. The energy envelope $E(m)$ of the signal $x(n)$ in decibel scale is calculated as

$$E(m) = \sum_{i=1}^{N} 20 \log_{10} |x_i[n]|^2 \tag{3.1}$$

where $x_i[n]$ is i:th frame and N is the total number of frames of the signal $x(n)$.

Noise level is calculated iteratively. First initial noise level estimate is set equal to the minimum energy of the envelope and threshold for onset and offset is set to the half of the noise level. The flow diagram of iterative syllable search algorithm is presented in figure 3.2. Noise estimate is the average energy of the non-signal frames ,i.e., frames that are not assigned to the syllable.

Merging syllable candidates

Border effect is common phenomenon in general audio segmentation and also in syllables of the birds sounds. Syllables that consist of temporally distinct pulses are common for some species. Both of these phenomena can be found in songs of Sedge Warbler (*Acrocephalus Schoenobaenus*). Figure 3.3 shows a sequence of eight syllables from the song of Sedge Warbler. First two syllables (energy envelope presented in top right panel of figure 3.3) of the song consist of two pulses which are detected as separate by the syllable search

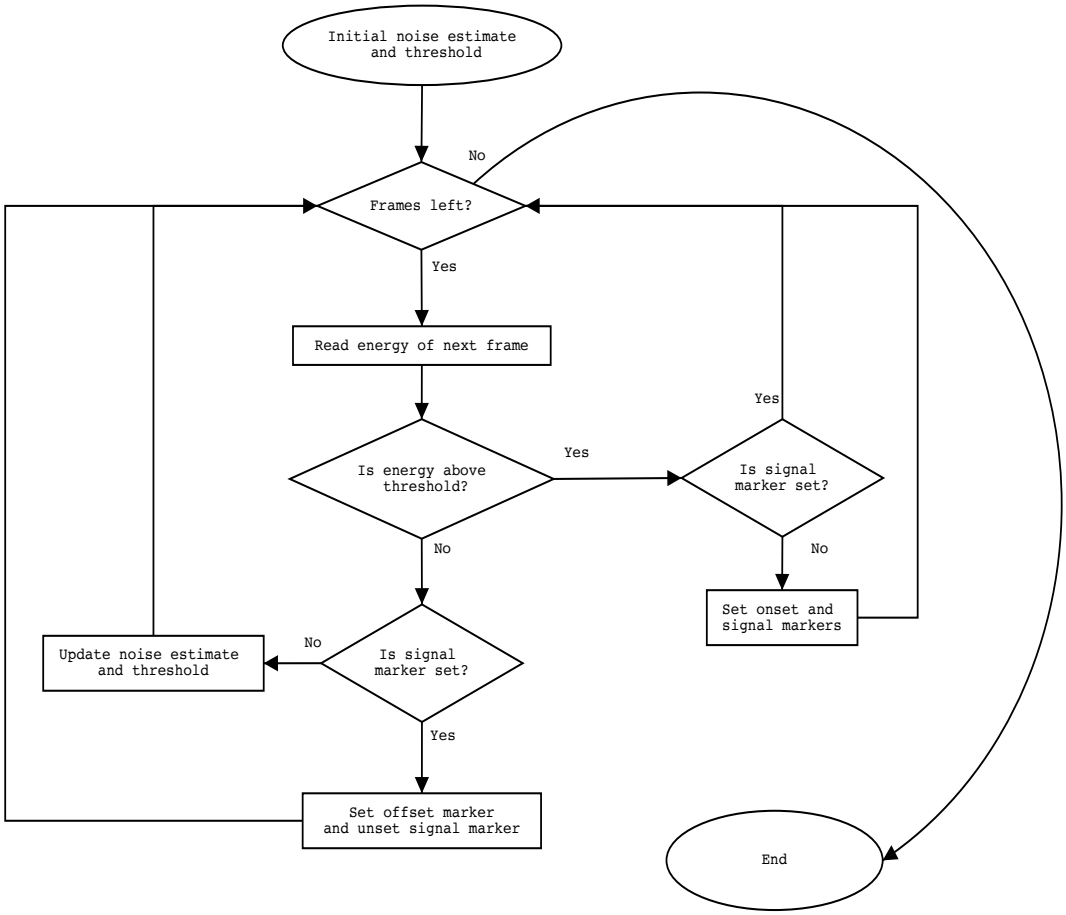

Figure 3.2: Flow diagram of syllable search algorithm.

algorithm. Border effect occurs in segmentation of two last syllables (energy envelope presented in bottom left panel) of the song.

In syllable merging phase syllable candidates that are very close to each other are grouped together. Distance between syllable candidates that belong together is typically very short. Threshold value used here is $15ms$ corresponding 10 overlapping frames. Segmentation result after syllable merging phase is presented in bottom right panel of the figure 3.3.

3.2.2 Segmentation based on the short-time spectrum maximum

Segmentation is performed by tracking the maximum of the spectrogram of the audio signal. Many bird sounds are sinusoidal or are constructed of sinusoidal components. Maximum of spectrum should detect such sounds well. For silent parts the value of the maximum of the spectrum is relatively low and it should not confuse the detection of syllables.

Noise level is estimated to adjust threshold for syllable-to-silence discrimination to the

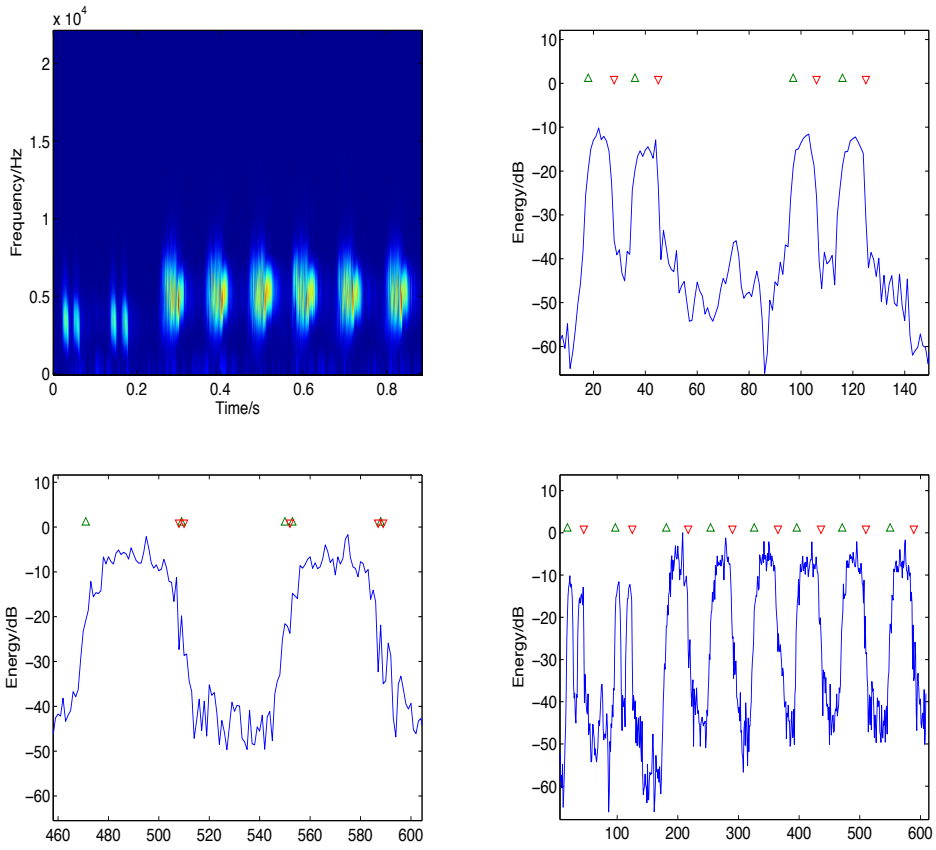

Figure 3.3: Segmentation of bird song using short time signal energy method. Top left panel shows spectrogram of eight syllables of a song of Sedge Warbler (*Acrocephalus Schoenobaenus*). Top right and bottom left panels shows segmentation result before merging syllable candidates phase. Panels illustrates respectively syllables consist of two pulses and the border effect in two pulses. Final segmentation result is presented in the bottom right panel.

suitable level. Syllable search is done iteratively starting from strongest syllable continuing to weaker ones. It is common for this method also that syllables that are constructed from distinct pulses are detected separate syllable candidates in the first place. Same method for merging of the syllable candidates that is used in previous segmentation method is used in this context also.

Maximum of spectrum

Spectrogram is calculated using the Short Time Fourier Transform (STFT). First, bird sound is divided into overlapping frames of 256 samples ($6ms$) with the step size of 64 samples,

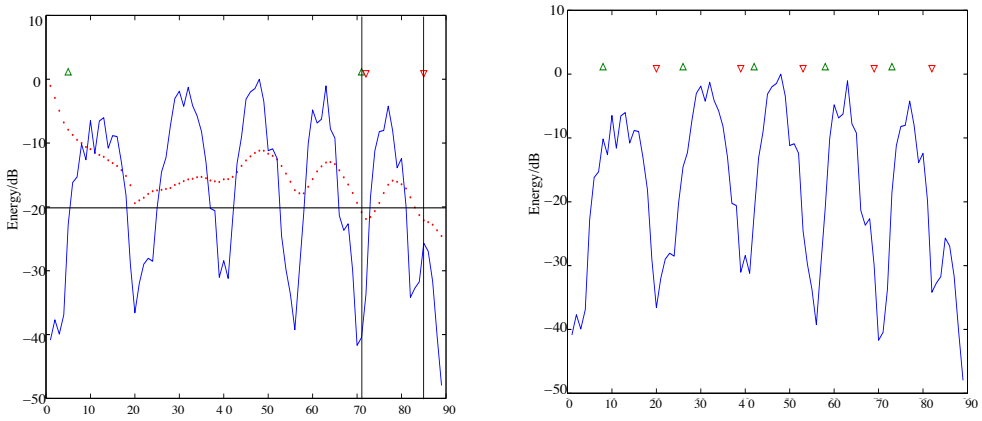

Figure 3.4: Segmentation of syllables of birds using short time spectrum maximum method. dotted curve in the left panel is filtered maximum of the spectrogram curve. New noise level estimate is set here to the value $-20dB$. Final segmentation result is presented in the right panel.

thus frames overlap by 75%. Longer frame size is used than in previous method in order to get a better spectral resolution. Each frame is windowed with hanning window and zero-padded to the size of 1024 samples. Frequency domain representation is calculated for each frame using the fast Fourier transform (FFT). For each transformed frame, the frequency bin with the highest absolute value is selected. The maximum of spectrum curve is represented on a decibel scale so that maximum value is normalized to $0dB$.

Syllable search

Syllable search is done in two phases. In the first phase preliminary syllable candidates are identified based on initial threshold for syllable, which is set to $-30dB$. Next, the estimate of the background noise level is calculated in the following way: The maximum of the spectrum curve is first low-pass filtered and then such value is chosen for the noise level that one or more preliminary syllable candidates are between lowest and highest indices of intersection of the noise level estimate and filtered maximum of the spectrogram curve (see left panel of figure 3.4). New threshold for onset and offset of the syllable is set to the half of the noise level estimate. Syllable search is done again using the same algorithm than in the first phase, but with a new threshold. The segmentation results after first and second phase are illustrated in figure 3.4. The new noise level is set here to the value $-20dB$ and threshold to the $-10dB$.

Syllable search algorithm starts with finding the maximum value of the maximum of spectrum curve. From this point algorithm reads values around it until four adjacent bins

are below threshold. By this border effect is reduced because algorithm does no stop at once the value of the maximum of spectrum curve is below threshold, but it needs to be below threshold for a short time. Once syllable is found it's start and end indexes are saved and segment is deleted before algorithm starts to search for the next syllable. Algorithm stops when highest peak of curve is below threshold.

3.2.3 Comparison of the segmentation methods

Both methods are able to discriminate sufficiently distinct and strong syllables accurately, regardless of the type of the syllable. Maximum of spectrum method does not use frequency information in the segmentation, only the magnitude of the maximum value of each frame, and from this perspective it can be regarded as temporal domain method. Because neighbourhood of the maximum of spectrum contains most of the energy of the sound for majority of the sounds energy envelope and maximum of the spectrum curves are very similar after normalization, thus performance of segmentation methods is quite similar for many typical bird sounds.

The biggest difference between segmentation methods is in syllable search algorithm and estimating the noise level. Essential difference is that in maximum of spectrum method syllable search is started from the strongest syllable whereas in energy based method search starts from the first syllable. Noise level estimate and threshold for syllable detection is updated during syllable search in the energy based method, but in maximum of spectrum method same threshold value is used throughout the syllable search. Iteratively updated threshold improves performance of segmentation result of bird songs. Because overall energy is often higher in the middle of the song than in the beginning and the end, higher threshold value is needed to discriminate syllables in the middle of the song..

Segmentation of syllables constructed of short pulses

In maximum of spectrum method short syllables are omitted in context of syllable search and before candidate merging phase. This causes segmentation errors within songs that are constructed of very short pulses like in songs of Grasshopper Warbler (*Locustella naevia*), whose syllables are consisted of two or three pulses. First or first two, in case when there is three pulses in the syllable, are very short, typically less than $4ms$. These are omitted in syllable search case, because pulses are detected separate and merging them is done after removal of the short candidates.

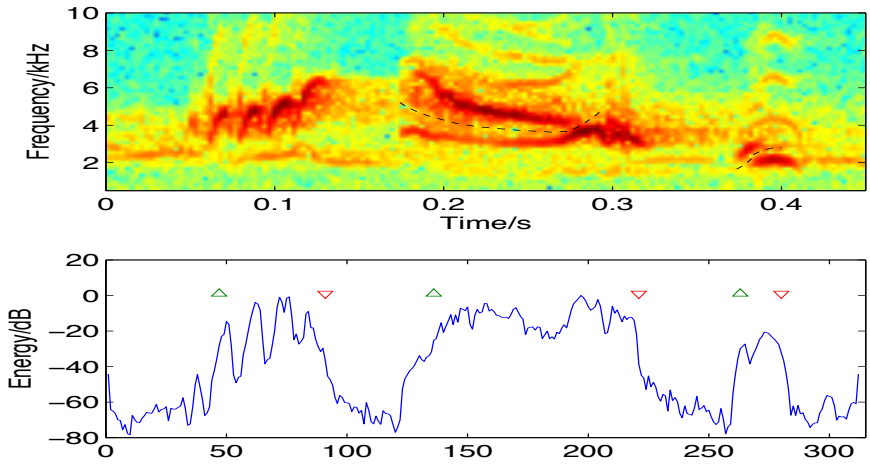

Figure 3.5: Segmentation of bird song consist of overlapping syllables. Used segmentation method have identified three syllables, but actual number of syllables is five. Actual syllable boundaries are denoted with dashed line.

Warbling and overlapping sounds

Syllable boundaries cannot be clearly defined in warbling and overlapping sounds. In warbling singing syllables follow each other without clear pauses in between them. Detection of syllable boundaries in these cases based on only temporal domain information is difficult. Syllables that overlap in the temporal domain usually overlap only partly. This occurs for example when a syllable begins before the end of the previous syllable. These syllables and also adjacent syllables that starts and ends at the same time instant are often separable in the frequency domain. There are two physical explanation for overlapping sounds: source based and environment based. In the first case syllables overlap already in the syrinx. This is possible with two independent sound sources in the syrinx. In the second case overlapping occurs due to reflections from trees and other environmental objects. This can be seen as expansion of the syllable in time domain.

In context of both, warbling and overlapping sounds, energy is usually slightly smaller between syllables, but it can be easily confused with other variation of syllables and cannot be the only feature for reliable segmentation. Part of song of Garden Warbler (*Sylvia borin*) presented in figure 3.5 shows examples of syllables that overlap in the time domain. In upper panel is presented spectrogram of the song and in lower panel is segmentation result with the energy method. Second identified syllable consist of two syllables, which cannot be detected separate with time domain methods because they are fully overlaping. Third identified syllable consist of two syllables whose offset and onset overlap.

Chapter 4

Features

The objective in pattern recognition or classification is to classify objects (patterns) into number of categories (classes) (Theodoridis & Koutroumbas 1998). In this work syllables extracted from songs and calls of birds are used as patterns. Classification is done based on the features, which are calculated from the syllables to be classified or recognized. Features constitute a feature vector, which is a representation of the syllable. Features are generated in three phases. First is simply calculation of features of patterns (raw data), which is followed by the removal of outliers, clearly errorneous syllables. In data normalization feature values are adjusted to the same dynamic range so that each feature has equal significance to the classification result. Classifier could be also trained with unnormalized data, but this may require more training data. Trainig of the classifier could also take more time with unnormalized data.

The number of possible features for a classification problem is usually large and it is often necessary to reduce. There are several reasons to reduce number of features to sufficient minimum so that classification result does not decrease. With less features computational complexity and number of free classification parameters decreases. Also generalization properties are better and noise of classification system decreases. Irrelevant features adds noise to the system and they can impair the classification result.

Selection of actual features used in classification is a critical part for the whole classification system. The aim is to select features with large between-class and small within-class discriminative power. Discriminative power of features or feature sets tells how well they can discriminate different classes. Feature selection is usually done by examining discriminative capability of individual features. Also linear and nonlinear transformations of feature vector can lead to the feature vector with better discriminative power. When number of classes is large selection of features may be challenging when certain features are able to discriminate certain classes but assign only small discriminative power to some other

classes.

4.1 Acoustical features

Features used in sound classification and recognition applications are usually chosen such that they represent some meaningful physical characteristics of the sound production, or some perceptually relevant aspects of an audio signal. Physical features, such as spectral centroid, signal bandwidth and zero crossing rate are mathematical characteristics of the sound and are calculated mathematically from sound wave. Perceptual features refers to the sensation of sound by humans. Examples of perceptual features are loudness, pitch and brightness. Usually there is close relation between a perceptual feature and one or more physical feature of the sound.

Further features can be divided into temporal (time domain) and spectral (frequency domain) features. Temporal domain features, such as zero crossing rate and signal energy, are calculated directly from the sound waveform. In the case of spectral features signal is first transformed to the frequency domain using Fourier transform and features are calculated from transformed signal.

Low-level acoustical signal parameters are used as features of the syllables. Characteristic to these features is that they are straightforward and fast to compute. Low-level features have been previously used for example in classification of general audio content (Li et al. 2001, Wold et al. 1996), music genre classification (McKinney & Breebaart 2003) and in speech to music discrimination (Scheirer & Slaney 1997). In these applications the number of classes is usually relatively low, which is not the case in classification of bird species. However diversity within bird sounds is large and features used in here could be used in classifying sounds into classes by their acoustical features.

Most of the features are calculated on frame basis. This is common in audio and speech analysis, because the amount and variability of data is reduced. First, syllables are divided into overlapping frames. Features are calculated from windowed frames, which results feature trajectories of the syllable. Mean and variance values of trajectories are calculated, thus each basic feature results in two actual features. Final feature vector include mean and variance values of frame based features plus parameters calculated from the entire syllable.

4.1.1 Spectral features

Frequency range is calculated from the entire syllable. All other spectral features are colculated on the frame basis and they provide short time spectral properties of the syllable. Frame size of 256 samples with 50% overlap is used. Fourier transform is applied to signal frames that are windowed with Hanning window.

Spectral Centroid (SC)

Spectral centroid is center point of spectrum and in terms of human perception it is often associated with the brightness of the sound. Brighter sound is related to the higher centroid. Spectral centroid for signal frame is calculated as:

$$SC = \frac{\sum_{n=0}^{M} n|X(n)|^2}{\sum_{n=0}^{M} |X(n)|^2} \tag{4.1}$$

where X is discrete Fourier transform (DFT) of signal frame and M is half of the size of DFT. Resolution of spectral centroid is the same than resolution of DFT, which is $F_s/(N-1)Hz$, where F_s is the sampling frequency and N is the size of the DFT frame.

Signal bandwidth (BW)

Signal bandwidth is defined as a width of the frequency band of signal frame around center point of spectrum. The bandwidth is calculated as

$$BW = \sqrt{\frac{\sum_{n=0}^{M}(n-SC)^2|X(n)|}{\sum_{n=0}^{M} |X(n)|^2}} \tag{4.2}$$

where SC is spectral centroid given in (4.1) and X and M are as in equation (4.1). The bandwidth of syllable is calculated as average of bandwidth of DFT frames of syllable.

Spectral rolloff frequency (SRF)

Spectral rolloff frequency is the point below which certain amount of power spectral distribution resides. Feature is related to "skewness" of spectral shape. The measure can distinguish sounds with different frequency ranges. Spectral rolloff frequency for a DFT frame is defined as

$$SRF = \max\left(K|\sum_{n=0}^{K}|X(n)|^2 < TH\sum_{n=0}^{M}|X(n)|^2\right) \tag{4.3}$$

where TH is the threshold between 0 and 1. Here we use a commonly used value 0.95.

Delta spectrum magnitude (spectral flux) (SF)

Delta spectrum magnitude measures difference in spectral shape. It is defined as the 2-norm of difference vector of two adjacent frame spectral amplitudes. It gives a higher value for syllables with a higher between-frame difference. Formula for delta spectrum magnitude calculations is given as

$$DSM_i = \sum_{n=0}^{M} |||X_i(n)| - |X_{i+1}(n)||| \tag{4.4}$$

Spectral flatness (SFM)

Spectral flatness measures the tonality of a sound. It gives a low value for noisy sounds and a high value for voiced sounds. Measure can discriminate voiced sounds from unvoiced also if they occupy same frequency range. Spectral flatness is the ratio of geometric to arithmetic mean (Markel & Gray 1976) of signal spectrum and it is given in dB scale as

$$SF = 10 \log_{10} \frac{G_m}{A_m} \tag{4.5}$$

where $G_m = \sqrt[1/M]{\prod_{i=0}^{M} |X_i|}$ is geometric mean and $A_m = 1/M \sum_{i=0}^{M} |X_i|$ is arithmetic mean of the magnitude values of the spectral points X_i. M is half of the size of the DFT.

Frequency range (range1, range2)

Frequency range gives low and high limit value of the frequency range that a syllable occupies. Frequency range is calculated for the whole syllable. The frequency range and the length of the syllable together defines boundaries of the syllable. Frequency range is calculated by means of normalized power spectrum of the syllable. Low and high limits are respectively the lowest and highest frequency bin whose power spectrum value in dB scale is above a threshold. The threshold value used here is $-40dB$.

4.1.2 Temporal features

In addition to the features described below, the temporal duration of the syllable (T) is also used as the feature of the syllable. The zero-crossing rate (ZCR) and short time signal energy are calculated on frame basis. The size of a frame is 256 samples and adjacent frames overlap 50% as it was also for the spectral features. Frames are windowed with rectangular window.

Zero-crossing rate (ZCR)

Zero-crossing rate (ZCR) is number of time domain zero-crossings in processing frame. A zero-crossing occurs when adjacent samples have different signs. ZCR is closely related to spectral centroid as they both measure construction of spectral shape of frame. It is defined for the frame as

$$ZCR = \sum_{n=0}^{M-1} |sgn(x(n)) - sgn(x(n+1))| \tag{4.6}$$

where x is time domain signal frame and M is the size of the frame. Signum function sgn is defined as

$$sgn(x(n)) = \begin{cases} 1, & x(n) \geq 0 \\ -1, & x(n) < 0 \end{cases} \tag{4.7}$$

Short time signal energy (EN)

Signal energy envelope of the syllable is calculated as given in equation (3.1) in the previous chapter. Maximum energy of the energy trajectory for the syllable is normalized to $0dB$. Without normalization energy depends on the recording gain and other recording conditions and would not assign much information on the energy content of the syllable. Normalized energy is able to discriminate syllables with different within-syllable energy content.

Modulation spectrum (MSm, MSf)

Modulation spectrum is not purely time domain measure, but it fits here, because signal envelope is calculated in the time domain. This feature is used to detect modulation frequency and index of amplitude modulation of the syllables. First envelope of the syllable is calculated from the analytic signal, which is formed using Hilbert transformation (Hartmann 1997). Envelope of the syllable is given as

$$E_{\mathcal{H}}(n) = |x(n) + i\mathcal{H}(x(n))| \tag{4.8}$$

where $x(n) + i\mathcal{H}(x(n))$ is the analytic signal and $\mathcal{H}$ denotes the Hilbert transform.

Modulation spectrum is Fourier transformed envelope signal. In this work position and normalized magnitude of the maximum peak of the modulation spectrum are used as features of a syllable. Position of maximum frequency is related to the modulation frequency and magnitude to the modulation index. Figure 4.1 shows the envelope of a syllable of Hooded Crow (*Corvus corone cornix*) calculated by equation (4.8), and modulation spectrum of that syllable.

4.2 Feature evaluation and selection

As mentioned earlier the number of features should be chosen as small as possible in order to make the classifier more efficient. Features usually provide different discriminative power, but usually goodness of a feature is not known *a priori*. Also diversity within bird

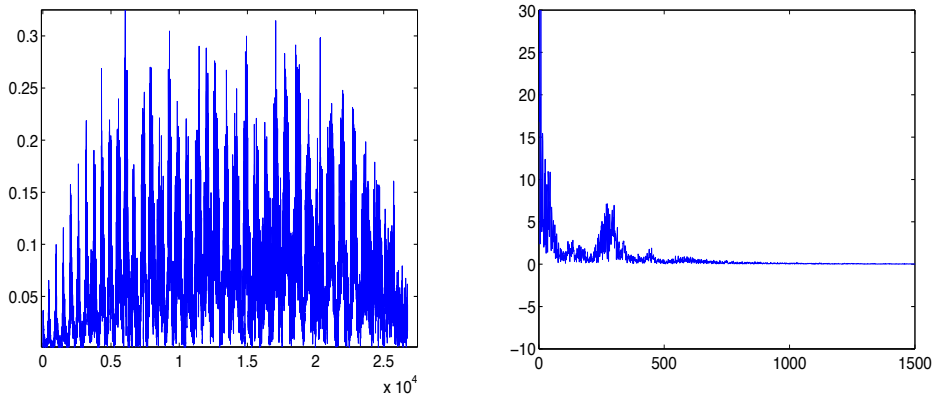

Figure 4.1: Envelope and modulation spectrum of the syllable of Hooded Crow.

sounds is large and it is likely that certain features that provide good discriminative power for one type of sounds may not be good for other types (McIlraith & Card 1997). Therefore it is justified to use species specific features

Discriminative power of individual features are evaluated using Linear Discriminant Analysis (LDA) method (Fukunaga 1990). In this work LDA is used for finding class separability measure for individual features. Advantage in LDA method compared to many other statistical methods is that information on distribution of features is not needed. Typically this distribution information in not known and Gaussian assumption is not suitable in most of the cases. Discriminative power is measured based on information how features are scattered in the space and how individual features are correlated. Discriminative power is calculated for a set of species but also for individual species. When evaluating species specific discriminative power of features species of intrest is set to one class and all other species are grouped together, thus it reduces to two class problem. The method is the same for both of the cases.

First we define scatter matrixes. *Within-class scatter matrix* is defined as

$$S_w = \sum_{i=1}^{M} P_i \Sigma_i \tag{4.9}$$

where Σ_i is the covariance matrix and P_i is *a priori* probability of the class or species i. *A priori* probability is defined here as $P_i = n_i/N$, where n_i is the number of syllables in the class i, N is total number of syllables among all classes and M is number of classes. *Between-class scatter matrix* is defined as

$$S_b = \sum_{i=1}^{M} P_i(\mu_i - \mu_0)(\mu_i - \mu_0)^T \tag{4.10}$$

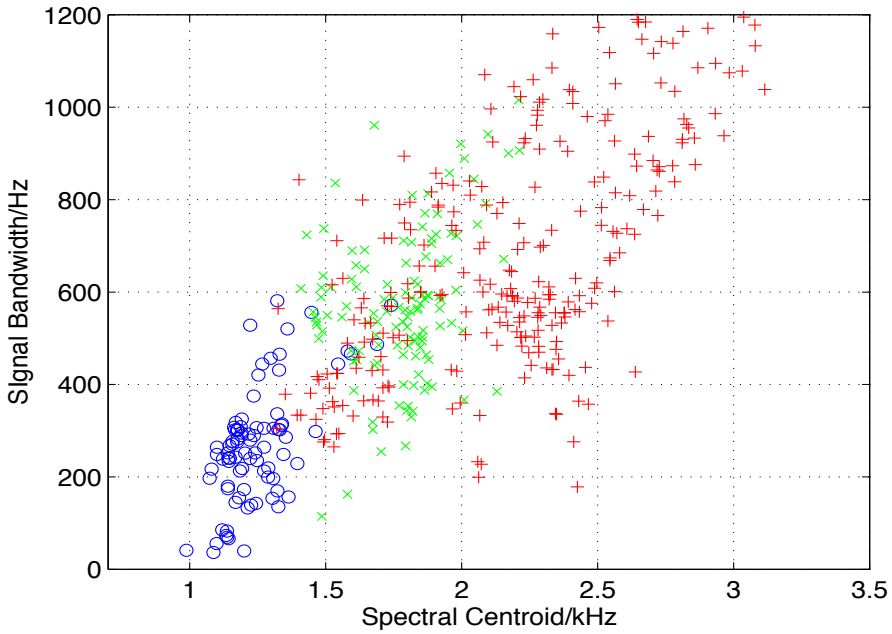

Figure 4.2: Distribution of features Spectral centroid and Signal bandwidth for species Common Raven(o), Hooded Crow(x) and Mapgie(+). Both axes represents frequency in Hz.

where μ_i is the mean vector of the feature vectors in the class i and μ_0 is mean of all feature vectors in all classes.

A measure for discriminative power of individual features or sets of featured is defined as

$$J = \frac{\det(S_b)}{\det(S_w)} \tag{4.11}$$

which gives large values when different classes are well separated.

In figure 4.2 is presented the distribution of features Spectral Centroid (SC) and Signal Bandwith (BW) for syllables of species Common Raven (*Corvus corax*), Hooded Crow (*Corvus corone cornix*) and Mapgie (*Pica pica*). Discriminative power in (4.11) gives $J = 1.85$ for SC and $J = 1.27$ for BW when they are calculated in general case, thus SC provide better overall discriminative power. However species specific discriminative power for SC gives $J = 1.50$ for species Common Raven, but only $J = 1.04$ for Hooded Crow. Indeed it can be seen in figure 4.2 that samples of Common Raven are clustered separate whereas samples of Hooded Crow overlap with samples of Mapgie.

Feature evaluation is followed by feature selection, where a subset of features is selected

from all available features. There is a number of ways to select the features, a good rewiev of different methods can be found for example in (Dash & Liu 1997). Complete feature selection method is in practise typically computationally too heavy, because it tests all possible feature combinations. In this work with $N = 19$ features this would be $2^N = 524288$ combinations in the general case. The simplest way is *scalar feature selection*, where features are treated individually (Theodoridis & Koutroumbas 1998). In this method a subset of features is selected based on the class separability measure. Advantage of this method is its computational simplicity. Major drawback is that the method does not take into account mutual correlations between features. Feature selection in this work is descibed in more detail in the next chapter.

4.3 Sinusoidal model of syllables

Syllables in the bird sounds can often be modelled by a single or small number of time-varying sinusoidal components, which encourage to use a sinusoidal model as a representation of a syllables in these cases. Multiple sinusoidal components of bird sounds are often in a harmonic ratio. Sinusoidal model of syllables have been demonstrated in context of AveSound project (Fagerlund 2004) previously in (Härmä 2003) and in (Härmä & Somervuo 2004). Currently in this work bird sounds are divided into four classes based on their harmonic structure: one for sounds with single sinusoidal component (class I) and three for sounds with harmonic structure. Sounds with harmonic structure are divided into three classes based on which component is the strongest. In class II sounds fundamental frequency is the strongest and in classes III and IV respectively first and second harmonics are strongest components.

In classical sinusoidal modelling method sound is represented with multiple time varying sinusoidal waves (McAulay & Quatieri 1986). Short Time Fourier Transform (STFT) is applied to detect sinusoid components of the sound. Sinusoidal components are assumed to be stationary within the frame and amplitude and frequency values are constant. Underlying sinusoidal components in the signal frames are detected from peaks of spectrum of the frame. From frame to another the frequency of a sinusoidal component is assumed to evolve slowly. When close enough neighbouring components cannot be found in adjacent frames sinusoidal components can also start or end within the sound to be modelled.

In this work simplified version of sinusoidal model is used. Each sinusoidal component (fundamental component and its harmonics in suitable cases) of syllable are modelled with one time varying sinusoid. Analysis-By-Synthesis/Overlap-Add (ABS/OLA) (George & Smith 1997) is used as sinusoidal modelling method. In this method residual is always degreased when new sinusoidal components are added to the model.

4.3.1 One sinusoid model

In Overlap-Add method sampled syllable signal $x(n)$ is first divided into finite length overlaping subsequences or frames $x_m[n]$ with equal length of N_a:

$$x(n) = \sum_{m=0}^{\infty} x_m[n - mN_s] \tag{4.12}$$

where N_s defines separation between adjacent frames and

$$x_m = \begin{cases} x(n + mN_a), & 0 \leq n \leq N_a \\ 0, & otherwise \end{cases} \tag{4.13}$$

Adjacent sequences overlap $N_a - N_s$ samples. In this work frame length of 256 samples is used. Adjacent frames overlap 50% (stepsize of 128 samples), thus here is assumed syllable to be stationary over 128 samples ($3ms$). Sinusoidal model of syllable can now be written for one sinusoid model as

$$s(n) = \sigma(n) \sum_{m=0}^{\infty} w(n - mN_s) s_m(n - mN_s) \tag{4.14}$$

where $w(n)$ is windowing sequence and $s_m(n)$ is constant-amplitude and constant-frequency sinusoid model of the k:th frame. The sequence $\sigma(n)$ is average of the magnitude of $x(n)$. Purpose of envelope sequence $\sigma(n)$ is to model global amplitude modulation of the syllable and reduce its effect to parameter estimation. In this work $\sigma(n) = 1$ in used because in that way the envelope information is included into the parameters of the sinusoidal model. The sinusoidal model $s_m(n)$ of the frame m is given as

$$s_m(n) = a_m \cos(\omega_m n + \phi_m) \tag{4.15}$$

where a_m, $\omega_m = 2\pi f_m / F_s$ and ϕ_m are respectively constant amplitude, frequency and phase values of the model s_m of the frame. Here F_s is sampling frequency and f_m is frequency in Hertz. Objective is to find model parameters σ, a, ω and ϕ so that modelling error is minimized. The Mean-Square Error (MSE) is used to evaluate performance of the model. It is calculated for the model $s(n)$ of the syllable $x(n)$ as

$$MSE = \sum_{n=0}^{\infty} [x(n) - s(n)]^2 \tag{4.16}$$

where $x(n) - s(n)$ is residual signal of the model.

Parameter estimation of the sinusoid is started by estimating frequency trajectory. It is started from the frame with highest energy and proceeding forward and backward until end and start of the syllable. Each frame is transformed to the frequency domain with 1024

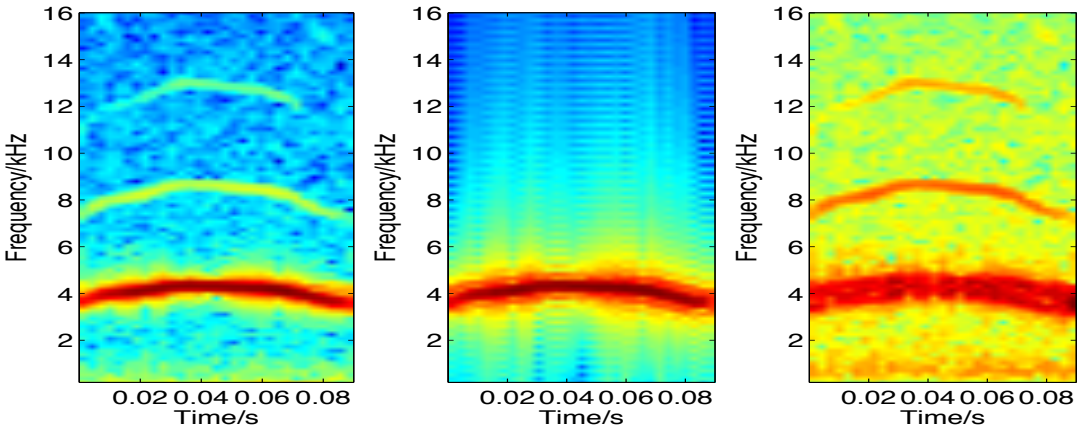

Figure 4.3: One sinusoid model of syllable of Willow Warbler (*Phylloscopus trochilus*). In the left panel is original syllable. In the middle is spectrogram of the model and in the right panel is the residual signal. The y-axis is frequency in kHz and x-axis is time in seconds.

point FFT. Frequency value of the frame is where transformed frame attain its maximum value. To avoid abrupt frequency jumps frequency value of the new frame is not allowed to overrun 5% deviation in relation to the frequency value of the previous frame.

Magnitude and phase parameters are estimated frame by frame so that MSE of the frame is minimized. Expressing sinusoidal model in (4.15) as

$$\begin{aligned} a_m \cos(\omega_m n + \phi_m) &= a_m \cos(\omega_m n)\cos(\phi_m) - a_m \sin(\omega_m n)\sin(\phi_m) \\ &= a\cos(\omega_m n) + b\sin(\omega_m n) \end{aligned} \tag{4.17}$$

leads to system of equations given as

$$\begin{aligned} a\gamma_{11} + b\gamma_{12} &= \psi_1 \\ a\gamma_{12} + b\gamma_{22} &= \psi_2 \end{aligned} \tag{4.18}$$

Solving a and b leads to closed form expression of a_m and ϕ_m. For details of solving a_m and ϕ_m see (George & Smith 1997).

One sinusoid model of a syllable from Willow Warbler (*Phylloscopus trochilus*) is shown in figure 4.3 Model with one sinusoidal component is feasible in this case. MSE error of the syllable is only $-7.6dB$ whereas energy of the original syllable is $36.2dB$.

4.3.2 Model of harmonic series of sinusoids

As mentioned in (Härmä & Somervuo 2004) syllables with harmonic structure are common. Here each harmonic component of a syllable is modelled with one time varying sinusoid.

Procedure for modelling harmonic structure of the syllable starts as in one sinusoidal model. First dominant sinusoidal component is modelled as described above. In relation to dominant components frequency curve frequency trajectory of harmonic component is selected from neigbourhood of 4 frequency bins ($180Hz$) of $k\omega_m$, where k specify the harmonic component and ω_m is frequency value of dominant component. Magnitude and phase of harmonic component is estimated as for dominant component. Model in (4.15) can now be written for multible sinusoids as

$$s_m(n) = \sum_k a_m^k \cos(\omega_m^k n + \phi_m^k) \tag{4.19}$$

where k specifies harmonics. For class II values $k = 1, 2, 3$ is used. Class III is specified by values $k = \frac{1}{2}, 1, \frac{3}{2}, ...$ and class IV with values $k = \frac{1}{3}, \frac{2}{3}, 1, \frac{4}{3}, ...$ MSE and residual of the model is calculated as given in (4.16).

4.3.3 Class measure

In ABS/OLA system residual and MSE of the model is reduced always when new sinusoids is added to the model. Therefore MSE of the model as such cannot discriminate syllables into harmonic classes. Modelling gain of the class C is defined in (Härmä & Somervuo 2004) as

$$G_C = 20 \log 10 \left(\frac{E(x(n)^2)}{MSE_C} \right) \tag{4.20}$$

where $E(x(n)^2)$ is energy of original syllable and MSE_C is MSE of the model $C = \{I, ..., IV\}$. Range measure is defined as $R = G_A - G_I$, where G_A is modelling gain of model that include all sinusoidal components modelled in classes I,...,IV. Finally likelihood that specify syllable to certain class is defined for classes II, III and IV as

$$H_C = \frac{G_C - G_I}{R} \tag{4.21}$$

For class I a heuristic likelihood measure is given as

$$H_I = \frac{1}{(1 + e^{(0.6R-3)})(1 + e^{(-0.2G_I-10)})} \tag{4.22}$$

Syllable is then classified to the class with highest likelihood. For example for syllable presented in figure 4.3 to belong to the class I is 95 whereas it is 88, 2 and 10 respectively for classes II, III and IV. Most of the syllables are cassified to the class I. Second largest is fourth class, but also syllables with inharmonic structure, like syllable from song of Blyth's Reed Warbler (*Acrocephalus dumetorum*) presented in figure 4.4, tends to fall into this

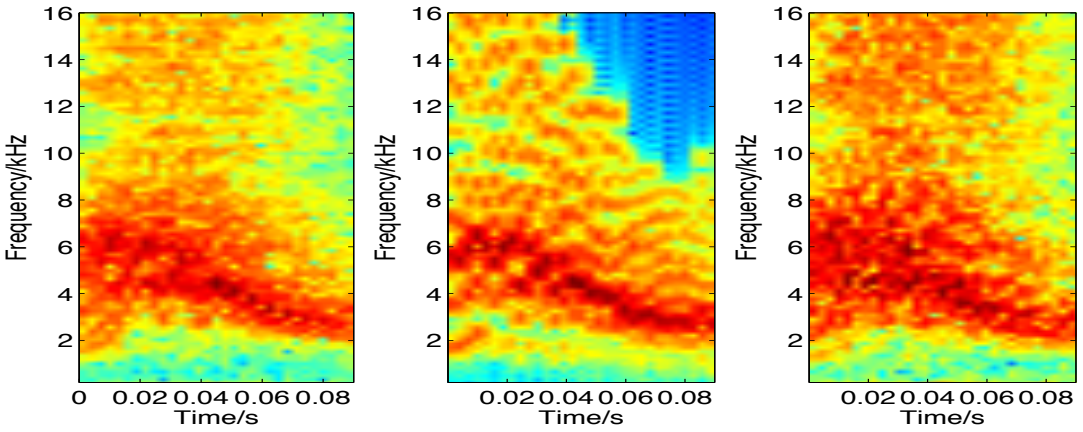

Figure 4.4: Model of syllable of Blyth's Reed Warbler (*Acrocephalus dumetorum*). Panels as in figure 4.3.

class. Likelihood for this syllable for class IV is 75. Model is however inaccurate, MSE of the model is $29dB$ whereas energy of the syllable is $38dB$.

Chapter 5

Classification and recognition results

In classification task feature vector space is divided into regions that correspond to different classes. If we have *a priori* information of the number of the classes in the classification problem, then it is called *supervised pattern recognition*. In this work the number of species to be recognized is known. However it is not known exactly, at least not for all species, the number of different syllables species can produce. In many classification schemes however models for each different syllable types would improve performance of classifier. Types of syllables can be detected by using *unsupervised pattern recognition* or *clustering* methods and algorithms. In this goal is to group similar patterns together and distinguish groups or clusters.

The role of a classifier is to decide which is the best possible class for the test pattern. This is done by comparing similarity between test pattern and model or target patterns of classes. Classifier does the decision based on the similarity or distance measure between test pattern and model patterns. Suitable distance measure depends on the problem and selected classification scheme. Simplest distance measure is minimum length measure in which Euclidean distance between feature vectors of test pattern and model patterns of classes is calculated.

There is a number of different classification methods. Supervised and unsupervised classification methods were mentioned above. This discrimination relates to the availability of the class information of the data. In supervised method class information of the data is known and models of classes are constituted patterns of classes. Simple example of supervised classification is *Nearest Neighbour (NN)* rule. In NN-rule test pattern is assigned to the class, where minimum distance between feature vectors of pattern and model of the class is achieved. In unsupervised classification (clustering) data is not labeled and aim is to find similar groups in the data. Clustering is sometimes used along with supervised classification for searching prototypes of classes.

Once the classification method has been selected its performance to pattern recognition task is evaluated using available data. First data is divided into training and testing data sets. There is few basic methods and a number of variations and combinations of these methods to select training and data sets (Theodoridis & Koutroumbas 1998). Suitable method depends also on the selected classification method. Three basic methods, shortly described in the following, are *resubtitution*, *holdout* and *leave-one/k-out*.

In resubtitution method same data set, typically all data available, is used first for training of the classifier and then for testing. It is clear that this method provide optimistic estimate of the true error probability because it is biased even for statistically independent samples. Advantage of the method is that training of the classification system must be done only once for all the test patterns. To obtain sufficiently good estimate of the classification error the number of samples N must be large enough, but also the ratio N/l, where l is the dimension of the feature vector, must be large enough and also variance of the features of the data must be low enough. In (Kanal 1974) was found out for N/l-ratio as high as five the estimate of classification error can be still optimistically biased. In this work N/l-ratio have ranged from about five to 20. Notable drawback of the system is also that by adding the features it is possible to reduce the classification error for the design set of patterns without any improvement in classification ability of independent patterns. Also this method is not possible with NN-classifier or its variants because patterns in the training data set are models of the classes.

Leave one/k out method reduce or cancel correlation of training and testing data sets in resubtitution method. In leave-one-out method apart from test pattern all the data is used for training the classifier. This is done for all patterns in the data set. A major advantage in this method is that basically all the data can be used for training still maintaining the independence between training and testing data sets. This method is useful if available data is reduced. Drawback of the method is its high computational complexity. Classifier must be trained separately for all test patterns (training data set is different for each test pattern). For many sophisticated classification schemes this method is not possible because training of the classifier is itself computationally very demanding. Leave-k-out is a sophisticated version of the leave-one-out method, whose purpose is to reduce computational complexity of the system and unwanted correlations between test patterns and training data set. At each test round not only the test pattern is left out from training part but k samples. The choice of which patterns are left out depends on the application. For example if a subset of patterns are correlated they can be left out from training part in order to maintain independence between training and testing data sets.

In holdout method data is divided into two independent non-overlapping data sets, one for training and one for testing the system. Like leave one/k out method holdout method

provide also unbiased estimate of error probability for independent samples. Advantage in relation to leave one/k out method is that classifier is trained only once for all test patterns. In this method the size of both training and testing data sets are reduced and number of required patterns for sufficient classification simulation is higher than in leave one/k out method. Another problem related to division into two data sets is to decide dimensions of data sets and which patterns are selected to the training set and which to the test set. Furthermore, both data sets should be independent and yet provide as good representation of the classes as possible.

In this work resubtitution method is not suitable because in many cases the N/l-ratio is relatively low. Also syllables withing classes can be highly correlated because some of them are from same individual bird. Leave one/k out method is better, because available data is almost maximally utilized and unwanted correlations are canceled. Also the choice of dimensions of the training and testing data sets is not needed as in holdout method. Leave one/k out method however sets a limitation to the acceptable classifier, because the classifier must be trained for each test pattern or set of patterns.

5.1 Methods for classification

In this work k-Nearest-Neighbour (kNN) nonlinear classifier is used. Compared to other classification methods kNN is convenient as preliminary method because it is simple to implement and it is very flexible. Major drawback is that method is computationally heavy, because the test feature vector is compared with all vectors in the training data set. For data selection leave-k-out method is used in order to obtain as much data for the train samples. This method fits well with kNN method, because very little preprocessing is needed during the training part of the kNN method. This also makes method efficient for simulating classification with different feature sets and also with different sets of species.

KNN method has other advantages that were mentioned above. In classification methods where a model of the class in constituted from training data, clustering is needed if class include several clusters. In kNN method however clustering is not needed, because nearest neigbours of the test vector is always in the right cluster. Here classes typically constitute clusters because individual species can produce different sounds and they describes into different areas in the feature space. KNN method does not require as much training data as other classification methods. Required amount of training data depends on how broad feature vectors are distributed. KNN method is also rather tolerant to outliers.

KNN method can bias towards classes with more samples in the training data set. This in expressed especially near decision boundaries if classes have different distribution densities. Method favors in this case a class with a higher density.

5.1.1 K-Nearest-Neighbour classifier

Values of the features used in this work lies within different dynamic ranges. Features with high value would have larger influence to the classification that features with small value. To obtain equal significance for the features they are normalized to the similar ranges. Here features are normalized to have zero mean and unit variance. For all features $k = 1, 2, ..., l$ new value of the feature k of the syllable i is given as

$$\widehat{x}_{ki} = \frac{x_{ki} - \overline{x}_k}{\sigma_k} \tag{5.1}$$

where $\overline{x}_k$ and σ_k^2 are mean and variance of the feature k given as

$$\overline{x}_k = \frac{1}{N} \sum_{i=1}^{N} x_{ki} \tag{5.2}$$

$$\sigma_k^2 = \frac{1}{N-1} \sum_{i=1}^{N} (x_{ki} - \overline{x}_k)^2 \tag{5.3}$$

where N is total number of syllables. The method above is linear and it assumes data to be evenly distributed around the mean. This is not always the case, for example outliers cause deviation to the distribution of the features. Here the number of outliers is assumed to be small so that they have only a small influence to the normalization.

Data is divided into training and testing data sets using leave k out method. Syllables or a set of syllables from same song and recording (see section 3.1 Bird sound database) are often from same species and it is likely that they are highly correlated. If these syllables would be in the training and test data sets would this give optimistic classification error in the similar way to the resubtitution method. This phenomenon is even higher for kNN classifier than for other types of classifiers, because feature vectors of syllables are compared directly. Here when certain syllable is classified, syllables from same recording are removed from training data set. In other words all syllables from certain recording can be classified with same training data set, which also reduces computational load.

Classification is done based on the distance measure between the test syllable feature vector and the model feature vectors. The model consists all of feature vectors in the training set. Two distance measures used in this work are Euclidean distance and Mahalanobis distance. Euclidean distance between vectors x and y is defined as

$$d_E(x, y) = \|x - y\| = \sqrt{(x - y)^T (x - y)} \tag{5.4}$$

and Mahaladonobis distance is defined as

$$d_M(x,y) = \sqrt{(x-y)^T \Sigma^{-1} (x-y)} \tag{5.5}$$

where Σ is the covariance matrix of the training data. Covariance matrix can be calculated for the whole training data or for each class separately. If covariance matrix is different for each class then separate matrices for each class should be used. Advantages using Mahalanobis distance metric is that it decorrelates different features. Also normalization of the data is not needed, because covariance matrix of the training data automatically scales the coordinate axes.

Distances between test syllable feature vector and feature vectors in the training data set are set in ascending order so that minimum distance is in the first place. The nearest neighbour of a test syllable is the syllable in the training data set, whose feature vector gives the minimum distance compared to the test feature vector Test syllable is assigned to the class, which is most often represented in k-nearest neighbour.

5.2 Simulation results

In this thesis main focus is in bird sounds that are inharmonic and do not fit to the sinusoidal model. Mainly this type of sounds are used in the following recognition experiments, but the recognizer is also tested with *Phylloscopus* family of birds in order to compare recognition results with low level acoustical features and sinusoidal model. Variables that were tested in recognition experinments of inharmonic sounds were the used distance measure and the number of neighbours in the classifier. Recognition was also tested with different number of features and with different feature sets.

5.2.1 Class of inharmonic sounds in birds

In (Härmä & Somervuo 2004) it was found out that a model with a small number of time varying sinusoidal components is sufficient representation for a large number of bird sounds. However many birds produces complex sounds that cannot be modeled with sinusoidal model. For example majority of the sounds produced by Hooded Crow (*Corvus corone cornix*) and species related to it does not fit to the sinusoidal model.

The harmonicity classes of bird sounds were described in section 4.3.3. It turned out that inharmonic sounds were often labeled to the harmonic class IV, but meanwhile likelihood to belong to the other harmonic classes were very small. In this thesis the definition of the class of inharmonic bird sounds is based on this observation. If likelihood to belong to harmonic classes I, II and III were less than 60%, sounds was labeled to the class of inharmonic sounds. In other cases it is assumed that sinusoidal model is good enough

Common name	Latin name	Lat. Abbr.	Recs.	Non-voiced	Syllables
Common Raven	Corvus corax	CORRAX	7	96%	91
Hooded Crow	Corvus corone cornix	CORNIX	8	98%	160
Magpie	Pica pica	PICPIC	7	99%	312
Eurasian Jay	Garrulus glandarius	GARGLA	9	99%	99
Sedge Warbler	Acrocephalus schoenobaenus	ACRSCH	6	65%	331
Marsh Warbler	Acrocephalus palustris	ACRRIS	8	34%	277

Table 5.1: A set of birds that produce regularly non-voiced sounds. Column are common English name, Latin name, abbreviation, number of recordings and percentage and number of non-voiced syllables extracted from these recordings.

representation of the sound.

In table 5.1 is listed a set of species that produce regularly inharmonic sounds. First three columns in the table give respectively common English name, Latin name and commonly used abbreviation derived from the Latin name. Next columns give number of recordings in the database, percentage and number of inharmonic sounds in these recordings. Mainly this set of species is tested in the recognition experiments in this work.

5.2.2 Classifi cation power of features

Classification power of individual features was tested with the method described earlier in this work in section 4.2. In table 5.2 is presented the classification power of the individual features for species presented in table 5.1. High values in the table refer to a better classification power. First column gives the abbreviation of the feature and the following columns express the classification power for all species together and species specific values. Table shows that mean of the Spectral Centroid (mSC) provides the best overall classification power. However for species Eurasian Jay (*Garrulus glandarius*, GARGLA) mSC does not provide good classification power.

Features for recognition experinments were selected by means of classification power of individual features using scalar feature selection method. In this method features are treated individually and possible correlations between features are not taken into account. Features are selected in on/off fashion, i.e. each feature is used or not used in classification, thus all features had equal weight. In this work the recognition was tested using different feature sets with the same features for all species and also with the species specific feature sets.

5.2.3 Recognition results

Recognition results for the simplest configuration of the recognition system for the same species is presented in the table 5.3 a). In this configuration all features were used and they

features	all species	CORRAX	CORNIX	PICPIC	GARGLA	ACRSCH	ACRRIS
mSC	4.1323	1.1816	1.1600	1.1455	1.0073	1.8811	1.0626
mBW	1.2666	1.1581	1.0038	1.0112	1.0136	1.0377	1.0445
mSRF	2.4188	1.2372	1.0924	1.0546	1.0007	1.6380	1.0097
mSF	1.2159	1.0136	1.0522	1.0103	1.0305	1.1095	1.0184
mSFM	1.2135	1.0966	1.0003	1.0076	1.0240	1.0453	1.0468
mZCR	3.7701	1.1924	1.1473	1.1345	1.0056	1.8414	1.0546
mEN	1.0537	1.0256	1.0038	1.0126	1.0005	1.0074	1.0119
vSC	1.0162	1.0022	1.0009	1.0056	1.0009	1.0044	1.0064
vBW	1.0068	1.0002	1.0008	1.0011	1.0013	1.0004	1.0046
vSRF	1.0313	1.0009	1.0276	1.0003	1.0027	1.0045	1.0001
vSF	1.0375	1.0002	1.0070	1.0084	1.0081	1.0145	1.0072
vSFM	1.0184	1.0008	1.0003	1.0069	1.0025	1.0004	1.0122
vZCR	1.0146	1.0029	1.0001	1.0053	1.0010	1.0036	1.0053
vEN	1.0156	1.0002	1.0000	1.0056	1.0024	1.0002	1.0111
T	2.0669	1.0011	1.4824	1.0480	1.1377	1.0457	1.0727
MSm	2.0646	1.8506	1.0227	1.0143	1.0243	1.0382	1.0136
MSf	3.0218	1.0802	1.0729	1.0789	1.0700	2.3537	1.0005
range1	2.5933	1.0583	1.1061	1.1561	1.0432	1.6634	1.0607
range2	1.5032	1.2477	1.0193	1.0008	1.0002	1.2187	1.0016

Table 5.2: Discriminative power of individual features. First column gives disciminative power of individual features for all species together. Latter columns gives species specific discriminative power of features. Features are identified by their abbreviation. Lower case m and v is related to the mean and variance of the feature calculated on the frame basis.

were compared with Euclidean distance measure. The Nearest Neighbour (1-NN) classifier were used for recognition. The average recognition result with this configuration was 49%, but the differencies in the recognition rates between species was large. Mahalanobis distance measure improved recognition result significantly. Recognition results using Mahalanobis distance measure are presented in table 5.3 b). The average recognition result was 71%. The best improvement in recognition rate was obtained with Marsh Warbler (*Acrocephalus palustris*), which was correctly recognized in 53% of the cases using Euclidean distance measure and in 82% of the cases with Mahalanobis distance measure.

The number of the neighbours in the k-NN classifier had only a small effect to the average recognition rate. The recognition results as the function of the number of the neighbours are ilustrated in the figure 5.1. Mahalanobis distance measure was used in this experiment. The average recognition rate was slightly reduced when the number of the neighbours was increased. The change in the average recognition rate was only five percentage units between 1-NN and 25-NN classifiers. The best average recognition rate was obtained with the 1-NN classifier. The change in the recognition rate of the individual species was higher than the change in average recognition rate. The biggest change in recognition rate was with species Marsh Warbler (*Acrocephalus palustris*), whose recognition rate decreased from 82% in 1-

a)

rec result	CORRAX	CORNIX	PICPIC	GARGLA	ACRSCH	ACRRIS
CORRAX	69	14	4	3	0	0
CORNIX	19	36	24	7	0	3
PICPIC	10	36	41	41	7	12
GARGLA	2	7	16	36	5	3
ACRSCH	0	1	6	4	56	29
ACRRIS	0	5	10	8	32	53

b)

rec result	CORRAX	CORNIX	PICPIC	GARGLA	ACRSCH	ACRRIS
CORRAX	74	5	5	0	0	0
CORNIX	10	56	12	21	2	1
PICPIC	14	28	67	5	4	5
GARGLA	0	9	7	73	0	1
ACRSCH	0	1	2	0	73	10
ACRRIS	2	2	6	1	23	82

Table 5.3: Recognition results for species described in the table 5.1 using 1-NN classifier with all features and a) Euclidean distance measure and b) Mahalanobis distance measure. Columns tells the percentage of the syllables of the species on the top row being recognized as a syllables of the species on the leftmost column.

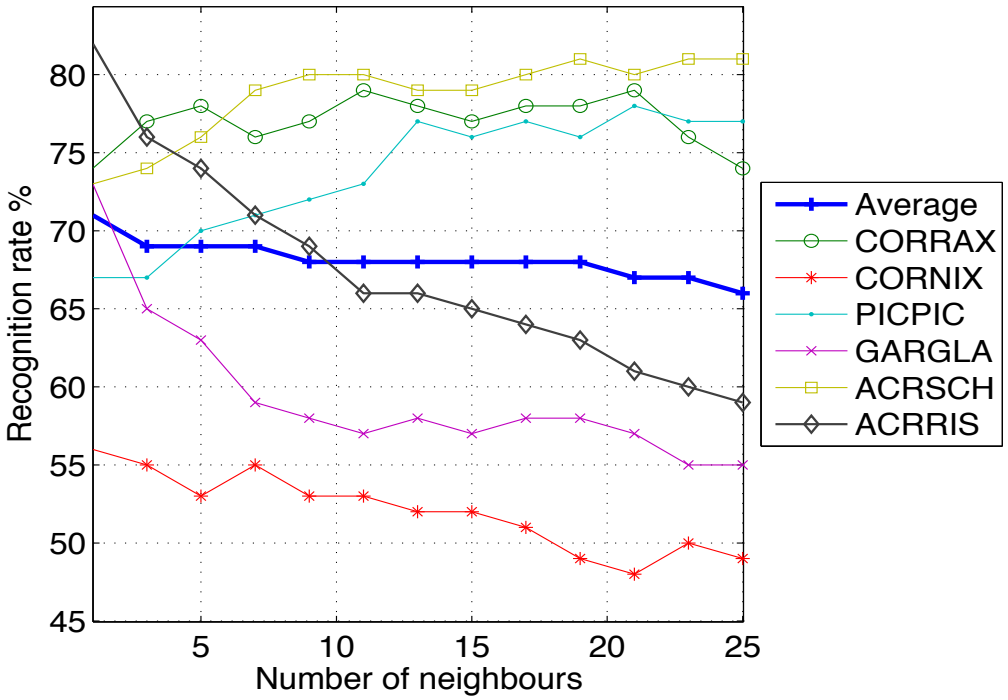

Figure 5.1: Recognition results as function of number of neighbours in the classifier.

NN classifier to 59% in 25-NN classifier. Result cannot be fully expounded on the different number of samples in the training feature set because recognition rate of Marsh Warbler (*Acrocephalus palustris*) was heavily reduced and it has third most of the samples. More probable explanation is different distribution dencities between species.

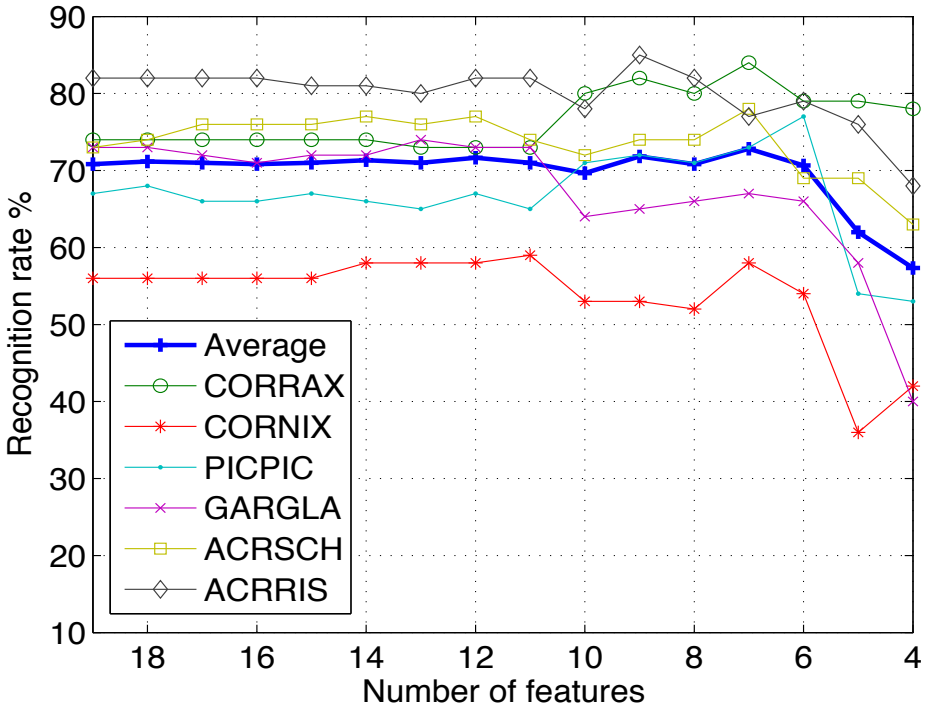

Figure 5.2: Recognition results as function of number of features. Features were selected by means of classification power of the features in the general case, i.e. first column in the table 5.2.

In table 5.2 is presented classification power of the individual features. Recognition was tested with different numbers of features. N features with the best classification powers were selected for the recognition experinments executed with the N dimensional feature space. In figure 5.2 is presented recognition results as function of the number of the features. Here features were selected based on the classification power in the general case and same features were used in context of all species. The best average recognition result (73%) were obtained with seven features, but the result was not much improved from the case where all features were used. Variance features had the lowest classification power and their effect to the recognition result in context of all species was negligible.

Recognition was also tested with different numbers of features using features specified by species specific classification power, i.e. features were removed by means of the classification power of the features related to each species (columns two to seven in the table 5.2). The equal number of features were used in context of each species in all recognition experinments. In figure 5.3 is presented recognition results as function of number of features used for classification. Also in this case removal of the worst features had only a small

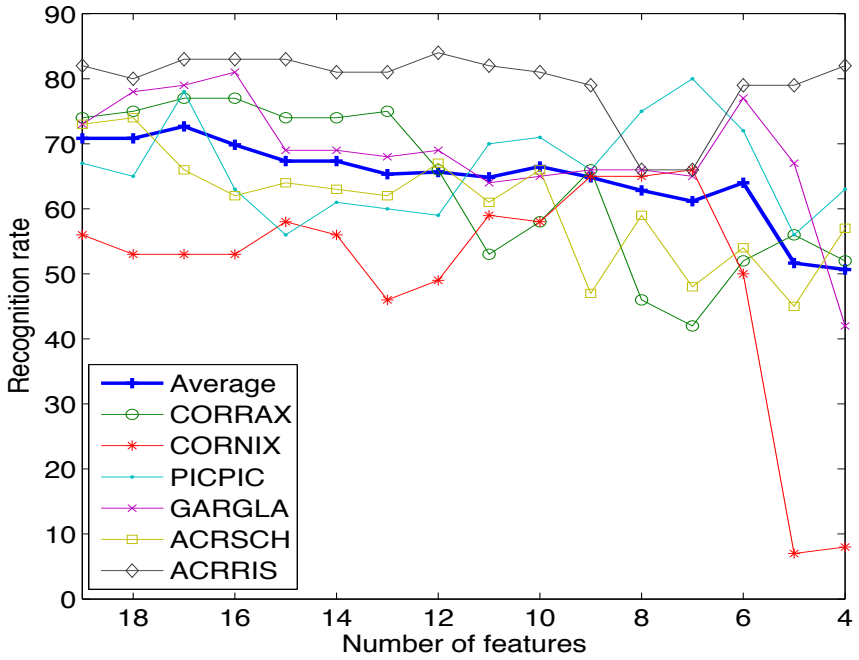

Figure 5.3: Recognition results as function of number of features. Features were selected by means of species specific classification power of the features, i.e. column two to seven in the table 5.2

effect to the average recognition rate. However, the change in recognition rate of individual species changed more than in the case where the same feature set was used in context of particular species. This may be due to some correlation between the features, which did not turned out in previous case where same features were removed in context of all species.

5.2.4 Recognition results for Phylloscopus family

Recognition was also tested with with *Phylloscopus* family of birds. Recognition of this family were previously studied in (Härmä 2003) and in (Härmä 2004) using the sinusoidal model of syllables. Recognition results in (Härmä 2004) are presented in the table 5.4 a) and recognition results using features described in section 4.1 of this work are presented in the table 5.4 b). Results show that the average recognition rate is slightly better in the latter case (from 65% to 68%). However recognition rate of some species was also reduced. Sounds of *Phylloscopus* family are often labelled to the harmonic class I, but they give often high likelihood to the class IV. It have been previously noted that sounds with complex spectrum fall into class IV. Visual inspection of spectrograms of syllables shows that a large number of

a)

rec result	PHYBOR	PHYCOL	PHYDES	PHYLUS	PHYSIB
PHYBOR	70	8	7	2	6
PHYCOL	8	60	6	22	10
PHYDES	17	9	59	11	3
PHYLUS	4	21	11	58	1
PHYSIB	1	2	17	7	80

b)

rec result	PHYBOR	PHYCOL	PHYDES	PHYLUS	PHYSIB
PHYBOR	83	2	7	6	6
PHYCOL	2	65	9	20	3
PHYDES	9	17	49	20	5
PHYLUS	4	19	7	66	4
PHYSIB	3	3	7	9	77

Table 5.4: Recognition results for Phylloscopus family. Columns tells the percentage of the syllables of the species on the top row being recognized as a syllables of the species on the leftmost column.

the syllables of *Phylloscopus* family are tonal but also frequency modulated. Also frequency context of these syllables is relatively stationary.

Chapter 6

Conclusions and Future Work

6.1 Conclusion

Automatic recognition of bird species by their sounds have been studied within the Avesound project. In this work focus has been in species that produce regularly sounds that are not tonal or harmonic in structure. The long term objective in this research is to develop methodology for a system that is capable to recognize majority of common Finnish bird species in field conditions.

The sounds of birds are produced mainly by the unique organ called syrinx. Diversity within structure of syrinx of different species is large, which evoke large number of different sounds birds can produce. Bird sounds can be divided by function into songs and calls. Songs are more spontaneous than calls and mostly produced by males during the breeding season. Call sounds are produced by both sexes throughout the year and they occur in some particular context with certain function.

Bird sound database have been developed within Avesound project. The current XML-based database includes nearly 2400 recordings from almost 300 bird species. The recordings are from many different sources, like from commercially available CD's and from recording experts of bird sounds. Typically recording and environmental conditions differ from one recording to another and information on these conditions is not available. The system needs to be invariant for these conditions. This holds also for the new recordings because environmental conditions can change abruptly even if recording conditions would be the same.

Segmentation is crucial for the following steps of classification because in this phase concurrent segments (syllables) are extracted from raw recordings. In this work two different segmentation methods have been tested. In these methods segmentation were done by means of short-time signal energy and short-time maximum of the spectrum. Performance

of these methods were quite similar. In both methods syllabic threshold were set by means of an iterative noise level estimate. Merging and omitting syllable candidates were executed in order to reduce errors in segmentation phase.

In this work syllables were represented with 19 low-level acoustical features. Seven features measured short time behavior of the syllable. Mean and variance of the feature trajectories calculated over a syllable were used as actual features in classification. Remaining five features described static properties over the entire syllable. Discriminative powers of individual features were calculated using LDA method. It shows that features related to the frequency band of the sound, such as SC, SRF and frequency range, provides good classification power within inharmonic sounds. Also in some species modulation measures give good classification power. Variance values of the feature trajectories of syllables are less important than average values of those. This may be due to fact that many sounds used in this work are relatively stationary over the syllables. Low classification power in mEN and mSF supports this assumption. It seems that features with low classification power hold some information on syllables, because recognition results were slightly reduced when features with low classification were removed.

The kNN method was used for classification. Two distance measures, Euclidean and Mahalanobis, were tested in this work. Features were normalized when Euclidean distance measure was used. In Mahalanobis distance measure this is done automatically. Other benefits with Mahalanobis distance measure is that it decorrelates features and scales distributions of features. Mahalanobis distance measure improved average recognition accuracy with 1-NN classifier by 22 percentage units. The number of neighbours had only small effect to the average recognition result, but effect was higher in context of individual species. One reason might be the different numbers of syllables in different species. However, this does not explain why for example recognition rate of Marsh Warbler (*Acrocephalus palustris*, ACRRIS) was heavily reduced as the number of neighbours was increased. Other reason to this phenomenon might be in different feature distributions between species. In kNN method classes with dense feature distribution is favored in relation to classes with sparse distribution.

6.2 Future work

All results presented in this thesis are in early stage in a way to the system capable to recognize majority of bird species in Finland. More research is needed at the all stages of this system.

Current database has a relatively large number of recordings. However for the majority of species the number of recordings and individuals is not sufficiently large for reliable

species recognition experiments. The database is updated regularly and more recordings will be added to it. Because the recordings in the current database have been taken from many sources there are large differences in the documentation of recordings. Especially some commercially available recordings lack detailed documentation of the recordings. Another problem is that different sounds are not systematically annotated at any level. This information would be useful when studying structural models of series of syllables.

As mentioned earlier segmentation is a crucial part for subsequent steps of classification. Currently a little of spectral information is used in segmentation of syllables. Usage of spectral information would improve performance of segmentation especially in the context of syllables that overlap in the time domain and syllables that do not have clear pauses between them. Performance of segmentation phase could be improved by making more accurate noise level detection and thresholding for syllables. Also different threshold could be used for onset and offset detection of the syllable instead of the same value, which is currently in use.

The kNN classifier is suitable as a preliminary method, but it is not feasible for any real time applications or as the number of classes and samples increase because the method is computationally heavy. More sophisticated methods are, for example, different types of neural networks. Within other classifiers than kNN clustering might be compulsory because typically bird species have more than one type of syllables in their repertoire. Different types of syllables map to different positions in the feature space and different models are needed for each type of the syllable. Also different parametric representations of syllables might be required for different types of syllables. In this case recognition of species would be done hierarchically so that in the first phase the type of a syllable would be detected, which would be followed by species recognition with more detailed representation of the syllable.

Bibliography

Anderson, S. E., Dave, A. S. & Margoliash, D. (1996), 'Template-based automatic recognition of birdsong syllables from continuous recordings', *J. Acoust. Soc. Am.* **100**(2), 1209–1219.

Beckers, G. J. L., Suthers, R. A. & ten Cate, C. (2003), 'Mechanisms of frequency and amplitude modulation in ring dove song', *The Journal of Experimental Biology* **206**(11), 1833–1843.

Beddard, F. E. (1898), *The Structure and Classification of Birds*, Longmans, Green, London.

Brittan-Powell, E. F., Dooling, R. J., Larsen, O. N. & Heaton, J. T. (1997), 'Mechanism of vocal production in budgerigars (*melopsittacus undulatus*)', *J. Acoust. Soc. Am.* **101**(1), 578–589.

Casey, R. M. & Gaunt, A. S. (1985), 'Theoretical models of the avian syrinx', *J. theor. Biol.* **116**, 45–64.

Catchpole, C. K. & Slater, P. J. B. (1995), *Bird Song: Biological Themes and Variations*, 1 edn, Cambridge University Press, Cambridge, UK.

Dash, M. & Liu, H. (1997), 'Feature selection for classification', *Intelligent Data Analysis* **1**, 131–156.

Doya, K. & Sejnowski, T. J. (1995), A novel reinforcement model of birdsong vocalization learning, *in* G. Tesauro, D. Touretzky & T. Leen, eds, 'Advances in Neural Information Processing Systems', Vol. 7, The MIT Press, pp. 101–108.

Eronen, A., Tuomi, J., Klapuri, A., Fagerlund, S., Sorsa, T., Lorho, G. & Huopaniemi, J. (2003), Audio-based context awareness - acoustic modelling and perceptual evaluation, *in* 'IEEE Int. Conf. Acoust. Speech and Signal Processing'.

Fagerlund, S. (2004), 'Avesound - automatic recognition of bird species by their sounds', http://www.acoustics.hut.fi/~sfagerlu/project/avesound.html. Avesound project website.

Fee, M. S., Shraiman, B., Pesaran, B. & Mitra, P. P. (1998), 'The role of nonlinear dynamics of the syrinx in the vocalizations of a songbird', *J. Acoust. Soc. Am.* **95**, 67–71.

Fletcher, N. H. (1988), 'Bird song - a quantitative acoustic model', *J. theor. Biol* **135**, 455–481.

Fletcher, N. H. (1992), *Acoustics Systems in Biology*, Oxford U.P., New York.

Fletcher, N. H. (2000), 'A class of chaotic bird calls', *J. Acoust. Soc. Am.* **108**(2), 821–826.

Fletcher, N. H. & Tarnopolsky, A. (1999), 'Acoustics of the avian vocal tract', *J. Acoust. Soc. Am.* **105**(1), 35–49.

Fukunaga, K. (1990), *Introduction to Statistical Pattern Recognition*, Academic Press, San Diego, California.

Gardner, T., Gecchi, G. & Magnasco, M. (2001), 'Simple motor gestures for birdsongs', *Physical Review Letters*.

Gaunt, A. S. (1983), 'A hypothesis concerning the relationship of syringeal structure to vocal abilities', *Auk* **100**, 853–862.

Gaunt, A. S., Gaunt, S. L. L. & Casey, R. M. (1982), 'Syringeal mechanics reassessed: Evidence from *streptopelia*', *Auk* **99**, 474–494.

Gaunt, A. S., Gaunt, S. L. L., Prange, H. D. & Wasser, J. S. (1987), 'The effects of tracheal coiling on the vocalization of cranes (aves: Gruidae)', *J. comp. Physiol.* **161**, 43–58.

George, E. B. & Smith, M. J. T. (1997), 'Speech analysis/synthesis and modification using an analysis-by-synthesis/overlap-add sinusoidal model', *IEEE Trans. Speech and Audio Processing* **5**(5), 389–406.

Goller, F. & Larsen, O. N. (1997*a*), 'In situ biomechanism of the syrinxand sound generation in pingeons', *J. exp. Biol* **200**, 2165–2176.

Goller, F. & Larsen, O. N. (1997*b*), A new mechanism of sound generation in songbirds, *in* 'Proceedings of the National Academy of Sciences', Vol. 94, pp. 14787–14791.

Goller, F. & Larsen, O. N. (2002), 'New perspectives on mechanism of sound generation in songbirds', *J. comp. Physiol. A* **188**, 841–850.

Greenewalt, C. H. (1968), *Bird Song: Acoustics and Physiology*, Smithsonian Institution Press, Washington D.C.

Hartmann, W. M. (1997), *Signals, Sound, and Sensation*, 1 edn, AIP Press, Woodbury, New York, USA.

Hoese, W. J., Podos, J., Boetticher, N. C. & Nowicki, S. (2000), 'Vocal tract function in birdsong production: Experimental manipulation of beak movemets', *J. Exp. Biol.* **203**, 1845–1855.

Härmä, A. (2003), Automatic identification of bird species based on sinusoidal modelling of syllables, *in* 'IEEE Int. Conf. Acoust. Speech and Signal Processing'.

Härmä, A. (2004), 'Avesound-memo: Phylloscopus-suvun lintujen laulujen elementtien vertailu', http://www.acoustics.hut.fi/~sfagerlu/project/pubs/phyllos.pdf. in Finnish.

Härmä, A. & Somervuo, P. (2004), Classification of the harmonic structure in bird vocalization, *in* 'IEEE Int. Conf. Acoust. Speech and Signal Processing'.

Kanal, L. (1974), 'Patterns in pattern recognition', *IEEE Trans. Information Theory* **20**, 697–722.

King, A. S. (1989), Functional analysis of the syrinx, *in* '(King & McLelland 1989)', chapter 3, pp. 105–192.

King, A. S. & McLelland, J., eds (1989), *Form and Function in Birds*, Vol. 4, Academic Press.

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

Krebs, J. R. & Kroodsma, D. E. (1980), 'Repertoires and geographical variation in bird song', *Adv. Study Behav.* **11**, 143–177.

Laje, R., Gardner, T. J. & Mindlin, G. B. (2002), 'Neuromuscular control of vocalization in birdsong: A model', *Physical Review E* **65**, 051921.

Larsen, O. N. & Goller, F. (1999), 'Role of syringeal vibrations in bird vocalisations', *Proc. Roy. Soc. Lond.B* **266**, 1609–1615.

Li, D., Sethi, I. K., Dimitrova, N. & McGee, T. (2001), 'Classification of general audio data for content-based retrieval', *Pattern Recognition Letters* **22**, 533–544.

Mace, R. (1987), 'The dawn chorus in the great tit paras major is directly related to female fertility', *Nature* **333**, 123–132.

Markel, J. D. & Gray, A. H. (1976), *Linear Prediction of Speech*, 1 edn, Springer-Verlag, Berlin Heidelberg New York.

McAulay, R. J. & Quatieri, T. F. (1986), 'Speech analysis/synthesis based on a sinusoidal representation', *IEEE Trans. Acoustics, Speech and Signal Processing* **34**(4), 744–754.

McIlraith, A. L. & Card, H. C. (1997), 'Birdsong recognition using backpropagation and multivariate statistics', *IEEE Trans. Signal Processing* **45**(11), 2740–2748.

McKinney, M. F. & Breebaart, J. (2003), Features for audio and music classification, *in* 'Int. Conf. on Music Information Retrieval'.

McLelland, J. (1989), Larynx and trachea, *in* '(King & McLelland 1989)', chapter 2, pp. 69–103.

Müller, J. P. (1878), *On certain variations in the vocal organs of the Passeres that have hitherto escaped notice*, London: Macmillan.

Nelson, D. A. (1989), 'The importance of invariant and distinctive features in species recognition of bird song', *Condor* **91**, 120–130.

Nowicki, S. (1987), 'Vocal tract reconances in oscine bird sound production: Evidence from birdsongs in a helium atmosphere', *Nature* **325**(6099), 53–55.

Nowicki, S. (1997), Bird acoustics, *in* M. J. Crocker, ed., 'Encyclopedia of Acoustics', John Wiley & Sons, chapter 150, pp. 1813–1817.

Patterson, D. K. & Pepperberg, I. M. (1994), 'A comparative study of human and parrot phonation: Acoustic and articulatory correlates of vowels', *J. Acoust. Soc. Am.* **96**(2, Pt.1), 634–648.

Scheirer, E. & Slaney, M. (1997), Construction and evaluation of a robust multifeature speech/music discriminator, *in* 'IEEE Int. Conf. Acoust. Speech and Signal Processing', pp. 1331–1334.

Suthers, R. A. (1990), 'Contributions to birdsong from the left and right sides of the intact syrinx', *Nature* **347**(6292), 473–477.

Theodoridis, S. & Koutroumbas, K. (1998), *Pattern Recognition*, 1 edn, Academic Press, San Diego, California, USA.

Westneat, M. W., Long, J. H., Hoese, W. J. & Nowicki, S. (1993), 'Kinematics of birdsong: Functional correlation of cranical movements and acoustic features in sparrows', *J. exp. Biol* **182**, 147–171.

Wold, E., Blum, T., Keislar, D. & Wheaton, J. (1996), 'Content-based classification, search, and retrieval of audio', *IEEE Multimedia* **3**, 27–36.