TAMPERE UNIVERSITY OF TECHONOLOGY
Department of Information Technology (Pori)

ARJA SELIN

BIRD SOUND CLASSIFICATION USING WAVELETS

The topic has been accepted in the Departmental Council meeting on 19.1.2005
Examiners: Prof. Juha T. Tanttu
Dr. Tech Jari Turunen

PREFACE

The thesis work was done at Department of Information Technology, in the Pori Unit of Tampere University of Technology. The research has been done within the Avesound project and it is funded by the Academy of Finland.

I would like to thank Professor Juha T. Tanttu for his guidance, and Dr. Tech, senior researcher Jari Turunen for his encouraging comments and important support. I would also like to thank the nature-recording expert, Pertti Kalinainen, for his recordings and specialized comments, and Docent Mikko Ojanen for his helpful discussions.

I would also like to thank Tampere University of Technology, Information Technology Pori personnel for a pleasant work environment and English lecturer Kirsti Honkasalo for corrections in my thesis.

Pori 26.4.2005

Arja Selin

Arja Selin
Kraftmanintie 4 B 12
28610 PORI

TABLE OF CONTENTS

ABSTRACT

TAMPERE UNIVERSITY OF TECHONOLOGY
Department of Information Technology
Information Technology (Pori)

SELIN, ARJA: Bird Sound Classification using Wavelets
Master's Thesis, 50 pages
Examiners: Professor Juha T. Tanttu
Dr. Tech Jari Turunen
Funding: The Academy of Finland
April 2005
Keywords: Bird vocalization, wavelet transform, automatic recognition, MLP, SOM

UDC: 004.93, 598.2

Birds and especially sounds produced by birds have always fascinated people. Many biologists have found bird song to be an interesting area for example for behaviour and recognition research. Birds produce sounds generally using the sound channel. Most sounds are produced by the syrinx, which is the avian vocal organ. The syrinx is bipartite, so the bird can produce two notes simultaneously.

The automatic recognition of bird species by their sounds is a very interesting field. Due to the specific properties of bird sound, it is necessary to develop new signal processing techniques. Bird sounds can be tonal or inharmonic. The main idea in this thesis was to study, how inharmonic and transient bird sounds can be recognized and classified efficiently. The wavelet analysis was selected due to its ability to preserve both frequency and temporal information, and ability to analyse signals which contain discontinuities and sharp spikes, such as inharmonic bird sounds. It is possible to present the essential information of the sounds with only few features in the wavelet analysis.

Artificial neural networks (ANN) have been proved to be very powerful tools for pattern recognition and have successfully been used in the automated classification of acoustic signals including natural sounds. The two well-known neural networks, the multilayer perceptron (MLP) and the self-organizing map (SOM) were used as classifiers in this thesis. The data consisted of sounds of eight bird species, which were the Mallard (*Anas platyrhynchos*), the Graylag Goose (*Anser anser*), the Quail (*Coturnix coturnix*), the Corncrake (*Crex crex*), the Pygmy Owl (*Glaucidium passerinum*), the River Warbler (*Locustella fluviatilis*), the Magpie (*Pica pica*), and the Spotted Crake (*Porzana porzana*).

The results are quite encouraging, for the MLP classified 96 % and the SOM 93,8 % of the test sounds correctly. Wavelet analysis proved to be an efficient tool in the bird sound classification. Thus, the features form clusters so well that the classification can also be done with other types of classifiers.

TIIVISTELMÄ

TAMPEREEN TEKNILLINEN YLIOPISTO
Tietotekniikan osasto
Signaalinkäsittely (Pori)

SELIN, ARJA: Wavelet-muunnos linnunlaulun luokittelussa
Diplomityö, 50 sivua
Tarkastajat: Professori Juha T. Tanttu
TkT Jari Turunen
Rahoittaja: Suomen Akatemia
Huhtikuu 2005
Avainsanat: Linnun äänet, wavelet-muunnos, automaattinen tunnistus, MLP, SOM

UDK: 004.93, 598.2

Linnut ja niiden äänet ovat aina kiehtoneet ihmisiä. Monet biologit ovat tutkineet lintujen käyttäytymistä ja niiden tunnistamista laulun perusteella. Linnut tuottavat ääniä yleensä äänikanavan kautta. Useimpien äänien tuottajana toimii syrinx, joka on lintujen äänentuottoelin. Koska syrinx on kaksiosainen äänentuottomekanismi, linnut voivat tuottaa kahta ääntä samanaikaisesti.

Automaattinen lintulajien tunnistus niiden äänien perusteella on hyvin mielenkiintoinen alue. Lintujen äänten erityisistä ominaisuuksista johtuen on tarpeellista kehittää uusia signaalinkäsittelyn ja hahmontunnistuksen tekniikoita. Linnun äänet voivat olla tonaalisia tai epäharmonisia. Tämän diplomityön perusajatuksena oli tutkia, kuinka epäharmonisia ja transientteja linnun ääniä voidaan tunnistaa ja luokitella tehokkaasti. Wavelet-analyysiä käytettiin, koska sen avulla voidaan tutkia signaalista sekä aika- että paikkatietoja sekä analysoida siinä esiintyviä epäjatkuvuuksia ja nopeita muutoksia. Äänten olennainen informaatio on mahdollista esittää wavelet-analyysissä vain muutamilla piirteillä.

Neuraaliverkot ovat osoittautuneet hyvin tehokkaiksi työkaluiksi hahmontunnistuksessa ja niitä on tuloksekkaasti käytetty akustisten signaalien, kuten luonnon äänien, automaattisessa luokittelussa. Tässä työssä luokittelijoina on käytetty kahta tunnettua neuraaliverkkoa, MLP- (multilayer perceptron) ja SOM-verkkoa (self-organizing map). Ääninäytteet valittiin kahdeksalta lintulajilta, joita olivat sinisorsa (*Anas platyrhynchos*), merihanhi (*Anser anser*), viiriäinen (*Coturnix coturnix*), ruisrääkkä (*Crex crex*), viirupöllö (*Glaucidium passerinum*), viitasirkkalintu (*Locustella fluviatilis*), harakka (*Pica pica*) ja luhtahuitti (*Porzana porzana*).

Tulokset ovat hyvin rohkaisevia, sillä MLP luokitteli 96 % ja SOM 93,8 % testiäänistä oikein. Näin ollen wavelet-muunnos osoittautui tehokkaaksi työkaluksi linnun äänien luokittelussa. Piirteet klusteroituivat niin hyvin, että hyviä tuloksia saataneen muillakin luokittelijoilla.

ABBREVIATIONS

ANN	Artificial Neural Network
coifN	Coiflet wavelets
dbN	Daubechies wavelets
FIR	Finite Impulse Response Filter
MLP	Multilayer Perceptron
MSE	Mean Square Error
PR	Perfect Reconstruction
SOM	Self-Organizing Map
SVD	Singular Value Decomposition
WPD	Wavelet Packet Decomposition
WPT	Wavelet Packet Transform

1. INTRODUCTION

Birds and their sounds have always fascinated people. Bird song has been an important source of inspiration for example to many composers and writers. Many people are interested in bird watching. They are able to recognize at least few most common species by their sound, whereas an experienced birdwatcher can recognize hundreds of species only by their sound.

There are several channels for birds to communicate and sound is only one of them. For example visual signalling is also important to birds, but it has several restrictions. Darkness or poor light make visual signalling difficult. On the other hand, visual information needs nearness of the partners. Therefore sounds are a better method for communicating, because with sounds it is possible to transmit rapidly and efficiently large amounts of information. Sounds can be heard over long distance and long after the bird has moved out of sight. Sounds serve a way to communicate with other conspecifics and also between different species. Sounds are only produced when needed, and so, all sounds have some meaning. Birds have many different types of sounds and many call types exist among them, for example alarm, mating or flight calls. Thus, each call type can be associated with a specific context. In addition to calls, many birds have very complex songs. The best time to hear bird song is at dawn, when birds spend such a considerable amount of time and energy on communicating with each other.

Birds produce sounds using the organ that resides in the sound channel. Most sounds are produced by the syrinx, which is the avian vocal organ. The syrinx is bipartite, so the bird can produce two notes simultaneously. For example, Zollinger et al. (2003) have studied the vocalization of Northern Mockingbirds (*Minus polyglottos*), and they have compared spectrographic phenomena in both sound sources and one side only. They have shown that each side of the duplex syrinx can be used independently. According to Suthers et al. (2003) pure-tone vocalizations in doves (*Streptopelia risoria* and *S. decaocto*) and likely in songbirds can be generated by a source filter mechanism. It appears that the respiratory muscles play a very important role in regulating the fundamental frequency. They have shown that the left and right sides of the syrinx have different (but overlapping) ranges of fundamental frequencies. Thus, birds are unique in possessing a twin sound-producing organ in contrast to the single source characteristic of mammals.

Since Charles Darwin, many biologists have found bird song to be a fascinating area for research. A lot of scientific study of bird song has been made and it has contributed to for example such areas as neurobiology and evolutionary biology. For example, Elowson and Hailman (1991) have studied the acoustically complex predator-elicited calls of the Florida Scrub Jay (*Aphelocoma coerulescens*). They have designed a dichotomous sorting paradigm that defines call types by their measured acoustical parameters. In many species there is high individual and regional variability in phrases and song patterns. Groth (1993) has studied the morphological variability in Red Crossbills (*Loxia curvirostra*) in North America and has illustrated that members of pairs show the call-matching and positive assortative mating phenomena. Robb (2000) has studied the vocalizations of Crossbills (*Loxia* spp.) in northwestern Europe. He has assumed that there exists regional variability. Both Groth and Robb have examined the fast evolution of the Crossbills (*Loxia* spp.), which is the reason why the Two-barred Crossbill (*Loxia leucoptera*), Common Crossbill (*L. curvirostra*), Scottish Crossbill (*L. scotica*) and the Parrot Crossbill (*L. pytyopsittacus*) are distinctly distinguishable with their different sounds. Bradbury (2003) has characterized the levels of variation in wild parrot (*Amazona auropalliata*, *Amazona albifrons*, *Aratinga canicularis*, and *Brotogeris jugularis*) vocal repertoires and found repeatable individual differences in loud contact calls. Lovell and Lein (2004) have studied individual song variation in a population of Alder Flycatchers (*Empidonax alnorum*). Their results have indicated that there is sufficient variation among males to permit the classification of the songs of individuals. In most of these studies the feature selection has been made manually. There exist only few studies where the identification of bird species is made automatically.

Human ear and brain constitute an effective voice recognition system. Consequently, for the human ear it is relatively easy to notice even subtle differences in sounds. For the computers the recognition task is much more difficult. Due to specific properties of bird sound, it is necessary to develop new signal processing and pattern recognition techniques. That is why the automatic recognition of bird species by their sounds is a very interesting and challenging field. For example, Härmä (2003) has studied automatic identification of bird species using sinusoidal modelling of syllables. In their further study (Härmä and Somervuo 2004) the syllables were modelled using a parametric line spectrum estimation method. The sounds were divided into four classes by their harmonic structure. Mesgarani and Shamma (2003) have shown that with a multiresolution spectrotemporal auditory model it is possible to classify the birdcall. Tanttu et al. (2003) have proposed a method for automatic classification of Crossbills (*Loxia* spp). This study was based on the same data that Robb (2000) has used. The proposed method was based on the tracking of the first harmonic components of the spectrogram. Somervuo and Härmä (2004) have shown the possibility of bird species recognition based on the syllable pair histogram of the song, where the nearest neighbour classifier was used. Baker and Logue (2003) have examined the call of the Black-capped Chickadee (*Poecile atricapillus*). They have used three bioacoustical analysis methods for comparing these complex sounds among different populations. Fagerlund (2004) has shown in his thesis, how inharmonic bird sounds can be classified using 19 low level parameters of syllables and k-Nearest-Neighbour nonlinear classifier.

Artificial neural networks (ANN) have been showed to be very powerful tools for pattern recognition and have successfully been used in the automated classification of acoustic signals including animal sounds. For example, Phelps and Ryan (1998) have studied the call of the Tungara Frog (*Physalaemus pustulosus*) with a neural network. They have shown that a mate recognition signal capable of eliciting responses from females can be recognized with a simple neural network. Deecke et al. (1999) have measured the similarities of discrete calls of Killer Whales (*Ornicus orca*) using artificial neural network. Placer and Slobodchikoff (2000) have illustrated that Gunnison's Prairie Dogs (*Cynomys Gunnisoni*) have different alarm calls for different species of predators. Thorn (2003) has used the modified self-organizing map (SOM) to classify the vocal repertoire of the Barbary macaque (*Macaca sylvanus L.*). Due to Thorn the methodology is intended to be general enough to extend to any species.

The ANN has also been used in the classification and recognition of bird sounds. In (McIlraith and Card 1997) the bird song recognition was made using backpropagation learning in two-layer perceptrons and using several methods from multivariate statistics. They have used spectral and time variables as features in recognition of six species of birds: the Song Sparrow (*Melospiza melodia*), the Fox Sparrow (*Passerella iliaca*), the Marsh Wren (*Cistothorus palustris*), the Sedge Wren (*Cistothorus platensis*), the Yellow Warbler (*Dendroica petechia*) and the Red-winged Blackbird (*Agelaius phoeniceus*). Terry and McGregor (2002) have studied the call of the Corncrake (*Crex crex*) and proposed that Corncrakes have individually distinctive vocalizations. They have studied how to simulate census and monitoring tasks involving these vocalizations. The backpropagation and the probalistic network have been used for reidentification tasks and the SOM network for discrimination between unknown numbers of individuals and counting the Corncrakes. Somervuo and Härmä (2003) have used the SOM when analysing the song syllables. These encouraging results have motivated the researcher to experiment with neural networks in this thesis.

In the efficient coding of natural sounds, for example bird sound, the goal is to encode the maximal amount of information about the stimulus by using a set of statistically independent features. According to Lewicki (2002) a bank of linear filters offers one way to code animal sound efficiently. The filter banks can model the frequency and phase responses of auditory nerve fibres of human and other animals. The basic idea of filter banks leads fluently to the wavelets. In the wavelet transform the connection between time and frequency remains strongly, which is important in analysing the natural sounds. This has inspired researcher to evolve a new method for automatic classification of bird sounds.

The wavelet transform can be compared to information gathered by our own eyes and ears, which give us location as well as frequency. The wavelets are being applied in a large and growing range of applications (Nason 1999), from which data and image compression (for example digitised fingerprints of the FBI), partial differential equation solving, transient detection (for example seismic data), pattern detection, texture analysis and noise reduction can be mentioned. The main advantages of wavelet methods are the use of localized basis functions and the faster computation speed than using other analysing methods. Localized basis functions are ideal for analysing real physical situations in which a signal contains discontinuities and sharp spikes.

The wavelet packet transform (WPT) has gained a great deal of attention in the field of digital signal processing (Rioul and Vetterli 1991). In the WPT the original signal is converted into wavelet coefficients. Because the parameter assessment using all wavelet coefficients will often turn out to be tedious or leads to inaccurate results, the extraction of the most important features is essential. For example, Pittner and Kamarthi (1999) have studied the feature extraction from wavelet coefficients. The wavelet packet decomposition (WPD) is able to trade time and frequency resolutions, maintaining the orthogonality, and perfect reconstruction (PR). The orthogonal packets can be designed by hierarchical association of PR paraunitary filter banks (Soman and Vaidyanathan 1992). Learned (1992) has used the WPT to the transient signal classification in her thesis. The energy maps of the wavelet coefficients were made from the whale clicks, the snapping shrimp and the background noise. The singular value decomposition (SVD) was used to reduce the information and the ANN was used as the final classifier.

Bird sounds can be tonal or inharmonic, which is one way to divide the bird species into different groups. Inharmonic sounds are often transients and their frequency contents are very near each other. The convenient spectral analysis methods are useful for the process of tonal and harmonic sounds classification, but inharmonic and transient sounds are not well matched to these methods.

The main purpose of this thesis was to study, how inharmonic or transient bird sounds can be recognized and classified efficiently. Hence, in this thesis the use of the WPT is illustrated in the context of automatic recognition of bird sounds. Two commonly known neural networks, multilayer perceptron (MLP) and SOM are used as classifiers. The thesis is divided into six Chapters. In Chapter 2 sound production of the birds and the different sound types are described. The theory of wavelets is presented in Chapter 3.This chapter first presents the scaling and the wavelet functions, second the wavelet families, and finally the wavelet decomposition. Chapter 4 focuses on the recognition methods applied in this thesis. First the whole process is illustrated as the block diagram and then the methods are discussed in detail. The data used in this thesis is introduced and the results is summarised in the Chapter 5. Conclusions and future work are discussed in the Chapter 6.

2. BIRD SOUND

The following discussion on the sound production and sound types of bird species is based on Greenewalt (1968, pp. 23-30, 111, 114) and Catchpole and Slater (1995, pp. 9-10, 21-26). Nearly all birds make different kind of sounds because these help them to communicate with each other. Birds have many different types of sounds, which are produced in the sound channel. These sounds vary from short and simple call notes to long and complex songs. In addition, birds use other voices, which are produced for example by clapping the wings or by clacking the mandibles together. Birds use their vocal chords to produce vocalisations, which can be divided into songs and calls. The sound communication system of the birds is divided into two parts. A bird sends the signal to the beak using a special sound-producing organ, the *syrinx* that propagates the sound into the air. The other half is hearing and perception, which are used in receiving signals.

2.1 Sound production

Birds use their lungs, bronchi, syrinx, trachea and beak in sound-production. Larynx is located in the upper end of the trachea. Unlike the humans, birds do not need the larynx in sound production but the larynx is needed to protect the respiratory system for example during the feeding. The avian double vocal organ, called the syrinx, produces songs and calls. It is located deep in the bird's chest. Figure 2-1 illustrates how the syrinx is located.

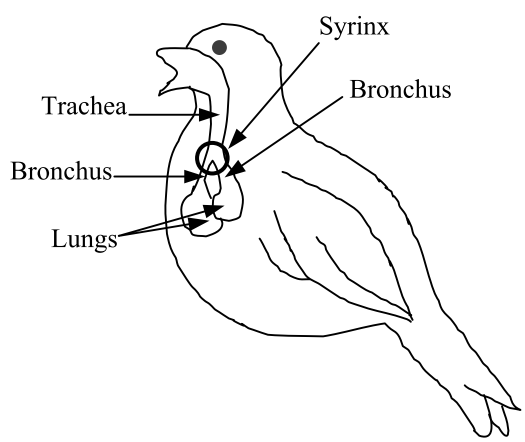


Figure 2-1. The location of the syrinx.

The syrinx is situated at the point where the trachea divides into two bronchi. An independently controllable part of the syrinx resides in each bronchus, so these two sound sources are used to generate the final sound. Some birds can sing two different notes or phrases at the same time. However, one side of the system appears to dominate the other. The section through the syrinx can be seen in Figure 2-2. The medial tympanic membranes are situated on the medial walls of the bronchus. They appear to be in the most important role in sound production. External labia and the medial part of the interclavicular airsac have also a very important role. An independently controllable part of the syrinx is outlined with the orange dash line in Figure 2-2.

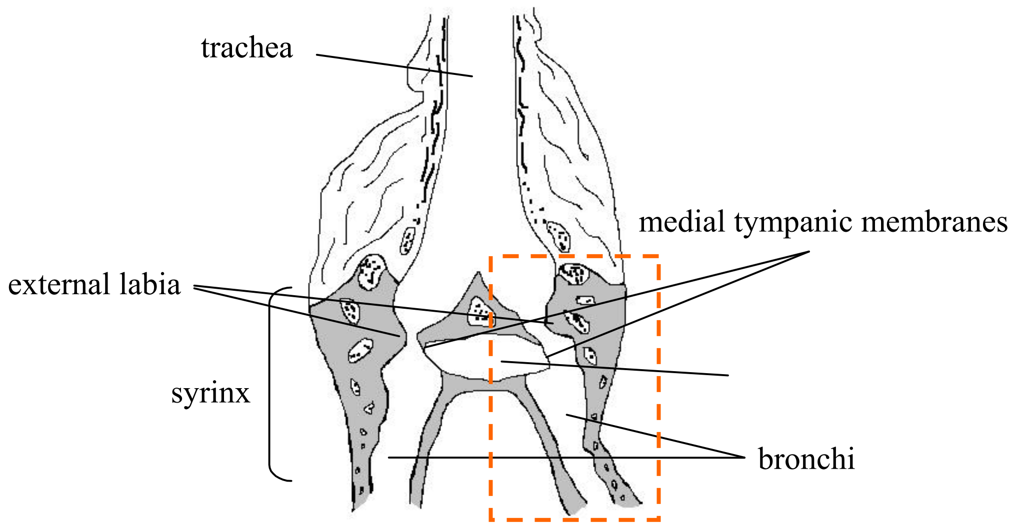
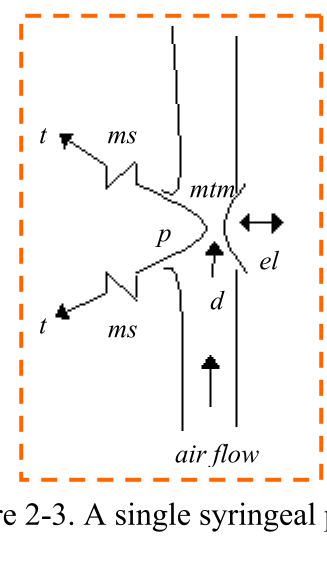


Figure 2-2. Section through the syrinx.

The diagrammatic representation of the syrinx is illustrated in Figure 2-3. The area bounded with the orange dash line represents a single syringeal passage. The same area is illustrated in Figure 2-2. First air flows from the lungs to bronchial passage. The pressure (*p*) is build up in the medial part of the interclavicular airsac, which is surrounded by the syrinx. This pressure forces the medial tympanic membrane (*mtm*) into bronchial lumen. The syringeal muscles apply tension (*t*) to the medial tympanic membrane. This tension is in opposition to the medial part of the interclavicular airsac pressure and produces a narrow passage (*d*) in the bronchus. The movement of the external labium (*el*) provides a method for controlling the amplitude. This does not affect the tension. The amplitude of vibration will increase when the passage increases. Both frequency and amplitude can be affected by changing the membrane tension.



d	passage
el	external labium
mtm	medial tympanic membrane
ms	membrane by syringeal muscles
p	pressure in the interclavicular sac
t	tension

Figure 2-3. A single syringeal passage.

Between the syrinx and the beak there is a vocal tract, which is used to modify the final sound. The tracheal tract acts like a selective filter, which can be adjusted in a variety of ways. The stretching and retracting of the neck lengthen and shorten the vocal tract and affect in this way the final sound. The quality of sound can be further influenced by constricting the larynx, by muscles in the throat or by the structure and movements of the beak.

2.2 Sound types

Nearly all birds use sounds to communicate. However, all birds do not sing and not all the sounds that birds produce are called songs. Vocalisation of the birds contains both songs and calls. Most birds use vocalisations, which are short and unmusical. Less than half of the birds on the planet are genuine songsters. Only the birds of the order *Passeriformes* sing songs. According to Catchpole and Slater (1995, p. 10) songs tend to be long and complex. Males produce song during the breeding season. Some birds have many versions of their species-specific songs and most of the birds have at least more than one. Consequently complexity is one of the main characteristics of songs. Recognition of the own species seems to be the most important function of the song.

Calls on the contrary are shorter and simpler than songs. Both sexes produce calls throughout the year. Calls have considerable functionality in the daily life of birds. Therefore they are less spontaneous. Calls can be divided into many different categories. Functions of calls can be flight, alarm, excitement, threat, begging, feeding, defence, pleasure and so on. Many birds have more than one call for one category but then some birds seem to use very similar calls in different circumstances to mean different things. The passerines have a greater repertoire of calls than non-passerines. However, it appears that most birds seem to have from 5 to 15 distinct calls. All these calls allow birds to communicate with each other and with the rest of the world.

Table 1-1. English and scientific names of the selected set of bird species used in this thesis

English name	Scientific name
Mallard	*Anas platyrhynchos*
Greylag Goose	*Anser anser*
Quail	*Coturnix coturnix*
Corncrake	*Crex crex*
Pygmy Owl	*Glaucidium passerinum*
River Warbler	*Locustella fluviatilis*
Magpie	*Pica pica*
Spotted Crake	*Porzana porzana*

Sounds of eight species, the Mallard, the Greylag Goose, the Quail, the Corncrake, the Pygmy Owl, the River Warbler, the Magpie, and the Spotted Crake, are studied in this thesis. Table 1-1 lists the English names and corresponding scientific names of this selected set of birds.

Most of these eight bird species have monotonous sounds, which remain almost the same all the time. For example, the Corncrake has nearly an unchangeable call. This distinctive call is often the only contact that it is possible to have with this elusive species. Also the whip-like calls of the Spotted Crakes and the diagnostic call of the Quail remain near to identical. The Pygmy Owl has many sound types from which only the monotonous territorial song of the male is used in this thesis. The River Warbler is a small passerine bird, whose song is a monotonous mechanical insect-like reeling. This song sounds like a sewing machine and also remains identical all the time. Squawk calls of the Mallard and the Greylag Goose are typical for those species. The calls of the Mallard remain quite unchangeable whereas the Greylag Goose has several types of calls. Many different calls of the Greylag Goose are included in this thesis. Also the Magpie has several types of sounds, from which only one type of call is used in this thesis. The idea was to choose such bird species whose sounds are inharmonic and similar with each other. The Mallard, the Greylag Goose, the Corncrake, the River Warbler and the Magpie are such species. The pure tonal sound of the Pygmy Owl was chosen to be a reference sound. The sound of the Quail and the Spotted Crake are harmonic, but they have some inharmonic features, for example irregular pitch period.

3. WAVELETS

The first recorded mention of wavelets was in Alfred Haar's thesis in 1909, although he did not call his function a "*wavelet*" (Strang and Nguyen 1996, p. 24). The wavelet analysis was introduced at the beginning of the 1980s, when Morlet evaluated seismic data (Mertins 1999, p. 210; Sheng 2000, p. 10-2). So, most of the applications using wavelets have emerged in the last twenty years (Berry 1999). In the late 1980's Ingrid Daubechies, who is the most well known wavelet researcher (Misiti et al. 2000, p. 1-33), developed the theory of wavelet analysis.

Fourier transform has long been the most widely used transform in many applications. In the Fourier analysis the sinusoidals are used as the basis functions, whereas in the wavelet analysis the basis function is in the form of a wavelet (Tan and Gan 1999). The wavelet analysis has an advantage over traditional Fourier analysis, for example if signals have spikes and discontinuities (Berry 1999). Signals with sharp changes can usually be better analysed with an irregular wavelet than with a smooth sinusoid (Misiti et al. 2000, p. 1-9).

Wavelet analysis is a general mathematical tool, which enables an efficient analysis of the small details and the long-term properties of the signal simultaneously (Nason 1999; Bruce et al. 2003). It is possible to find out both frequency and temporal information with wavelet analysis, so that an event stays connected to the time occurrence (Umbaugh 1998, p. 125). That is why the wavelet transforms are much more local than the classical transforms. The track of time and frequency information can be maintained during wavelet analysis (Boggess and Narcowich 2001, p. 156). Wavelet analysis represents a windowing technique with variable-sized regions. The long windows are used at low frequencies and short windows at high frequencies (Rioul and Vetterli 1991; Misiti et al. 2000, p. 1-7). In Figure 3-1 the time-scale region of wavelet analysis is presented (Strang and Nguyen 1996, p. 4).

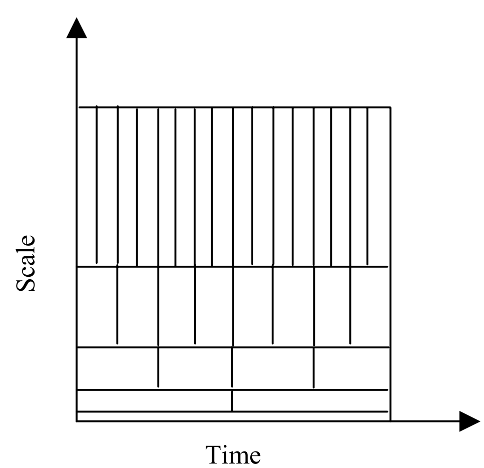


Figure 3-1. Time-scale region in the wavelet analysis.

3.1 Scaling and wavelet functions

In the wavelet analysis a signal is decomposed into weighted coefficients of wavelet functions (Boggess and Narcowich 2001, p. 254; Bruce et al. 2003). The whole transform is based on two functions called the *scaling function* ϕ and the *wavelet function* ψ. (Boggess and Narcowich 2001, p. 156; Rioul and Vetterli 1991). The scaling function ϕ and the wavelet function ψ generate a family of functions and they can be used to break down (decomposition) or build (reconstruction) a signal (Boggess and Narcowich 2001, p. 156). The scaling function ϕ is commonly called the father wavelet and the wavelet function ψ the mother wavelet (Nason 1999; Tan and Gan 1999; Berry 1999; Boggess and Narcowich 2001, p. 156; Jin et al 1994).

The scaling function $\phi(t)$ defines the scales at which the wavelets operate (Tan and Gan 1999). So, it allows short wavelets to be used with transient signals and precisely locate discontinuities and irregularities, and correspondingly, longer wavelets to be used to represent lower frequencies in the sampled signal. (Berry 1999) According to Sheng (Sheng 2000, p. 10-29) the scaling function may be decomposed as a linear combination of the scaling functions at the higher resolution.

A wavelet is a small wave and it has to have two properties. Function ψ is a wavelet, if the two equations

$$\int_{-\infty}^{\infty} |\psi(t)|^2 \, dt < \infty, \qquad \psi \in L^2(\mathbb{R}), \tag{3-1}$$

$$\int_{-\infty}^{\infty} \psi(t) dt = 0 \tag{3-2}$$

are true, where t indicates time (Boggess and Narcowich 2001, p 254). Equation (3-1) means that if wavelet ψ is a physical signal, its energy is finite. A wavelet function ψ is usually continuous and has exponential decay, which is defined as

$$\left|\psi(t)\right| \le Me^{-C|t|} \qquad \forall t \in \mathbb{R} \tag{3-3}$$

for some constants C and M (Boggess and Narcowich 2001, p. 254; Nason 1999). Wavelets are oscillating functions that are confined within a finite time span (Tan and Gan 1999).

The L^2 inner product of the functions, $f(t)$ and $g(t)$, is defined as

$$\langle f, g \rangle = \int_{m_1}^{m_2} f(t)\overline{g(t)} \, dt \qquad \text{for } f, g \in L^2\left(\left[m_1, m_2\right]\right) \tag{3-4}$$

where $\overline{g(t)}$ is a complex conjugate of the $g(t)$ (Boggess and Narcowich 2001, p. 5). The norm of the wavelet ψ is defined as

$$\|\psi\|^2 = \langle \psi, \psi \rangle . \tag{3-5}$$

The wavelet basis is orthogonal, if

$$\langle \psi_i, \psi_j \rangle = 0 \qquad i \neq j. \tag{3-6}$$

Hence, the inner product is zero over interval $(-\infty, \infty)$. If, in addition,

$$\int_{-\infty}^{\infty} |\psi(t)|^2 \, dt = 1, \tag{3-7}$$

the wavelet is normalized and orthogonal. Hence, the perfect basis is orthogonal and orthonormal, if the inner product $\langle \psi_i, \psi_j \rangle = \delta_{ij}$. (Boggess and Narcowich 2001, p. 12; Strang and Nguyen 1996, p. 25)

It is important that a wavelet has c*ompact support*, which means that it is zero outside a bounded set. *Vanishing moments* are the key factors in many wavelet applications. The *k*:th moment M_k of wavelet ψ is defined as

$$M_k = \int_{-\infty}^{\infty} t^k \psi(t) dt . \tag{3-8}$$

A wavelet has *N* vanishing moments, if

$$M_k = \int_{-\infty}^{\infty} t^k \psi(t) dt = 0, \quad \text{for } \; 0 \leq k \leq N-1 . \tag{3-9}$$

If (3-10) holds polynomials of degree *N*-1,

$$p(t) = \sum_{k=0}^{N-1} d_k t^k \tag{3-10}$$

have zero wavelet coefficients, where d_k denotes the coefficients of the polynomial signal (Misiti et al. 2000, p. 3-17). This means that smooth parts of the functions will have zero wavelet coefficients associated with them. (Nason 2000; Misiti et al. 2000, p. 6-49) By shifting and scaling the wavelet ψ it is possible to make the whole *wavelet family* (Daubechies 1992, p. 24; Rioul and Vetterli 1991; Nason 1999). The wavelet family is defined as

$$\psi_{a,b}(t) = \frac{1}{\sqrt{|a|}}\psi\left(\frac{t-b}{a}\right) \tag{3-11}$$

where $a,b \in \mathbb{R}$ and $a \neq 0$.

The wavelet family function $\psi_{a,b}(t)$ depends on parameters *a* and *b*. Scaling a wavelet means stretching or compressing it (Misiti et al. 2000, p. 1-11). The time and frequency resolution of the wavelet transform depends on the scaling parameter *a* (Mertins 1999, p. 211). When the scale factor *a* decreases, the wavelet becomes more compressed (Misiti et al. 2000, p. 1-12; Boggess and Narcowich 2001, p. 255). At the same time when *a* decreases, the high frequency content of $\psi_{a,b}(t)$ increases. The stretched wavelet tracks coarse changes in the signal and is related to low frequencies (Tan and Gan 1999). Figure 3-2 illustrates the changes of the scaling parameter *a* in the graphs of $\psi_{1,0}$, $\psi_{1/2,0}$ and $\psi_{1,1}$ with given $\psi(t) = te^{-t^2}$.

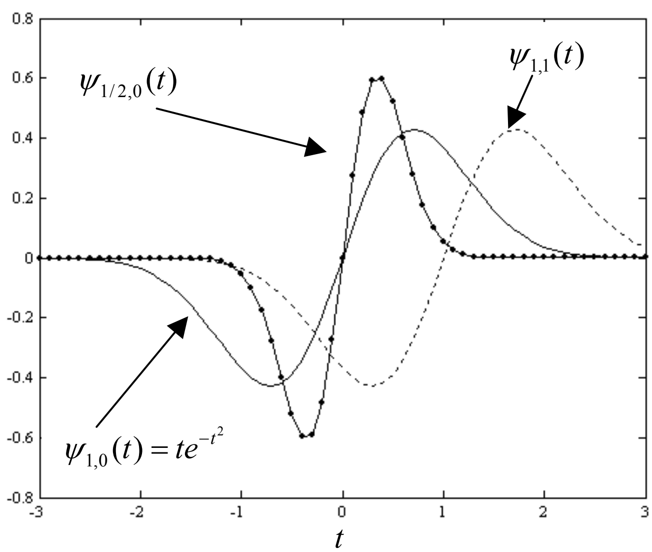


Figure 3-2. Graphs of $\psi_{1,0}$, $\psi_{1/2,0}$ and $\psi_{1,1}$.

The wavelet function can be shifted with the parameter *b* (Boggess and Narcowich 2001, p. 255; Bruce et al. 2003). The variation of shift factor *b* simply means a translation in time (Mertins 1999, p. 211), as shifting a wavelet means delaying or hastening its onset (Misiti et al. 2000, p. 1-12). The graph of $\psi_{1,1}$ in Figure 3-2 illustrates that if the value of *b* is changed, the graph is shifted.
Both scaling parameter *a* and shifting parameter *b* can be discretized. In that case the discrete wavelet family $\psi_{j,k}(t)$ becomes

$$\psi_{j,k}(t) = 2^{j/2}\psi(2^j t - k) \qquad j,k \in \mathbb{Z}. \tag{3-12}$$

Correspondingly, the scaling family function $\phi_{j,k}(t)$ is defined as

$$\phi_{j,k}(t) = 2^{j/2}\phi(2^j t - k) \qquad j,k \in \mathbb{Z}. \tag{3-13}$$

(Sheng 2000, p. 10-29; Boggess and Narcowich 2001, p. 187) These discrete scaling $\phi_{j,k}(t)$ and wavelet family $\psi_{j,k}(t)$ will be used in this thesis.

3.2 Wavelet families

There are several wavelet families that have been proved to be particularly useful (Misiti et al. 2000, p. 1-32). The prototype of a wavelet is the *Haar wavelet*, which is discontinuous, and resembles a step function (Kammler 2000, p. 594; Misiti et al. 2000, p. 1-33). The Haar family constitute an orthonormal basis (Daubechies 1992, p. 10). The Haar wavelet $\psi(t)$ is defined as

$$\psi(t) = \begin{cases} 1, & \text{if } 0 \le t < \frac{1}{2} \\ -1, & \text{if } \frac{1}{2} \le t < 1 \\ 0, & \text{otherwise} \end{cases} \tag{3-14}$$

and is depicted in Figure 3-3.

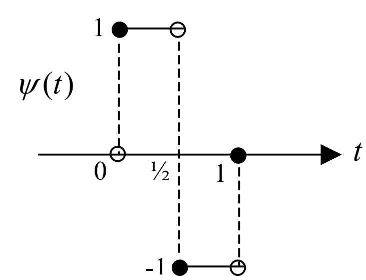


Figure 3-3. Graph of the Haar wavelet $\psi(t)$.

Ingrid Daubechies discovered the hierarchy of the wavelets. The *Daubechies wavelets* (dbN, where N is the order) become increasingly smooth by going up the hierarchy (Boggess and Narcowich 2001, p. 227). The Daubechies wavelet basis is a family of orthonormal, compactly supported scaling and wavelet functions (Sheng 2000, p. 10-74).

It is possible to control the smoothness of the wavelet analysis with the Daubechies wavelets (Nason 1999). The db1 wavelet is the same as the Haar wavelet. The Daubechies wavelets are classified according to the vanishing moments they have. The orthonormal Daubechies wavelets have a maximum number of vanishing wavelet moments for a given support (Mertins 1999, p. 253). So, the smoothness in the wavelet function increases with the number of vanishing moments (Boggess and Narcowich 2001, 232). Most dbN wavelets are not symmetrical, but the regularity increases with the order N. In Figure 3-4 the graphs of the three different Daubechies' scaling functions are shown. The corresponding wavelet functions are shown in Figure 3-5.

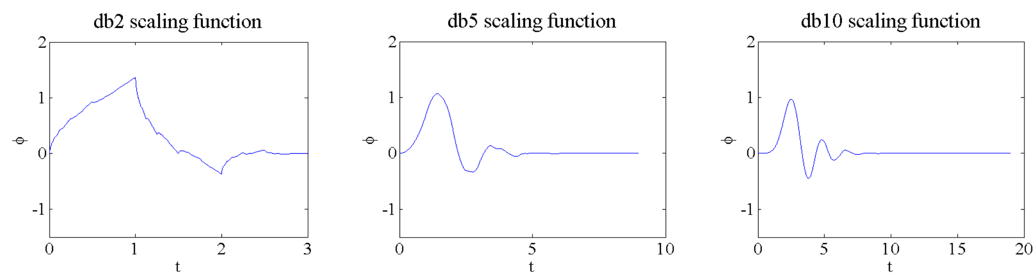


Figure 3-4. Daubechies' scaling functions db2, db5 and db10.

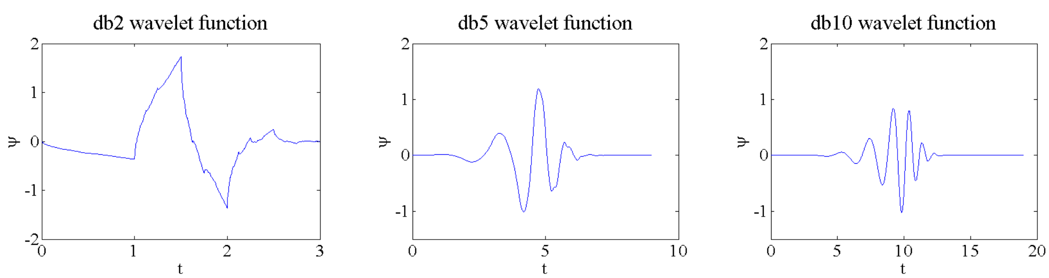


Figure 3-5. Daubechies' wavelet functions db2, db5 and db10.

Daubechies built the *Coiflet family* at the request of R. Coifman. In the Coiflet wavelet family (coifN, where N is the order) wavelet ψ and scaling functions ϕ are more symmetrical than the dbNs. (Misiti et al. 2000, p. 6-63) The idea is to trade off some vanishing wavelet moments to the scaling function. *Symlets* are Daubechies wavelets with maximal symmetry. Their properties are similar to the Daubechies wavelets. (Mertins 1999, p. 253) Many other wavelet families exist, for example *Biorthogonal*, *Morlet*, *Mexican hat* or *Meyer* that may be applied to signal and image processing applications (Misiti et al. 2000, pp. 1-32 – 1-37).

According to Berry (1999) there is no wavelet family which is best suited for all types of measurements. The data plays a part when the most appropriate wavelet is chosen for analysis. For example, some wavelets are better for edge detection or noise reduction than others. It is also possible to create a new wavelet family for special purposes. In that case carefulness is needed because it has to be checked that the self-created wavelet meets all the mathematical requisites (Misiti et al. 2000, p. 7-2).

3.3 Wavelet decomposition

The following discussion on the wavelet decomposition is based on Boggess and Narcowich (2001, pp. 155-200). In the wavelet *decomposition* a signal is broken down with the scaling function ϕ and the wavelet function ψ. An effective way to decompose a signal is to use a *multiresolution analysis* together with a continuous scaling function ϕ and the wavelet function ψ. In the multiresolution analysis *subspaces* V_j and W_j are defined, where V_j is the scaling or approximation space and W_j is the wavelet or detail space. Each V_j is contained in the next subspace V_{j+1}. Thus, the following holds:

$$V_0 \subset V_1 \subset \ldots \subset V_j \subset V_{j+1} \subset \ldots \tag{3-15}$$

The space $L^2(\mathbb{R})$ can be decomposed as an infinite orthogonal direct sum

$$L^2(\mathbb{R}) = V_0 \oplus W_0 \oplus W_1 \oplus \ldots \tag{3-16}$$

and so a signal $s \in L^2(\mathbb{R})$ has a piece in each subspace. In order to carry out the decomposition algorithm, V_j is decomposed into an orthogonal direct sum of V_{j-1} and its orthogonal complement, which is denoted W_{j-1}:

$$V_j = V_{j-1} \oplus W_{j-1} \tag{3-17}$$

(Boggess and Narcowich 2001, pp. 156, 161, 183; Strang and Nguyen 1996, p. 165, 175)

Suppose $\{V_j; j \in \mathbb{Z}\}$ is a multiresolution analysis with a scaling function

$$\phi(t) = \sum_{k \in \mathbb{Z}} p_k \phi(2t-k) \tag{3-18}$$

that is compactly supported, and where p_k are the coefficients

$$p_k = 2\int_{-\infty}^{\infty} \phi(t)\overline{\phi(2t-k)}\, dt\,. \tag{3-19}$$

Let W_j be the span of $\left\{\psi\left(2^j t-k\right); k \in \mathbb{Z}\right\}$, where

$$\psi(t)=\sum_{k\in\mathbb{Z}}(-1)^k\overline{p_{1-k}}\phi(2t-k)\ . \tag{3-20}$$

Then $W_j\subset V_{j+1}$ is the orthogonal complement of V_j in V_{j+1}. By successive orthogonal decompositions,

$$\begin{aligned}V_j&=W_{j-1}\oplus V_{j-1}\\&=W_{j-1}\oplus W_{j-2}\oplus V_{j-2}\\&\quad\dots\\&=W_{j-1}\oplus W_{j-2}\dots W_0\oplus V_0.\end{aligned} \tag{3-21}$$

At the beginning of the decomposition it is supposed that a signal *s* belongs to the approximation space V_j. The signal *s* can be presented with two primary orthonormal bases. The natural scaling function basis is defined in (3-13). In that case the signal *s* has the form

$$s=\sum_{k\in\mathbb{Z}}\left\langle s\ ,\ \phi_{jk}\right\rangle\phi_{jk}\ . \tag{3-22}$$

The second basis has the form

$$s=\sum_{k\in\mathbb{Z}}\left\langle s\ ,\ \phi_{j-1,k}\right\rangle\ \phi_{j-1,k}+\sum_{k\in\mathbb{Z}}\left\langle s\ ,\ \psi_{j-1,k}\right\rangle\ \psi_{j-1,k}\ . \tag{3-23}$$

The decomposition formula starts with the coefficients relative to the basis in (3-22) and uses them to calculate the coefficients relative to the basis in (3-23). The whole decomposition formula can be defined with two equations as

$$\left\langle s\ ,\ \phi_{j-1,\,l}\right\rangle=2^{-1/2}\sum_{k\in\mathbb{Z}}\overline{p_{k-2l}}\ \left\langle s\ ,\ \phi_{jk}\right\rangle, \tag{3-24}$$

$$\left\langle s\ ,\ \psi_{j-1,\,l}\right\rangle=2^{-1/2}\sum_{k\in\mathbb{Z}}(-1)^k\ p_{1-k+2l}\ \left\langle s\ ,\ \psi_{jk}\right\rangle \tag{3-25}$$

Equation (3-23) can be rewritten in the form:

$$s=\sum_{k\in\mathbb{Z}}a_k^{j-1}\ \phi\left(2^{j-1}t-k\right)+\sum_{k\in\mathbb{Z}}b_k^{j-1}\ \psi\left(2^{j-1}t-k\right) \tag{3-26}$$

where a_k^j's are called the approximation coefficients and the b_k^j's are called the detail coefficients and they are defined as

$$a_k^{j-1}=2^{(j-1)/2}\left\langle s,\ \phi_{j-1,\,k}\right\rangle, \tag{3-27}$$

$$b_k^{j-1} = 2^{(j-1)/2} \left\langle s,\ \psi_{j-1,\,k} \right\rangle. \tag{3-28}$$

The decomposition algorithm of the signal *s* consists of three major steps: initialisation, iteration, and termination (Boggess and Narcowich 2001, pp. 197-200). In the *initialisation* step the approximation space V_j has to be selected so that it best fits the information available on signal *s*, and the approximation $s_j \in V_j$ has to be chosen to best fit *s* itself. With the sampled signal *s* it is enough to determine the coefficients a_k^j as follows:

$$a_k^j = 2^j \int_{-\infty}^{\infty} s(t)\overline{\phi(2^j t - k)}dt \approx m\, s(k/2^j) \tag{3-29}$$

where $m = \int_{-\infty}^{\infty} \overline{\phi(t)}dt$. The accuracy of the approximation increases with increasing j. Hence, it is important to pick j large enough so that all of the coefficients are accurately calculated.

After the initialisation the signal $s \approx s_j \in V_j$, the second and the most effective step with the wavelets is the *iteration*. The approximation signal s_j is decomposed into a sum of a lower level approximation part, $s_{j-1} \in V_{j-1}$, and a wavelet part (detail), $w_{j-1} \in W_{j-1}$. Thus,

$$s_j = s_{j-1} + w_{j-1}\,. \tag{3-30}$$

The iteration is illustrated in Figure 3-6.

$$\begin{tikzcd}
s \approx s_j \arrow[r] \arrow[dr] & s_{j-1} \arrow[r] \arrow[dr] & s_{j-2} \arrow[r] \arrow[dr] & \cdots & \arrow[r] \arrow[dr] & s_1 \arrow[r] \arrow[dr] & s_0 \\
& w_{j-1} & w_{j-2} & \cdots & & w_1 & w_0
\end{tikzcd}$$

Figure 3-6. The iteration step of the decomposition.

The whole wavelet decomposition can be done with the approximation and wavelet coefficients, the *a*'s and *b*'s. Using the two discrete filters H_0 (low-pass) and H_1 (high-pass), and the downsampling by two $(\downarrow 2)$ after the filtering the decomposition can be done effectively. The iteration of the decomposition algorithm using filters H_0 and H_1 is illustrated in Figure 3-7.

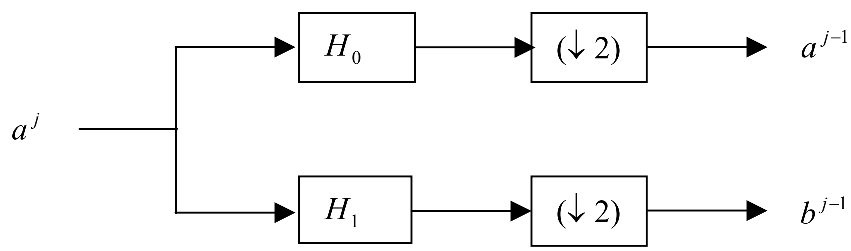


Figure 3-7. Decomposition diagram for a multiresolution analysis.

The impulse responses of low-pass filter H_0 and the high-pass filter H_1 are defined as

$$h_0 = \frac{1}{2}\overline{p_{-k}}\,, \tag{3-31}$$

$$h_1 = \frac{1}{2}(-1)^k p_{k+1} \tag{3-32}$$

where the coefficients p_k come from (3-19).

The iteration of the decomposition algorithm can now be formulated in a convolution form as

$$a^{j-1} = (\downarrow 2)\left(h_0 * a^j\right), \tag{3-33}$$

$$b^{j-1} = (\downarrow 2)\left(h_1 * a^j\right), \tag{3-34}$$

where the sequences a^j and the b^j are defined as

$$a^j = \left(\ldots, a^j_{-1}, a^j_0, a^j_1, \ldots\ \right), \tag{3-35}$$

$$b^j = \left(\ldots, b^j_{-1}, b^j_0, b^j_1, \ldots\ \right). \tag{3-36}$$

It is important to note that the discrete filters and downsampling operator do not depend on the level j.

The filtering process can be iterated on either or both subsequences. The signal can be broken down into many lower resolution components. The third step of the decomposition is *termination*, where the decomposition is finished. Basically the decomposition can be continued until there is a single sample in the highest signal part box. On the other hand, this is unnecessary in many cases. The choice of termination point usually depends on the purpose it is used for.

In this thesis the parts of the signal are called bins, and the direction of growth is downward. In Figure 3-8 the three level (N=3) wavelet decomposition tree, where the scheme is iterated on the lower band, is illustrated. The wavelet coefficients are marked with letter *c*. The approximation (low-pass) coefficients a_k^j correspond to A's and the detail (high-pass) coefficients b_k^j correspond to D's, when for example a^{j-1} corresponds to cA_1 and b^{j-1} corresponds to cD_1. The number in the lower index denotes the level of the wavelet decomposition. The highest signal part boxes are cA_3 and cD_3 in Figure 3-8.

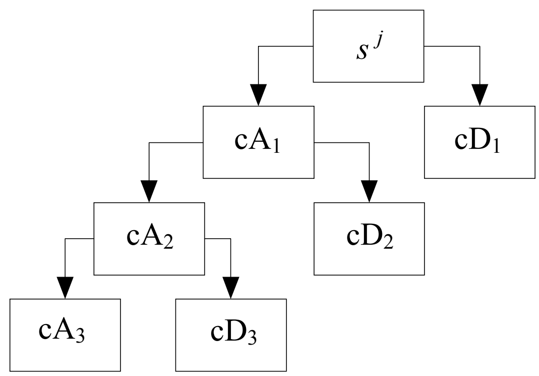


Figure 3-8. The wavelet decomposition tree.

In the WPD the detail part, D, is decomposed as well, as can be seen in Figure 3-9. This complete binary tree offers the richest analysis.

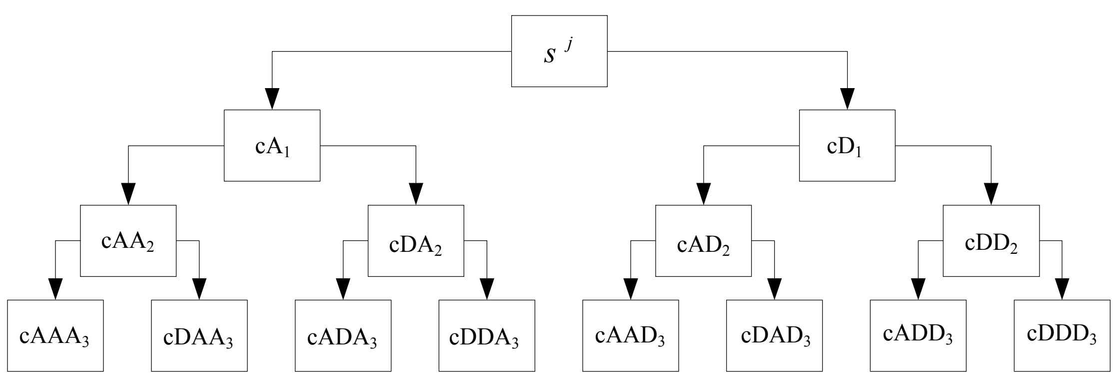


Figure 3-9. Three level wavelet packet decomposition tree.

In the WPD the number of bins in the highest level is 2^N, where N is the level of the decomposition. Hence, the frequency scale of the signal is broken down to 2^N parts. For example, in Figure 3-9 the bin $cAAA_3$ contains the wavelet coefficients of frequencies

$0 \ldots F_s / 16$ and the bin cDDD3 $7F_s / 16 \ldots F_s / 2$, where F_s denotes the sampling rate of the signal. Due to downsampling in each level, each bin vector contains half as many elements as its parent bin vector, so the downsampling by two halves the time resolution (Sheng 2000, p. 10-28). The number of the coefficients in each bin is $L/2^N$, where L is the length of the signal s.

Due to the downsampling, aliasing occurs in the WPD tree, and it can be seen exchanging the frequency ordering of some branches of the tree (Akansu and Haddad 1992, p. 128). In Figure 3-10 the WPD tree is presented in an increasing frequency order from the left to the right. The arrows depict the signal decomposition with low-pass (H_0) and high-pass (H_1) filters.

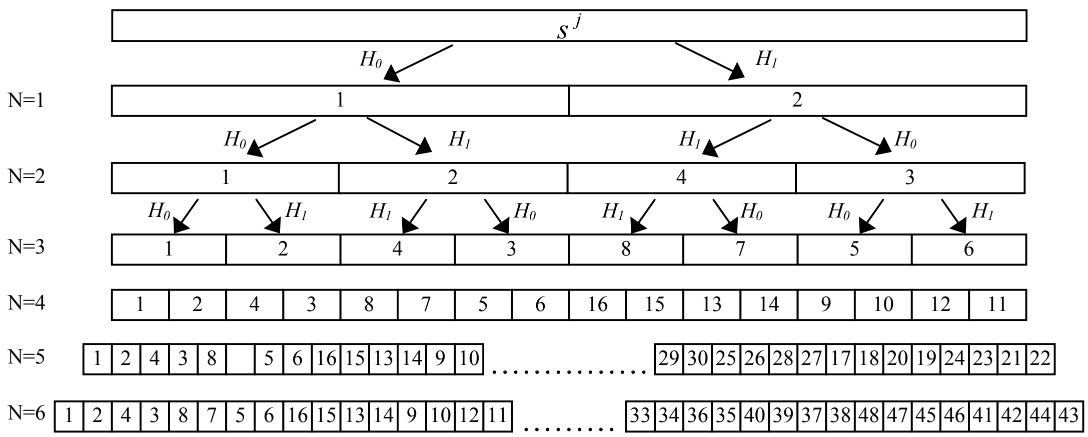


Figure 3-10. The six level WPD in increasing frequency order.

The wavelet transform is an efficient method for local analysis of nonstationary and transient wideband signals (Sheng 2000, p. 10-1). As the traditional Fourier analysis breaks down a signal into constituent sinusoids of different frequencies, the wavelet analysis uses the shifted and scaled versions of the original wavelet. Although the Fourier analysis is efficient and nowadays fast to calculate, the wavelet seems to be more efficient and useful especially with transient signals, such as inharmonic bird sounds. It is possible to represent complicated signals with a few coefficients with wavelets. That makes for example the task of classifiers easier. (Nason 1999)

Although the wavelet transform is effective and powerful for multiresolution local spectrum analysis, the dyadic wavelet transform unfortunately has some weaknesses. The values of the wavelet coefficients will change dramatically, if the input signal is shifted by one sampling interval distance. Hence, the wavelet transform is time dependent. The shifting of the signal and changing the coefficients, for example in noise reduction, often produce the aliased version of the signal, which is also a weakness of the wavelet transform. For example, the nonorthonormal quadratic spline wavelets have been used, because the orthonormal quatrature mirror filters have been found sensitive to the time shifting of the input. (Sheng 2000, p. 10-82)

4. CLASSIFICATION ALGORITHM

In Figure 4-1 the *classification model* of the whole recognition process is described. First the soundtracks were segmented into smaller pieces, which are called sounds in this thesis. Then the sounds were preprocessed, which included for example noise reduction. All sounds were decomposed to the wavelet coefficients with the WPD and after that the features were calculated from these coefficients. These feature vectors of the training data were introduced to the MLP and to the SOM network during the training session. Then both the networks were tested with the testing data and after that the classification results were examined.

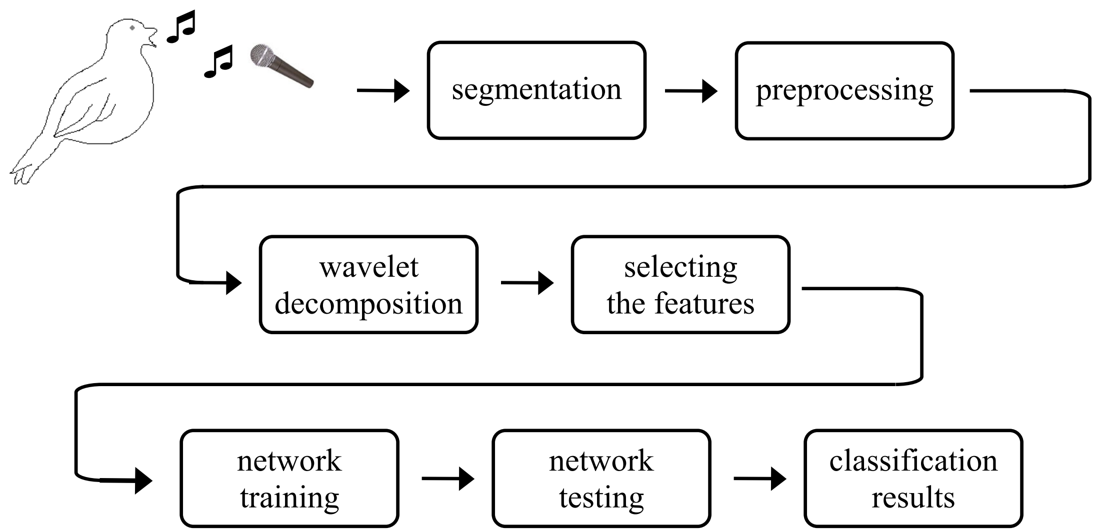


Figure 4-1. The classification model used in this thesis.

4.1 Segmentation and preprocessing

The sampling rate of the sound data F_s was 44.1 kHz and 16-bit accuracy was used. First the zero mean data was normalized in the range [-1, 1], and the low frequency wind noise was reduced using long moving average filter. In the segmentation phase, the soundtracks were extracted into smaller pieces. The noise level varied a lot between the sound tracks. Therefore the noise threshold level was calculated adaptively from long-term mean energy value. In Figure 4-2 an example of the segmentation of the track of the Quail sound is illustrated, and the red box in the Figure contains only one sound after successful segmentation. The onset of the sound exceeded the adaptive threshold level and the end of the sound dropped under that threshold value. Consequently, the one segment of data consisted only one sound. The sounds, which were recorded in a very noisy environment or which were in inseparable groups, were rejected.

A part of the soundtrack of the Quail

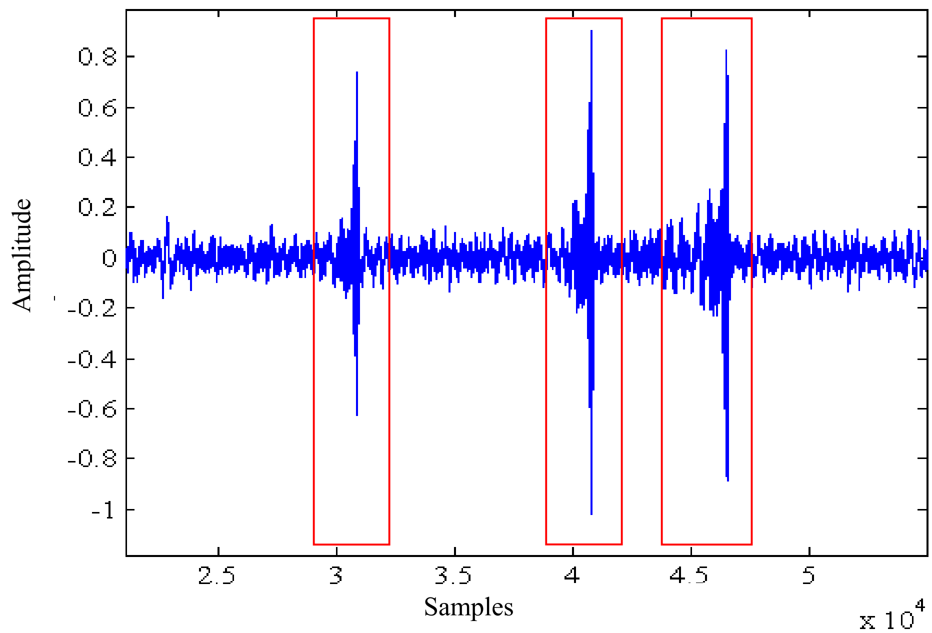


Figure 4-2. An example of the segmentation.

In the preprocessing phase the broadband noise was reduced from the sounds. This noise reduction was made with filter bank, and it is illustrated in Figure 4-3. The sound was split into eight separate frequency bands. After splitting the signal s to these frequency bands, the threshold value T_{h0} was calculated from the output of band number eight, because the information of the highest frequencies proved to be the least essential due to its noise-like contents. After preliminary tests the threshold value T_{h0} was defined as 2 times the standard deviation of the output s_8. The coefficients, of which amplitude dropped under the calculated threshold level T_{h0}, were set to zero in every band. The outputs from the thresholding blocks were then calculated as

$$\tilde{s}_i(n) = \begin{cases} 0 & \text{if } s_i(\text{n}) < \text{T}_{\text{h0}} \\ sgn\left(s_i\left(n\right)\right)\left(\left|s_i\left(n\right)\right| - T_{h0}\right) \quad \text{for } i = 1\ldots8 & \text{else} \end{cases} \tag{4-1}$$

After reducing the irrelevant noise information from each band, all eight frequency bands were added into one signal $\tilde{s}$. The sound signal contained now much less noise and the essential information of the bird sound was emphasized.

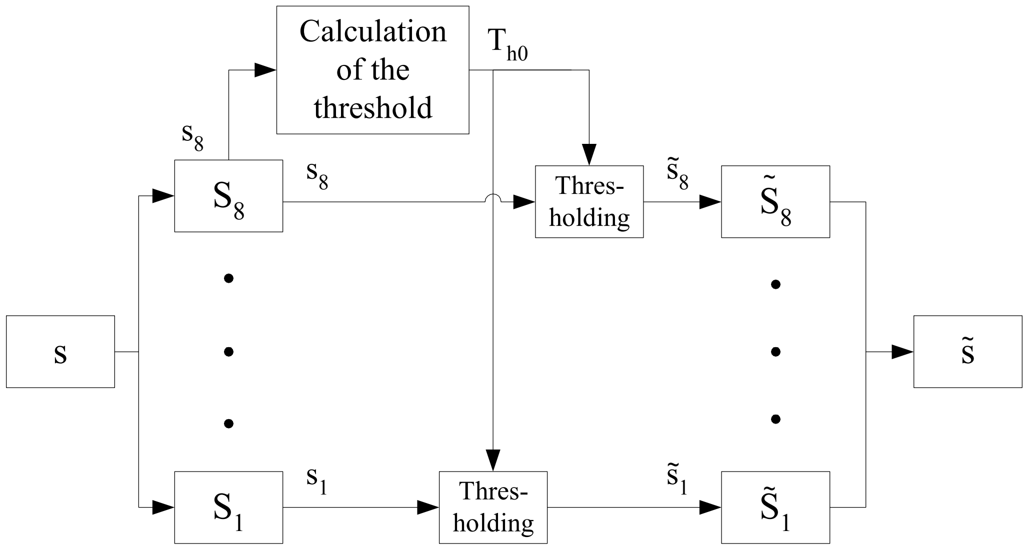


Figure 4-3. The noise reduction with filter bank.

4.2 Wavelet Packet Decomposition

In this study the wavelet packet analysis was used for the signal decomposition. In the *wavelet packet decomposition* (WPD) the signal s is split into approximation (A) and detail (D) parts. The symmetric wavelet decomposition tree, which was used in this thesis, is illustrated in Figure 4-4. For this case, after preliminary tests it turned out that the best decomposition level (N) was six. Thus, the signal s was split into $2^6 = 64$ bins.

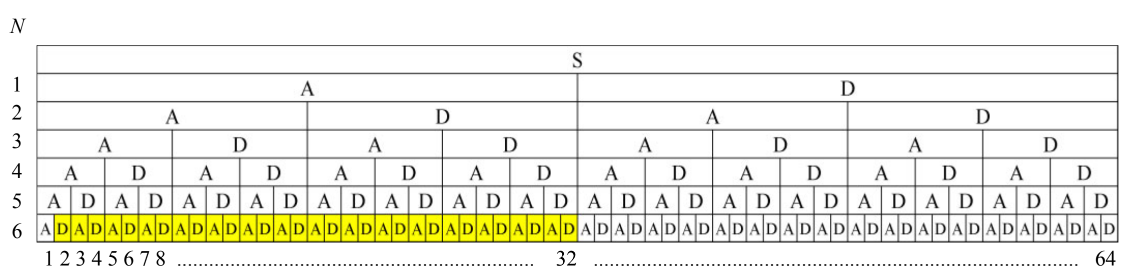


Figure 4-4. The WPD tree of this thesis.

The ordering of the WPD tree was made as in Figure 3-10. The bin number 1 contained so low frequencies that they proved to be irrelevant for the classification. The bins 33-64 also proved to be irrelevant; so over 11025 Hz ($F_s/4$) frequencies were not needed in the classification. Thus, the wavelet coefficients were calculated from bins 2-32, which are marked yellow in Figure 4-4. The number of the coefficients in each bin is $L/2^N$, where L is the length of the original signal s.

There are several wavelet families that have proven to be particularly useful, from which Daubechies, Coiflets, Symlets, discrete Meyer, Biorthogonal, and reverse Biorthogonal wavelets are suitable for the WPD (Misiti et al. 2000, pp. 6-75–6-76). The Daubechies wavelet family was selected, because the analysis with dbN's is orthogonal, and both scaling and wavelet function are compactly supported. In addition, the dbN's have other pleasant properties for example given number of vanishing moments and that they are implemented as FIR (Finite Impulse Response) filters. The only difficulty with dbN's is the poor regularity. However, the regularity increases with the order. In Figure 4-5 three WPD coefficient figures are illustrated, where Haar (db1), Daubechies 5 (db5) and 10 (db10) wavelets are used in the decomposition. In all cases the decomposition level N was six.

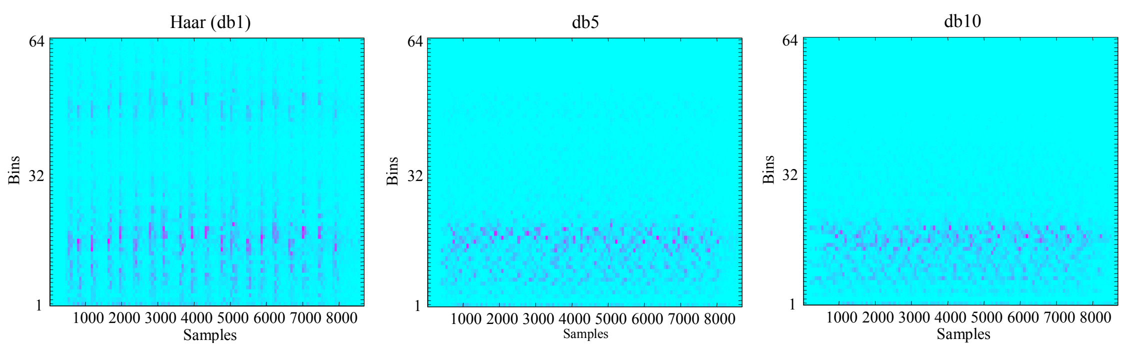


Figure 4-5. The WPD coefficient figures of example wavelet functions.

After preliminary tests, it turned out that the db10 wavelet function compromised the best decomposition results with selected bird sounds. That is why the db10 was selected for the wavelet function. The impulse responses $h_0(n)$ and $h_1(n)$ of two db10 FIR filters, low-pass H_0 and high-pass H_1 filters of the decomposition are illustrated in Figure 4-6. The length of the impulse responses is 2N, where N is the order of the wavelet function. In this study N is 10, therefore there are 20 coefficients in each filter. These filters are derived from the wavelet ψ and scaling ϕ functions as discussed in the Chapter 3.

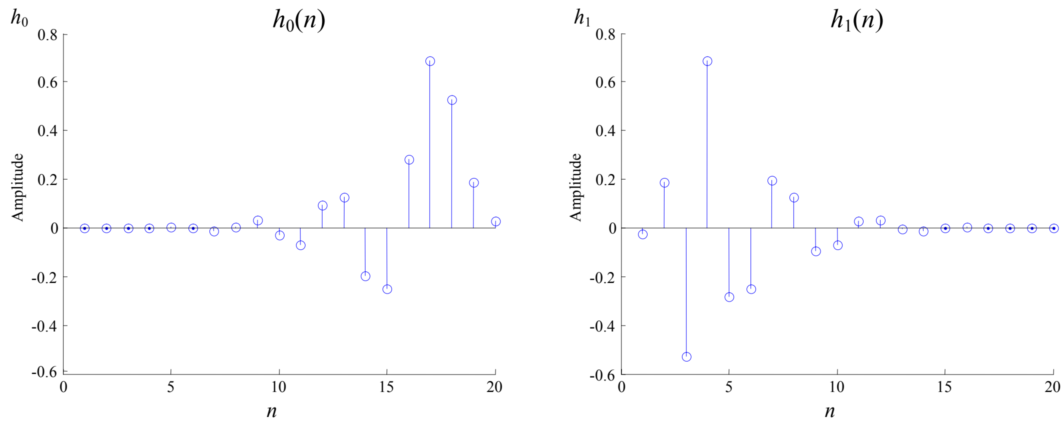


Figure 4-6. The impulse responses of the decomposition filters of the db10.

4.3 Features

After the WPD, the number of the coefficients of each bin was $L/64$, where L is the original length of the sound. There were too many coefficients per sound for the efficient pattern recognition, so the coefficient data of each sound was reduced to a feature vector of four real numbers. The four features are called *maximum energy*, *position*, *length*, and *width* in this thesis. These four features are illustrated in Figure 4-7.

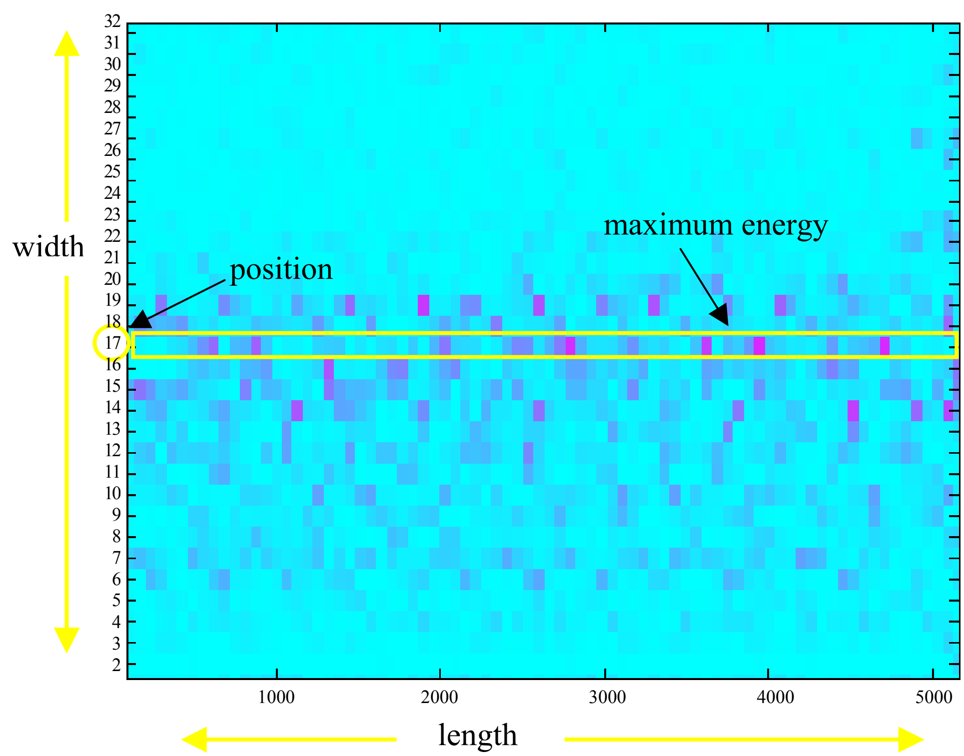


Figure 4-7. The four features: maximum energy, position, length, and width.

The average energy E_B of the wavelet coefficients (c) of each 31 bins was calculated as

$$E_B = \frac{\sum_{n=1}^{n_c} c^2[n]}{n_c} \tag{4-2}$$

(Oppenheim et al. 1997, p. 6), where n_c is the number of the coefficients of the bin. Then the largest energy value was chosen and it is called the *maximum energy* of the sound. The *position* represents the number of the bin, in which the maximum energy was located.

After preliminary test from all bins the threshold value T_{h1} was calculated as

$$T_{h1} = \frac{E_B}{6} \tag{4-3}$$

from the energy of each bin E_B. The number of those coefficients of each bin, which exceeded the threshold value T_{h1}, was added together and then it was averaged between all the bins that contained sound energy. This feature was called the *length*. The other threshold value T_{h2} was defined as 1.3 after preliminary test from all bins. If the equation

$$T_{h2} = 1.3 < \sum_{n=1}^{n_c} c^2[n] \tag{4-4}$$

was satisfied, those bins were added together. So, the *width* represents the number of those bins, whose sum of the energy exceeded the T_{h2}.

Finally all four features were normalized, in order to be comparable with each other. The normalization levels were defined after a preliminary test from all bins. The *maximum energy* E_m was normalized as

$$\widetilde{E_m} = \frac{E_m}{n_B} \tag{4-5}$$

where n_B is the number of the coefficients of the bin, where most coefficients were exceeding the T_{h1}. The *position* P was normalized as

$$\widetilde{P} = \frac{P}{2^N / 4} = \frac{P}{16}. \tag{4-6}$$

The *length l* was normalized

$$\tilde{l} = \frac{l}{100} \tag{4-7}$$

and the *width w* as

$$\tilde{w} = \frac{w}{20}. \tag{4-8}$$

The reason for the normalization was the SOM, where the recognition results are better, if the inputs are in the same scale. The training time of the SOM network is also shorter with normalized inputs.

4.4 Neural networks

In neural networks the information is distributed among many links that connect the simple processing units, neurons, together in the network. A network gains information in training by example as data sets are repeatedly presented to it. During this training period the network itself adjusts its links in order to retain information. A network is trained in this manner until it is able to function at an acceptable level of performance.

A *neuron* is a fundamental information-processing unit of the neural network. There are three basic elements in the model of neuron, which are illustrated in Figure 4-8.

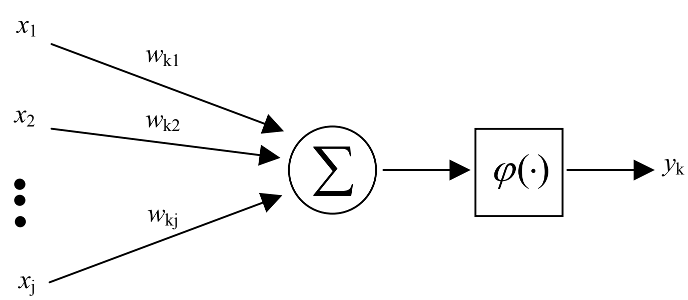


Figure 4-8. The basic elements of the model of neuron k.

Each of the connecting links or synapses is characterized by a weight of its own, which means that a signal x_j at the input of synapse j connected to neuron k is multiplied by the synaptic weight w_{kj}. Then the adder sums the weighted input signals and eventually the activation or logistic function $\varphi(\cdot)$ limits the output y_k of a neuron. In the MLP network the logistic function must be differentiable, so the sigmoidal functions are often used. (Haykin 1994, pp. 8, 148) A typical representative of the logistic function is defined as

$$\varphi(x) = \frac{1}{1+e^{-ax}} \tag{4-9}$$

where a is a slope parameter (Theodoridis and Koutroumbas 1999, p. 92).

In this thesis two well-known neural networks, the MLP and the SOM, were used for the classification. The MLP is based on supervised learning and the SOM on unsupervised learning. When using neural networks in the pattern classification, there has to be a fixed number of classes into which patterns are to be classified.

In supervised learning the network has to undergo the training session first, which means that a set of input patterns along with the known class is repeatedly presented to the network. Then in the testing session a new pattern is presented to the network and the task is to classify this new pattern correctly. The MLP consists of the input layer, one or more hidden layers and an output layer. In the MLP training the error back-propagation algorithm is used, which consists a forward pass or function signal and backward pass or error signal. The synaptic weights of the network are all fixed during the function signal, and they are all adjusted after the error-correction rule during the error signal. (Haykin 1994, p. 138) In Figure 4-9 two signal flows of the MLP are illustrated.

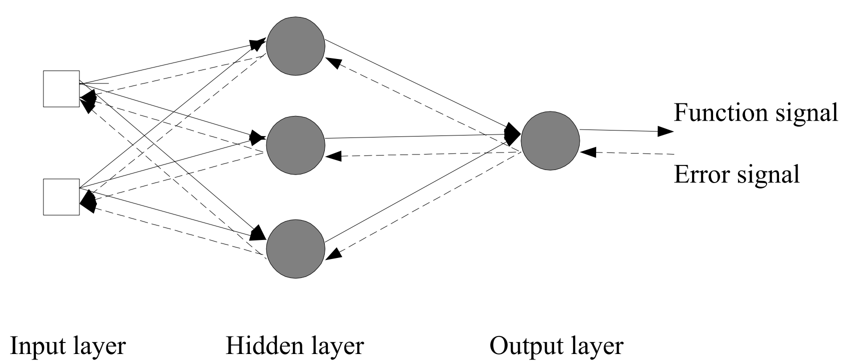


Figure 4-9. The directions of two basic signal flows of the MLP.

In back-propagation the main goal is that the neural network will be able to classify novel patterns correctly with respect to the training patterns. In that case the network will generalize well. Hence, the generalization means that the network will produce a correct input-output mapping even when the input is slightly different from the examples used to train the network. The network is overtrained, if it learns too many specific input-output relations. The overtraining should be avoided. (Haykin 1994, pp. 177-178) In Figure 4-10 good generalization and bad generalization (overtraining) are illustrated.

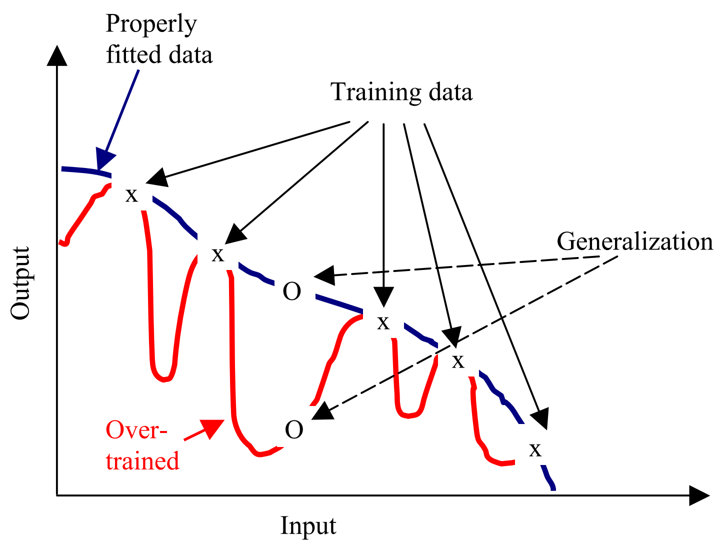


Figure 4-10. Good generalization in blue and poor generalization in red.

Figure 4-11 shows the structure of the MLP used in this study. This chosen structure was selected as a result of preliminary tests. The testing was started with a very small network. The nodes were added one by one until the good generalization was reached, and at the same time it was checked that the network was not overtrained. The overtraining is in direct relation to the results of the network. When the results of the unknown test data get worse, the network is overtrained and poor generalization exists. After the training, it was examined that all the used nodes, and the weighting and bias parameters were needed, which means that none of the outputs of the nodes was too close to zero.

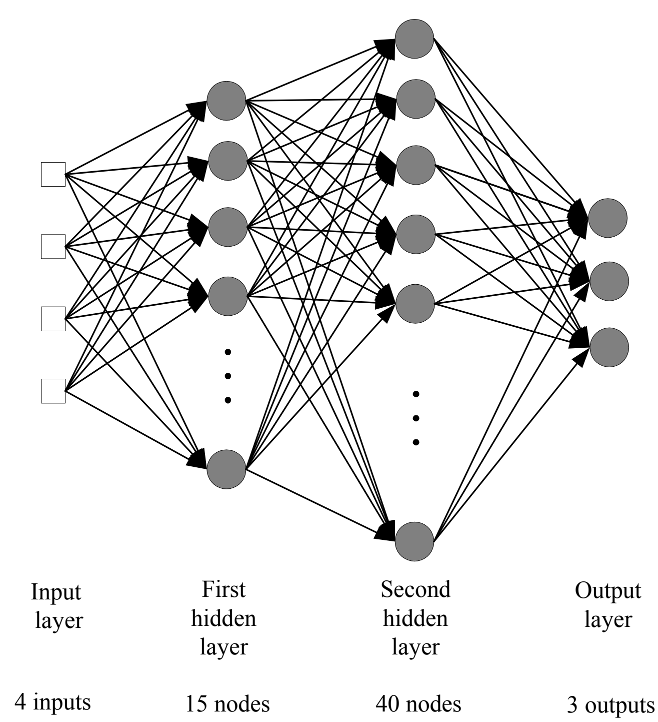


Figure 4-11. The MLP architecture used in this thesis.

The four features (maximum energy, position, length, width) described in the Section 4.3 were used as inputs at the input layer. The first hidden layer had 15 and the second hidden layer had 40 nodes. Between the input and the first hidden layer, and the first and the second hidden layer hyperbolic tangent sigmoid function (tansig) was used as the activation function. The range of the tansig activation function is [-1, 1] and it is defined as

$$\varphi_{\text{tansig}}(x) = \frac{2}{(1+e^{-2x})-1} \tag{4-10}$$

Finally, in the output layer three outputs were used, and log sigmoid function (logsig) as activation function was used between the second hidden layer and the output layer. The range of the logsig function is [0, 1] and it is defined as

$$\varphi_{\text{logsig}}(x) = \frac{1}{1+e^{-x}} \tag{4-11}$$

In Figure 4-12 the tansig and logsig activation functions are illustrated.

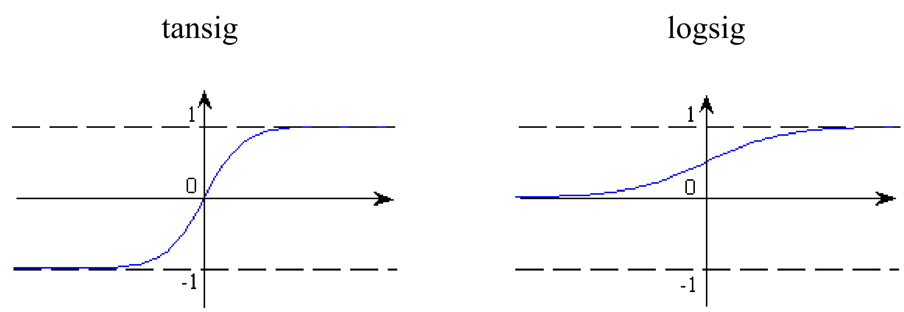


Figure 4-12. The tansig and logsig activation functions.

After that each output was rounded to 0 or 1. Thus, it was possible to present decimal numbers 0-7 with three output bits, and that was enough for classes of eight bird sounds. Finally, the binary output was converted into numbers between 1-8.

The MLP network was trained for up to 65 epochs and the mean square error (MSE) goal was 0.0001. After the training the network was tested with the testing data.

The second network, which was used in the experiments, was the SOM. It is a clustering and visualization tool, which enables the organization of the database in an unsupervised manner (Kohonen 2001, p. 105). In unsupervised learning there is no prior knowledge of the categories into which the activations are to be classified (Haykin 1994, p. 67). In Figure 4-13 one example of the model of the SOM is illustrated. The SOM consists of a one-dimensional input layer and a two-dimensional competitive layer. The idea is to project and visualize high-dimensional signal spaces on such a two-dimensional display (Kohonen 2001, p. 105). Hence, the output neurons are usually arranged in a two-dimensional lattice that ensures that each neuron has a set of neighbours. Each neuron in the competitive layer holds a weight reference vector that after training resembles a different input pattern.

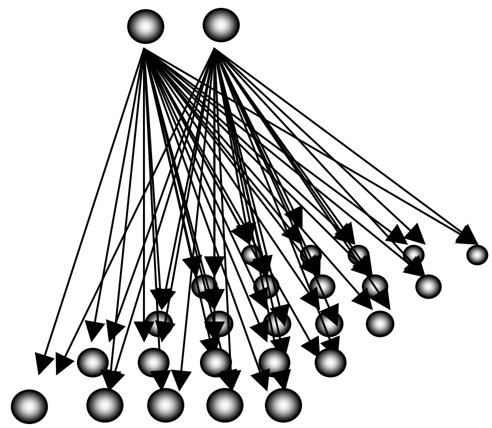


Figure 4-13. The example of the model of the SOM. Both input nodes of the input layer are connected to each neuron of the 5 x 5 output layer.

The construction of the SOM is based on competitive learning and the use of neighbourhood when adapting the model. The input pattern vectors are presented to all nodes in parallel. In the competitive learning the output neurons of the network compete among themselves to be activated. Only one output neuron is on at any one time after the competition, so the winner is called *winner-takes-all* neuron. (Haykin 1994, p. 397) The individual neurons of the network become feature detectors, because they learn to specialise in sets of similar patterns. The synaptic weights w_{ji} connect input node i to neuron j. A fixed amount of synaptic weight is allocated each neuron. That means the synaptic weight is distributed among its input nodes as

$$\Delta w_{ji} = \begin{cases} \eta(s_i - w_{ji}), & \text{if neuron } j \text{ wins the competition} \\ 0, & \text{if neuron } j \text{ loses the competition} \end{cases} \tag{4-12}$$

where η is the learning-rate parameter and s_i is the i:th component of the input signal. All neurons of the network are constrained to have same Euclidean length or norm by requiring that

$$\sum_j w_{ji}^2 = 1, \quad \text{for all } j \tag{4-13}$$

By moving its synaptic weight vector to center of gravity of the discovered cluster each of the output neurons has discovered a cluster of inputs. (Haykin 1994, p. 54-55)

In this thesis the SOM network was trained using the feature vectors (maximum energy, position, length, width) as inputs. These feature vectors were introduced to 10 x 10 size SOM. The other sizes, for example 6 x 6, 8 x 8, and 12 x 12, of the network were also tested, but the chosen size presented best classification results.

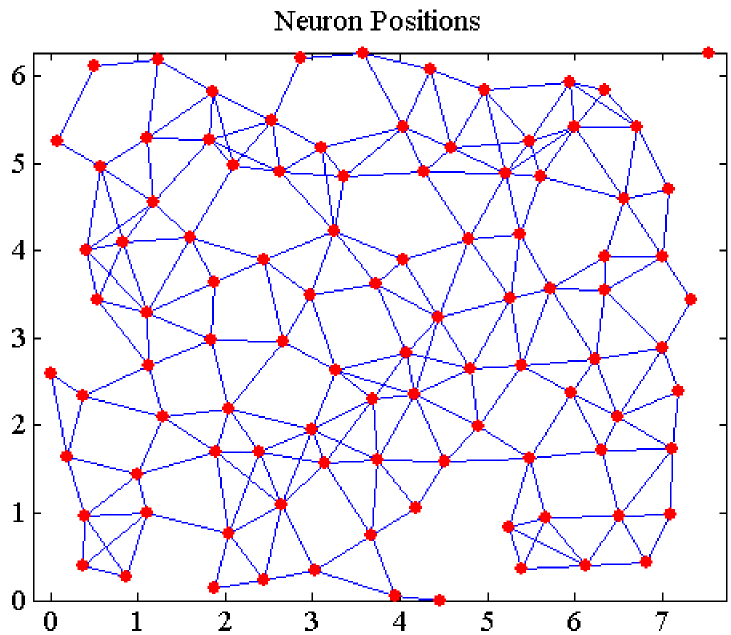


Figure 4-14. A 10 x 10 set of neurons after the training of the SOM.

When initialising the SOM, the random layer topology function (randtop) was used. That specified the topology for the original neuron locations. The training data was introduced to the SOM network 3000 times. Other number of epochs was also tested, but the results did not improve. Figure 4-14 illustrates the locations of the 100 nodes in the 2-D plot. The values of the axes are suggestive for the purpose of visualisation.

5. TEST DATA AND RESULTS

This Chapter focuses on the data and the results of this thesis. The data was analysed in MATLAB environment, and the Wavelet Toolbox (Misiti et al. 2000, p. 1-33) was utilised. In Section 5.1 it is clarified how the data was selected, what were the sources of the data and how the data can be interpreted. The results of both networks, MLP and SOM, are summarized in Section 5.2.

5.1 Data

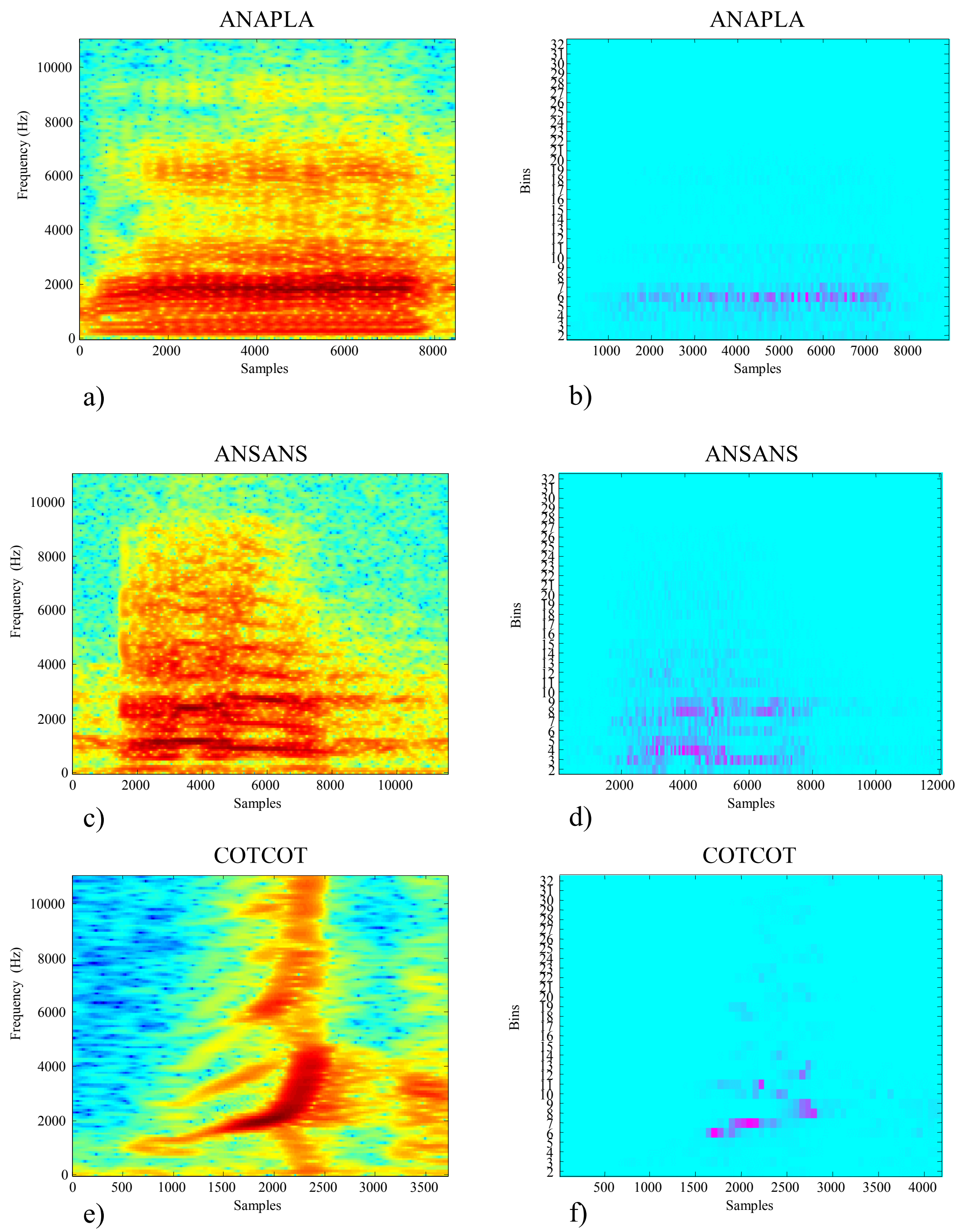
In this thesis the main purpose was to study the automatic classification of the inharmonic or transient bird sounds. This is the reason why the sounds of the Mallard, the-Greylag Goose, the Corncrake, the River Warbler and the Magpie were selected. The pure tonal territorial song of the male Pygmy Owl was chosen as a reference sound. The sounds of the Quail and the Spotted Crake are tonal, but contain some transient features. In addition, recordings of those eight species were available quite easily and the quality of the recordings was sufficient. Table 5-1 illustrates the selected eight species set of bird sounds. The table contains scientific abbreviations, English names and sound types. Also the number of the sounds in the training and testing is indicated.

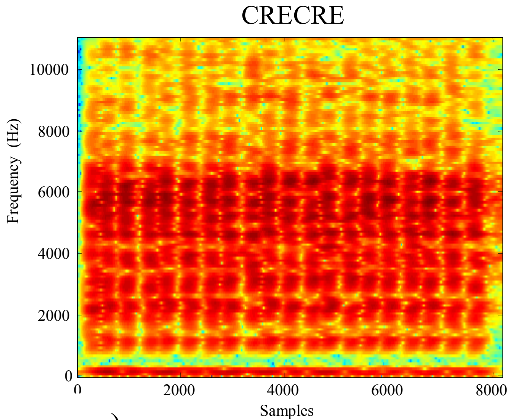
Table 5-1. Selected set of bird sounds used in this thesis

Scientific Abbr.	English name	Sound type	MLP Training	SOM Training	Testing
ANAPLA	Mallard	inharmonic	138	113	60
ANSANS	Greylag Goose	inharmonic	135	113	59
COTCOT	Quail	tonal	190	113	83
CRECRE	Corncrake	inharmonic	443	113	110
GLAPAS	Pygmy Owl	pure harmonic	113	113	48
LOCFLU	River Warbler	inharmonic	890	113	328
PICPIC	Magpie	inharmonic	203	113	97
PORPOR	Spotted Crake	tonal	166	113	69
			2278	904	854

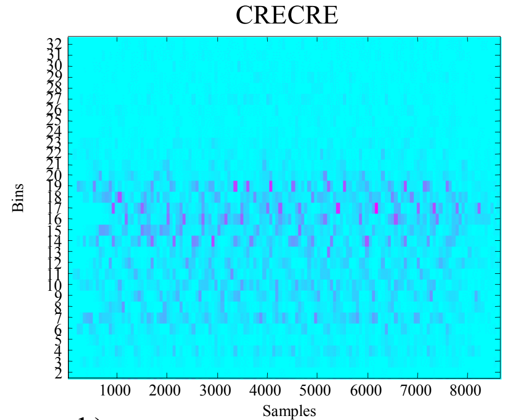
In this study data from six sources was used. The species, recorders, references, and number of the sounds and tracks can be found in Appendix 1. Most of the sounds are from Pertti Kalinainen (2004). The other sounds were from Schulze (2003), Saurola (1995), Heiskanen (1993), Heiskanen (2004) and Roche (2001). There were totally 3132 sounds, which were divided into training data, 2278 sounds, and testing data, 854 sounds. The training and testing data were from different tracks. It turned out that if there were the same number of training data, the SOM network functioned better. So, in the training of the SOM network the training data was reduced to 904. The sampling rate F_s of the data was 44.1 kHz.

The typical spectrograms and corresponding wavelet coefficient figures of eight species that are used in this thesis are presented in Figure 5-1a-p. The frequency and bins are bounded to 11025 Hz ($F_s/4$), because at the higher frequencies there was no essential information. In the spectrograms the darker colours represent the higher energies of the sound. Correspondingly, the larger absolute values of the coefficient are presented with the pinker colour in the adjacent wavelet coefficient figures. The range of the coefficients is [-5, 5].

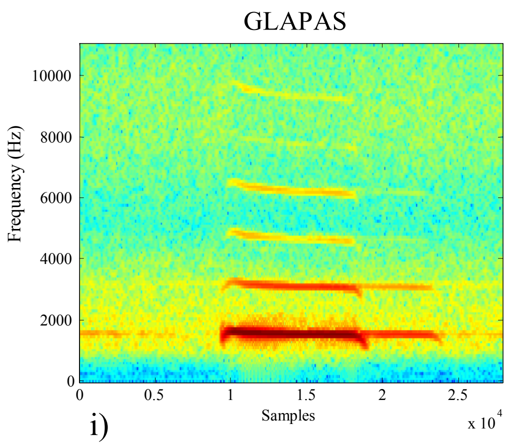



CRECRE
Frequency (Hz)
Samples

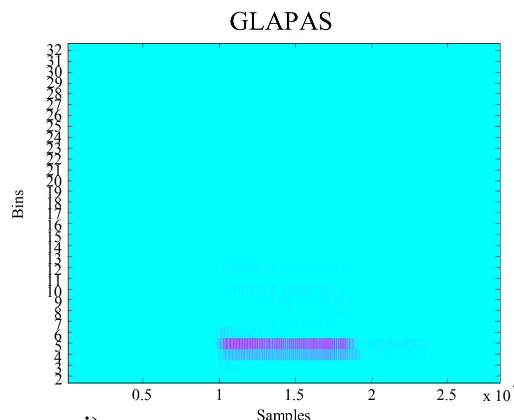
g)


CRECRE
Bins
Samples

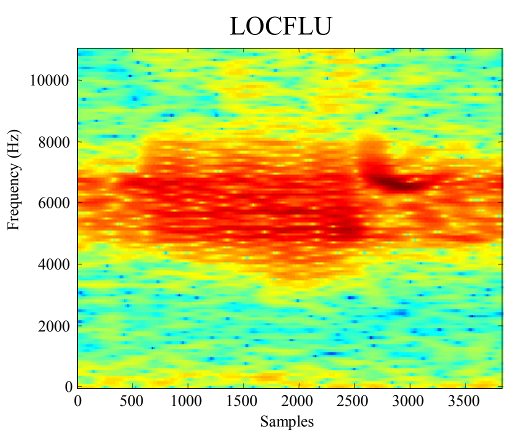
h)


GLAPAS
Frequency (Hz)
Samples
x 10^4

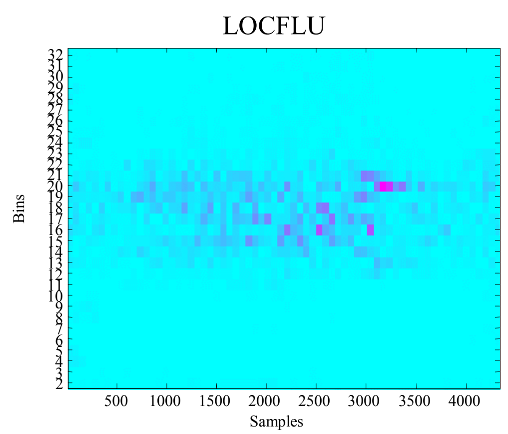
i)


GLAPAS
Bins
Samples
x 10^4

j)


LOCFLU
Frequency (Hz)
Samples

k)


LOCFLU
Bins
Samples

l)

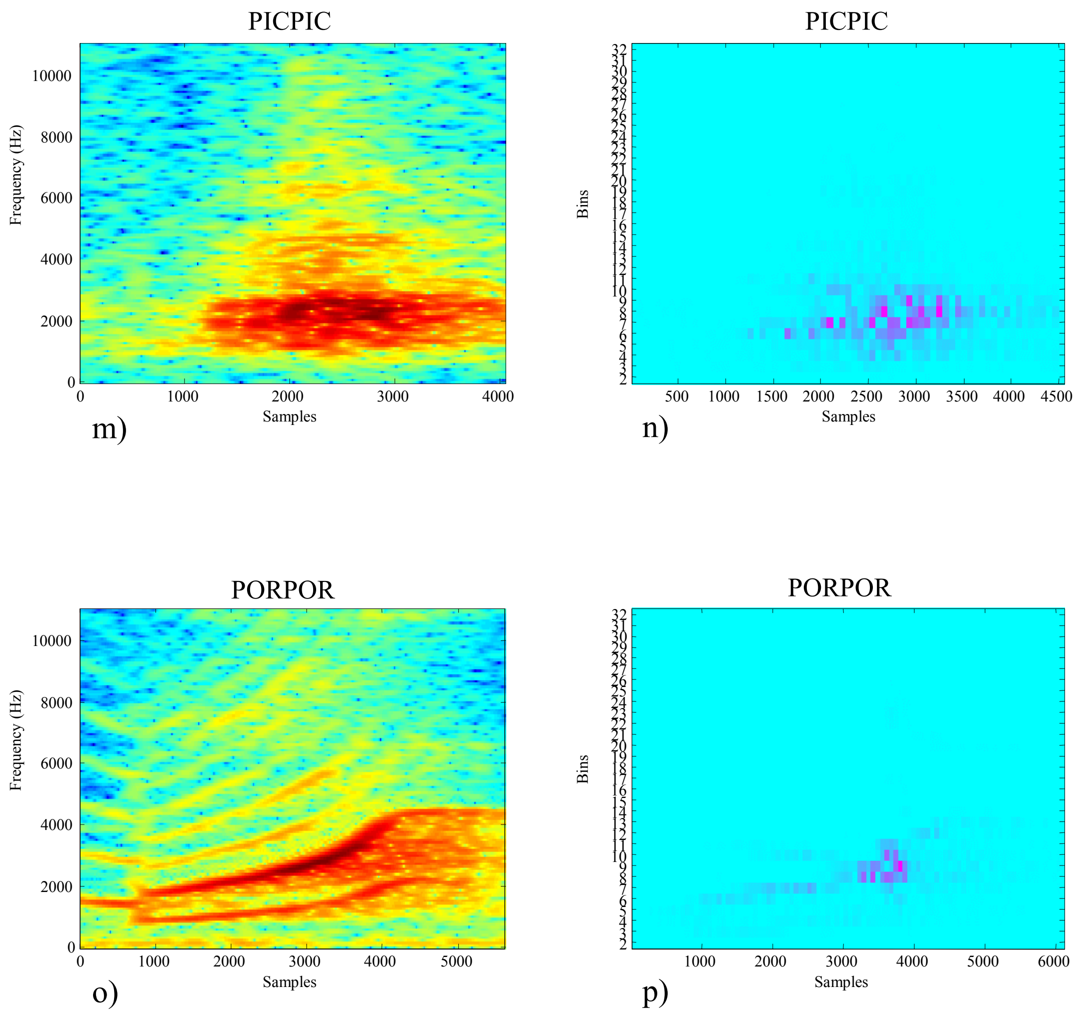


Figure 5-1a-p. Typical spectrograms and corresponding wavelet coefficients of the eight species used in this thesis. The frequency and bins are bounded to 11025 Hz, because at higher frequencies there was no essential information.

As it can be seen from Figure 5-1a-p, the wavelet transform compresses the energy of the coefficients more than traditional Fourier transform in spectrograms. Only the very essential information is preserved after the wavelet transform.

5.2 Results

The four features, *maximum energy*, *position*, *length*, and *width*, were calculated from the wavelet packet coefficients. The features can be visualized in groups of three features in $\mathbb{R}^3$. Hence, four feature spaces are illustrated in Figures 5-2 to 5-5. Each feature was normalized (see Section 4.3), so that they were comparable with each other. Consequently, they are pure numerical values and have no quantity. The eight colours correspond to the eight bird species used in this study.

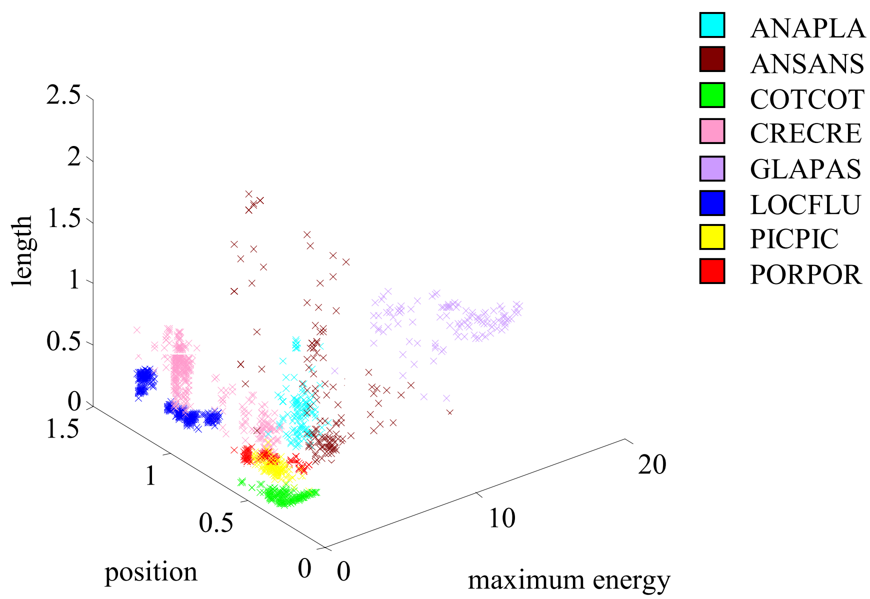


Figure 5-2. The feature space in terms of maximum energy, position and length.

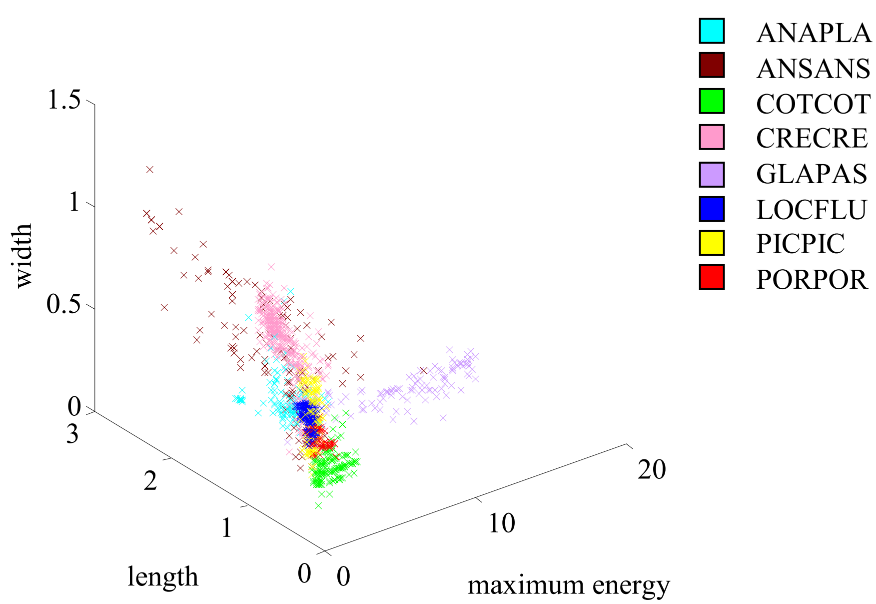


Figure 5-3. The feature space in terms of maximum energy, length, and width.

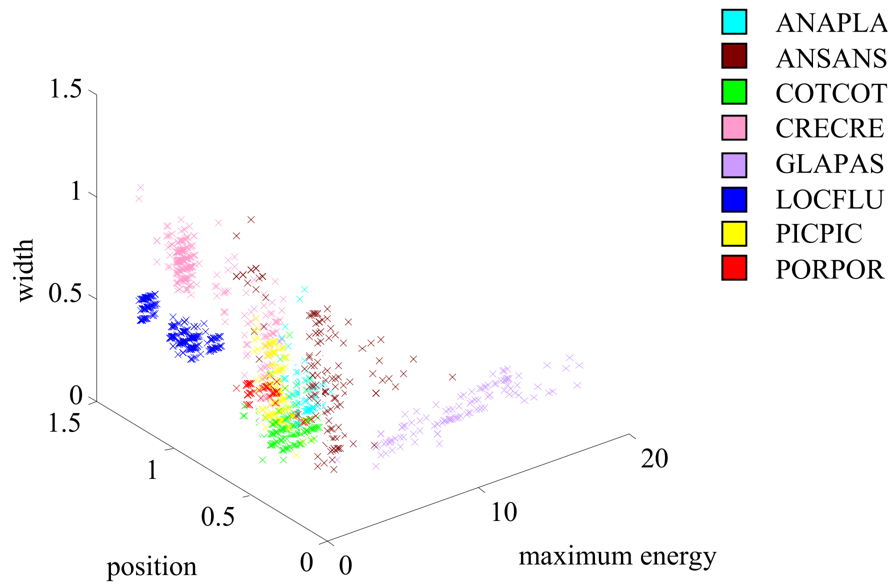


Figure 5-4. The feature space in terms of maximum energy, position, and width.

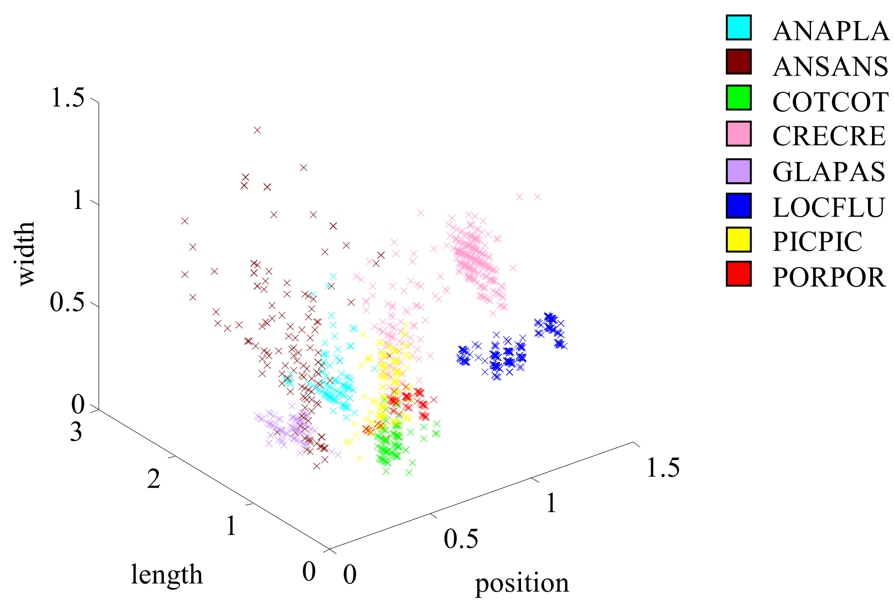


Figure 5-5. The feature space in terms of position, length and width.

Figures 5-2 to 5-5 illustrate that distinct clustering exists between the eight bird species. The final classification was made with the neural networks using the four features. Table 5-2 contains the classification result of the MLP network. The rows of the confusion matrix show how each species is classified. All the test sounds of the Quail (COTCOT) and the Spotted Crake (PORPOR) were recognized correctly. Twenty-four sounds of all the test sounds were classified wrongly.

Table 5-2. The confusion matrix of the testing data when using the MLP.

	ANAPLA	ANSANS	COTCOT	CRECRE	GLAPAS	LOCFLU	PICPIC	PORPOR	Number of the sound in testing
ANAPLA	**59**	1	0	0	0	0	0	0	**60**
ANSANS	1	**49**	1	3	1	3	1	0	**59**
COTCOT	0	0	**83**	0	0	0	0	0	**83**
CRECRE	1	2	0	**106**	0	0	1	0	**110**
GLAPAS	0	1	0	0	**46**	1	0	0	**48**
LOCFLU	0	1	0	0	0	**327**	0	0	**328**
PICPIC	0	0	5	1	0	0	**91**	0	**97**
PORPOR	0	0	0	0	0	0	0	**69**	**69**
									854

Table 5-3 illustrates the classification result of the MLP network in percentage terms. Altogether 96 % of the test sounds of the eight bird species were recognized correctly with the MLP network.

Table 5-3. The confusion matrix in percentage when using the MLP.

%	ANAPLA	ANSANS	COTCOT	CRECRE	GLAPAS	LOCFLU	PICPIC	PORPOR
ANAPLA	**98**	2	0	0	0	0	0	0
ANSANS	1.7	**83**	1.7	5.1	1.7	5.1	1.7	0
COTCOT	0	0	**100**	0	0	0	0	0
CRECRE	1	2	0	**96**	0	0	1	0
GLAPAS	0	2	0	0	**96**	2	0	0
LOCFLU	0	0.3	0	0	0	**99.7**	0	0
PICPIC	0	0	5	1	0	0	**94**	0
PORPOR	0	0	0	0	0	0	0	**100**

The classification result with the SOM is illustrated in Figure 5-6. The coloured areas present how each test species are situated in the 10 x 10 SOM network after the overlapping nodes have been analysed. The species who had most sounds in a particular node won and the possible other sounds were classified as outliers. If two or more different species had a same number of sounds in the particular node, all were classified as outliers. Appendix 2 presents the number and location of all the test sounds in the 10 x 10 SOM network. The outliers and their handling process are discussed in Chapter 6.

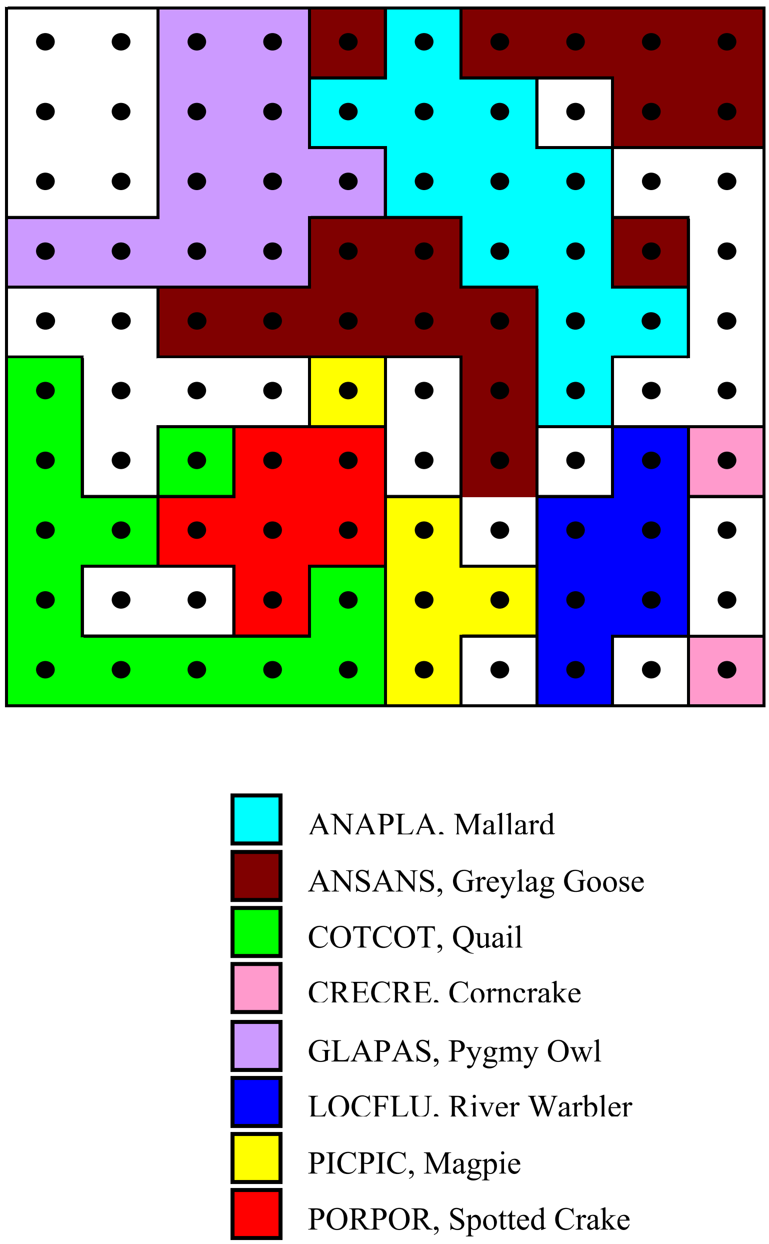


Figure 5-6. The classification result of the 10 x 10 SOM network

The confusion matrix of Table 5-4 illustrates the classification result of the SOM network. All test sounds of the Pygmy Owl (GLAPAS) and the River Warbler (LOCFLU) were recognized correctly, as can be seen from the diagonal of the matrix.

Table 5-4. The confusion matrix of the testing when using the SOM.

	ANAPLA	ANSANS	COTCOT	CRECRE	GLAPAS	LOCFLU	PICPIC	PORPOR	Testing
ANAPLA	**50**	10	0	0	0	0	0	0	**60**
ANSANS	12	**45**	0	1	0	0	1	0	**59**
COTCOT	0	0	**81**	0	0	0	2	0	**83**
CRECRE	0	0	0	**109**	0	0	1	0	**110**
GLAPAS	0	0	0	0	**48**	0	0	0	**48**
LOCFLU	0	0	0	0	0	**328**	0	0	**328**
PICPIC	1	0	2	0	0	0	**94**	0	**97**
PORPOR	0	0	1	0	0	0	1	**67**	**69**
									854

Table 5-5 illustrates the classification result of the SOM network in percentage terms. Totally 93.8 % of the test sounds was recognized correctly with the SOM network, which means that 32 sounds of all the test sounds were classified wrongly.

Table 5-5. The confusion matrix in percentage terms when using the SOM.

%	ANAPLA	ANSANS	COTCOT	CRECRE	GLAPAS	LOCFLU	PICPIC	PORPOR
ANAPLA	**83**	17	0	0	0	0	0	0
ANSANS	20	**76**	0	2	0	0	2	0
COTCOT	0	0	**98**	0	0	0	2	0
CRECRE	0	0	0	**99**	0	0	1	0
GLAPAS	0	0	0	0	**100**	0	0	0
LOCFLU	0	0	0	0	0	**100**	0	0
PICPIC	1	0	2	0	0	0	**97**	0
PORPOR	0	0	1.5	0	0	0	1.5	**97**

6. CONCLUSIONS AND DISCUSSION

The aim of this thesis was to study, how inharmonic and transient bird sounds can be recognized and classified efficiently. Many technologies, which are developed for the analysis of speech and musical signals, can be applied to bird sounds. However, there are important differences between speech or music and bird sound. One fundamental difference is the rate of sound events in bird sound, which is often much higher than in speech. For this reason the wavelet packet transform was applied in this study. The wavelet transform has been used in many applications, for example in noise reduction and image compression, but very few studies, in which the wavelet transform has been used in the context of automatic classification, can be found

Segmentation plays an important role in the pattern recognition and also in the sound classification. Incorrectly segmented sounds will probably be classified wrongly. Segmentation emerged as a very essential part of the whole classification process also in this study, where the semiautomatic segmentation method was used. Although the segmentation would be automatic, which has proven to be difficult, it is the most time consuming part of the whole recognition process. All segments of sounds must be checked before calculating the features. In many cases the manual checking is the safest way and it takes naturally time. Noise reduction goes hand in hand with successful segmentation. The segmentation is even more difficult, if the sound tracks are very noisy. Hence, efficient noise reduction is also a very crucial part of the preprocessing. In the future work, the possibilities of efficient noise reduction and automatic segmentation will be examined.

The selection of the wavelet function and the decomposition level are the most important phases of the WPD. In this study the db10 was selected for the wavelet function after preliminary testing, because there is no reliable algorithm for selecting the wavelet function properly. It is possible to calculate the best decomposition tree for the signal with MATLAB software. However, the best tree turned out to be asymmetric and unique for every bird sound. That is why the symmetric tree and the bins of the highest level were used. The level of the decomposition was selected to be six after preliminary testing. In the future work, effective algorithms will be developed for the selection of the wavelet function and of the decomposition level.

The wavelet analysis has many advantages, for example its ability to find out both frequency and temporal information, and to analyse signals which contain discontinuities and sharp spikes. These properties are appropriate for inharmonic and transient bird sounds. However, the wavelet transform is time variant, which is its main disadvantage in bird sound recognition. To avoid this problem, the time independent features were used in this thesis. The shifting of the signal and the changing of the coefficients causes often the aliasing, which is a weakness of the WPT. This property hinders particularly the reconstruction of a signal, for example in the noise reduction with wavelets. Anyway, the aliasing was not a problem in this study, because the signals were only decomposed with wavelet transform and so the reconstruction part was not used.

The selection of feature vectors has a strong effect to the pattern classification results. In this thesis four features were used. The features were calculated from the wavelet packet decomposition coefficients. Many kinds of other features were calculated from the coefficients and they were also tested. However, the chosen four features, *maximum energy*, *position*, *length*, and *width*, proved to describe and separate best the sounds of the eight bird species. In the future work, the possibilities of the statistical analysis tools, as variance, standard deviation and moment, will be examined so that the results could be improved and the selection of bird species could be increased.

The data of the eight bird species that was used in this thesis was divided so that there were about 70 % training data and 30 % testing data. This kind of division is common when the neural networks are used. Both networks, the MLP and the SOM were first trained and then tested. The training data of the SOM network was reduced, because it was observed that the SOM network functions best when there is approximately the same number of sounds from each group. In the beginning, the number of sounds of the River Warbler was 1218, which was many times more than what the other seven species had. In that case, the River Warbler conquered proportionally most nodes of the SOM network. The training data of the SOM was reduced randomly so that all the eight species had the same number of training data, and then the SOM was retrained. The presented results of this study are very encouraging and they proved that it is possible to recognize bird sounds with only four features with the MLP and the SOM networks.

The training data contained very probably sounds of seven Mallard, nine Graylag Goose, three Quail, eight Corncrake, five Pygmy Owl, two River Warbler, six Magpie, and three Spotted Crake individuals. The testing data was selected from different tracks than the training data (see Appendix 1). Hence, the testing data was also very probably from different individuals. So, in the testing data there were sounds of two Mallard individuals, four Graylag Goose, two Quail, Corncrake, and Pygmy Owl individuals, and one River Warbler, Magpie, and Spotted Crake individuals. Most of the sounds were taken from (Kalinainen 2004), and almost all the other from (Heiskanen 1993), (Schulze 2003) and (Heiskanen 2004). In the future work, more data is needed in order to generalize the results with more individuals.

Besides the classification of different bird species, also discrimination of the individuals among the species by their sounds is a very interesting topic. For example, Terry and McGregor (2002) have studied the call of the Corncrake (*Crex crex*) and showed that Corncrakes have individually distinctive vocalizations. In this thesis there were for example test sounds of two Corncrake individuals. Conspicuously, two nodes, (7,10) and (10,10), had been activated for Corncrake in the result of the SOM network (see Figure 5-6 and Appendix 1). Instantly, it roused a suspicion, whether the SOM had discriminated those two individuals. When listening to these sounds of two Corncrake individuals, they appeared to sound different. However, after examining these mentioned nodes, it turned out that the SOM had not separated these two individuals correctly. All sixty-nine sounds of the other individual, recorded by Heiskanen, situated in the node (10,10). Twenty-two of the other individual, recorded by Kalinainen, situated in the same node (10,10) and the rest eighteen sounds situated in the node (7,10). It is a very interesting question, why the SOM has classified these sounds in this way. One reason might be that the selected features were inadequate for the classification of individuals. On the other hand, the purpose of this thesis was the classification of the bird species. However, there exist references (Tanttu et al. 2005) that it might be possible to identify groups of birds and even individuals by their sounds with the SOM network. In the future work, our purpose is to study what kind of features are needed for the identification of the individuals, and if it is possible to recognize the bird species and to identify individuals simultaneously and how it can be done.

After the MLP network testing session, all wrongly classified sounds were manually examined and labelled. It turned out that 24 sounds were classified wrongly: one sound of the Mallard, ten sounds of the Greylag Goose, four sounds of the Corncrake, two sounds of the Pygmy Owl, one sound of the River Warbler, and six sounds of the Magpie. After plotting and examining all wavelet packet coefficients figures the reason for misrecognition became obvious. Figure 5-2 illustrates the wavelet packet coefficients of one misrecognized Greylag Goose and the two references, the Mallard and the Greylag Goose. This particular Greylag Goose (in the down middle) was classified as the Mallard.

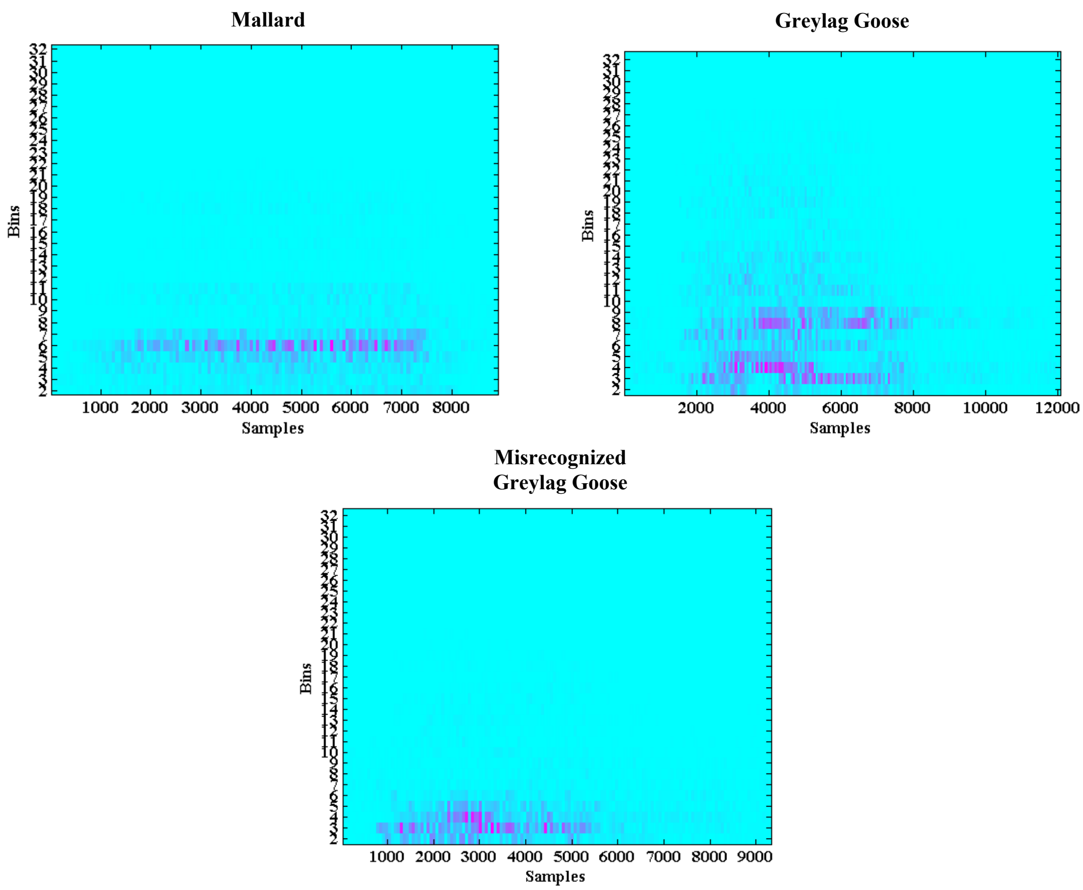


Figure 5-2. The wavelet packet coefficients of one misrecognized (in the down middle), the reference Mallard (on the up left) and the reference Greylag Goose (on the up right).

Because the features were reduced from the wavelet packet coefficients in this thesis, the shape of the coefficient pattern of the highest level (level 6) of the wavelet packet tree was very significant. As it can be seen from Figure 5-2, the coefficient pattern of the outlier Greylag Goose resembles the Mallard more than the Greylag Goose. Similar coefficient pattern error might be the reason for the misrecognition in 18 other cases. The wrong recognition was presumably caused by the bit pattern error in six cases. One bit might have flipped due to false segmentation or poor sound quality. Hence, that can be the reason, why for example three sounds of Greylag Goose (bit pattern 001) were recognized as River Warblers (bit pattern 101).

Also the SOM network was checked after the testing session. The SOM network was examined node by node (see Appendix 2). The species which had most sounds in a particular node, won and the possible other sounds were classified as outliers. If there were same number of sounds in the particular node, all were classified as outliers. Hence, it turned out that 32 sounds were classified as outliers: ten sounds of the Mallard, 14 sounds of the Greylag Goose, two sounds of the Quail, one sound of the Corncrake, three sounds of the Magpie, and two sounds of the Spotted Crake. Five sounds were same among the outliers of the MLP and SOM networks. In Figure 5-2 one of them is illustrated.

Most of the SOM network misrecognitions might result from the variation of the coefficient pattern shape of the highest level (level 6) in the wavelet packet tree. The value of the feature vectors varied too much and it might have caused the outlier in 23 cases. An other presumable reason was the segmentation error, which might cause the misrecognition in nine cases. For example, the coefficient patterns of two sounds of the Magpie resembled the Quail more than the Magpie. It should be noticed that the SOM classified all the ten outliers of the Mallard as the Greylag Goose and 12 of 14 outliers of the Greylag Goose as the Mallard. That was very interesting, because both of those species belong to the *Anseriformes* and so they belong to the same order and family.

It can be seen both from the feature space figures (see Figures 5-2 to 5-5) and the results of the MLP network (see Table 5-2) and the SOM network (see Table 5-4) that the cluster of the Graylag Goose spreads out most and it also has most misrecognitions. That might be due to the very different sound types of the Graylag Goose. The other interesting fact is that the misrecognitions of the Magpie were classified in most cases as the Quail. That might result from the fact that some sounds of the Magpie were much shorter than the other. Hence, the wavelet coefficients became different from the reference sounds of the Magpie. Thus, more research of the feature selection is needed.

In conclusion, the MLP classified 96 % and the SOM 93.8 % of the test sounds correctly. The difference between the results might be a consequence of the MLP being a supervised and the SOM an unsupervised network. The supervised network can perhaps learn and generalize the data better and faster than unsupervised network. On the other hand, the data might have not been entirely suitable for the SOM network. Although the data had been checked, there were a few sounds, which had false segmentation or poor sound quality. That might have disturbed the unsupervised SOM more than supervised

MLP. Anyway, it turned out that both neural networks classified the data used in this study very well.

In conclusion, the presented results of this thesis are very encouraging. It turned out that it is possible to recognize bird sounds using neural networks with only four features calculated from the wavelet packet coefficients. Although the neural networks have many benefits, like their ability to learn and therefore generalize, there is a long way to go before the recognition system beats the human ear. When using neural networks in the pattern classification, there has to be a fixed number of classes into which activations are to be classified. This thesis involved the study of the sound recognition of eight bird species. However, there exist 444 bird species only in Finland (BirdLife Suomi 2005). Hence, the disadvantage of the neural networks is the fixed number of output classes. When more species should be classified, the network has to be retrained all over again before it can be tested with a new set of birds. In the future work, our purpose is to study the possibilities of other classifiers. For example k-Nearest-Neighbour nonlinear classifier has been used successfully in automatic recognition of bird sound (Fagerlund 2004). As it can be seen from Figures 5-2 to 5-5, the features used in this thesis were clustered quite well. This might suggest that also other classifiers can yield as good or even better results as presented in this study.

The automatic classification presents new method for identifying and differentiating bird species by their sounds, and may give new tools also for bird researchers. However, the automatic recognition and classification of bird species is by no means an easy task. The fact that sounds and calls vary among the species and the same species might have many call types make automatic classification even more difficult. The wavelet transform has proved to be an efficient method to be taken into consideration. The tools presented in this thesis provide quite robust an approach for recognition and classification purposes, particularly among the inharmonic and transient bird sounds. In the future work, the long-term goal of the research is to develop a system capable of recognizing the majority of Finnish bird species from a continuous recording. Automatic recognition of bird sounds would facilitate the study of the vocal communications of birds and the development of song in young birds in an objective and systematic way. A bird recognition system would also have a high commercial potential among bird-watchers.

REFERENCES

Akansu, A.N., Haddad, R.A. (1992), Multiresolution Signal Decomposition: Transforms, Subbands, and Wavelets. Boston: Academic Press. 376 pages.

Baker, M.C., Logue, D.M. (2003), Population Differentiation in a Complex Bird Sound: A Comparison of Three Bioacoustical Analysis Procedures. Ethology 109, 3 (March). pp. 223-242.

Berry, S. (1999), Practical Wavelet Signal Processing for Automated Testing. In Proceedings of the IEEE Systems Readiness Technology Conference (AUTOTESTCON '99), San Antonio, 30 August - 2 September 1999. pp. 653-659.

BirdLife Suomi (2005), Suomessa tavatut lintulajit. http://www.birdlife.fi/lintuharrastus/suomessa_tavatut_lintulajit.shtml. (10.3.2005)

Boggess, A., Narcowich, F.J (2001), A First Course in Wavelets with Fourier Analysis. New Jersey: Prentice-Hall, Inc. 283 pages.

Bradbury, J.W. (2003), Parrots and Technology. In Proceedings of the 1st International Conference on Acoustic Communication by Animals, Maryland, USA, 27-30 July 2003. pp. 29-30.

Bruce L.M., Cheriyadat, A., Burns, M. (2003), Wavelets: Getting Perspective. Potentials, IEEE 22, 2 (April-May). pp. 24-27.

Catchpole, C.K., Slater, P.J.B. (1995), Bird Song: Biological Themes and Variations. Cambridge: Cambridge University Press. 248 pages.

Daubechies, I. (1992), Ten Lectures on Wavelets. Philadelphia: Society for Industrial and Applied Mathematics. 357 pages.

Deecke, V.B., Ford, J.K.B., Spong, P. (1999), Quantifying Complex Patterns of Bioacoustic Variation: Use of a Neural Network to Compare Killer Whale (*Ornicus orca*) Dialects. Journal of Acoustical Society of America 105, 4 (April). pp. 2499-2507.

Elowson, A.M., Hailman, J.P. (1991), Analysis of Complex Variation: Dichotomous Sorting of Predator-elicited of the Florida Scrub Jay. Bioacoustics Vol. 3. pp. 295-320.

Fagerlund, S. (2004), Automatic Recognition of Bird Species by Their Sounds. Master's Thesis, Helsinki University of Technology. 56 pages.

Greenewalt, C.H. (1968), Bird Song: Acoustics and Physiology. Washington: Smithsonian Institution Press. 194 pages.

Groth, J.G. (1993), Call Matching and Positive Assortative Mating in Red Crossbills. The Auk 110, 2 (April). pp. 398-401.

Haykin, S. (1994), Neural Networks: a Comprehensive Foundation. New York: Macmillan College Publishing Company, cop. 696 pages.

Heiskanen, I. (1993), Yön linnut [CD-ROM]. Heinola: LYL:n lintuvaruste. CD and appendix 12 pages.

Heiskanen I. (2004), Unpublished Recordings.

Härmä, A. (2003), Automatic Identification of Bird Species Based on Sinusoidal Modelling of Syllables. In Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '03), Hong Kong, 6-10 April 2003. pp. V-545-548.

Härmä, A., Somervuo, P. (2004), Classification of the Harmonic Structure in Bird Vocalization. In Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '04), Montreal, Canada, 17-21 May 2004. pp. V-701-704.

Jin, Q., Wong, K.M., Luo, Z.Q. (1994), Wavelet Function, Scaling Function and Discrete Wavelet Transform. In Proceedings of the IEEE-SP International Symposium on Time-Frequency and Time-Scale Analysis, Philadelphia, USA, 25-28 October 1994. pp. 409-412.

Kalinainen, P. (2004), Unpublished Recordings.

Kammler, D.W. (2000), A First Course in Fourier Analysis. Upper Saddle River: Prentice Hall. 790 pages.

Kohonen, T. (2001), Self-Organizing Maps. Berlin: Springer. 501 pages.

Learned, R. (1992), Wavelet Packet Based Transient Signal Classification. Master's Thesis, Massachusetts Institute of Technology. 93 pages.

Lewicki, M.S. (2002), Efficient Coding of Natural Sounds. Nature neuroscience 5, 4 (April). pp. 356-363.

Lovell, S.F., Lein, M.R. (2004), Song Variation in a Population of Alder Flycatchers. Journal of Field Ornithology 75, 2 (April). pp. 146-151.

McIlraith, A.L., Card, H.C. (1997), Birdsong Recognition Using Backpropagation and Multivariate Statistics. In the Proceedings of the IEEE Transactions on Signal Processing 45, 11 (November). pp. 2740-2748.

Mertins, A. (1999), Signal Analysis: Wavelets, Filter Banks, Time-Frequency Transforms and Applications. Chichester: Wiley, cop. 317 pages.

Mesgarani, N., Shamma, S. (2003), Bird Call Classification Using Multiresolution Spectrotemporal Auditory Model. In Proceedings of the 1st International Conference on Acoustic Communication by Animals, Maryland, USA, 27-30 July 2003. pp. 155-156.

Misiti, M., Misiti, Y., Oppenheim, G., Poggi, J.-M. (2000), Wavelet Toolbox for Use with Matlab. Natick (MA): MathWorks. 572 pages.

Nason, G.P. (1999), A Little Introduction to Wavelets. IEE Colloquium on Applied Statistical Pattern Recognition (Ref. No. 1999/063), 20 April 1999. pp. 1/1-1/6.

Oppenheim, A.V., Willsky, A.S., Nawab, S.H. (1997), Signals and Systems. Upper Saddle River: Prentice-Hall. 957 pages.

Phelps, S., Ryan, M.J. (1998), Neural Networks predict Response Biases of female Tungara Frogs. Proceedings of the Royal Society of London. Series B, Biological Sciences 265, 1393 (February). pp. 279-285.

Pittner, S., Kamarthi, S.V. (1999), Feature Extraction from Wavelet Coefficients for Pattern Recognition Tasks. IEEE Transaction on Pattern Analysis and Machine Intelligence 21, 1 (January). pp. 83-88.

Placer, J., Slobodchikoff, C.N. (2000), A Fuzzy-neural System for Identification of Species-specific Alarm Calls of Gunnison's Prairie Dogs. Behavioural Processes 52,1 (October). pp. 1-9.

Rioul, O., Vetterli, M. (1991), Wavelets and Signal Processing. IEEE Signal Processing Magazine 8, 4 (October). pp. 14-38.

Robb, M.S. (2000), Introduction to Vocalizations of Crossbills in Northwestern Europe. Dutch Birding 22, 2. pp. 61-107.

Roche, J.C., Chevereau, J. (2001), Guia Sonora de las Aves de Europa [CD-ROM]. Barcelona: Lynx Edicions. 10 Audio-CDs and appendix 74 pages.

Saurola, P. (1995). Suomen pöllöt [CD-ROM]. Helsinki: Kirjayhtymä. CD and appendix 8 pages.

Schulze, A. (2003). Die Vogelstimmen Europas, Nordafrikas und Vorderasiens [CD-ROM]. Germering: Musikverlag Edition Ample. 17 Audio-CDs and appendix 63 pages.

Sheng, Y. (2000), Wavelet Transform. Poularikas, A.D. (eds.), The Transforms and Application Handbook. Boca Raton: CRC Press. pp. 10-1–10-88.

Soman, A.K., Vaidyanathan, P.P. (1992), Paraunitary Filter Banks and Wavelet Packets. In Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP ‘92), San Francisco, USA, 23-26 March 1992. pp. 397-400.

Somervuo, P., Härmä, A. (2003), Analyzing Bird Song Syllables on the Self-Organizing Map. In Proceedings of the Workshop on Self-Organizing Maps (WSOM ’03), Hibikino, Japan, 11-14 September 2003. Proceedings on CD-ROM.

Somervuo, P., Härmä, A. (2004), Bird Song Recognition Based on Syllable Pair Histograms. In Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP ‘04), Montreal, Canada, 17-21 May 2004. pp. V-825-828.

Strang, G., Nguyen T. (1996), Wavelets and Filter Banks. Wellesley: Wellesley-Cambridge Press. 490 pages.

Suthers, R.A., Beckers, G., Zollinger, S.A., Vallet, E., Kreuzer, M. (2003), Mechanisms of Vocal Complexity in Birds. In Proceedings of the 1st International Conference on Acoustic Communication by Animals, Maryland, USA, 27-30 July 2003. pp. 237-238.

Tan, C.-J., Gan, W.-S. (1999), Wavelet Packet Decomposition for Spatial Sound Conditioning. Electronics Letters 35, 21 (October). pp. 1821-1823.

Tanttu, J.T., Turunen, J., Ojanen M. (2003), Automatic Classification of Flight Calls of Crossbill Species (*Loxia* sp.). In Proceedings of the 1st International Conference on Acoustic Communication by Animals, Maryland, USA, 27-30 July 2003. pp. 239-240.

Tanttu, J.T., Turunen, J., Selin, A., Ojanen M. (2005), Automatic Feature Extraction and Classification of Crossbill (*Loxia* spp.) Flight Calls. Submitted to Bioacoustics.

Terry, A.M.R., McGregor, P.K. (2002), Cencus and Monitoring Based on Individually Identifiable Vocalizations: the Role of Neural Networks. Animal Conservation 5, 2 (May). pp. 103-111.

Theodoridis, S., Koutroumbas, K. (1999), Pattern Recognition. San Diego: Academic Press. 625 pages.

Thorn, A. (2003), Artificial Neural Networks for Vocal Repertoire Analysis. In Proceedings of the 1st International Conference on Acoustic Communication by Animals, Maryland, USA, 27-30 July 2003. pp. 245-246.

Umbaugh, S.E. (1998), Computer Vision and Image Processing: A Practical Approach Using CVIPtools. USA: Prentice-Hall Inc. 504 pages.

Zollinger, S.A, Riede, T., Suthers, R.A (2003), Production of Nonlinear Phenomena in Northern Mockingbirs (*Minus polyglottos*). In Proceedings of the 1st International Conference on Acoustic Communication by Animals, Maryland, USA, 27-30 July 2003. pp. 283-284.

APPENDIX 1

The species, recorders, references, and number of the sounds and tracks used in this thesis are listed in this Table.

Species	Recorder	Reference	Number of Sounds		Number of tracks		
				total	Training	Testing	Total
ANAPLA	Kalinainen Pertti	Kalinainen (2004)	124	198	5	1	6
		Schulze (2003)	74		2	1	3
ANSANS	Kalinainen Pertti	Kalinainen (2004)	157	194	9	3	12
		Schulze (2003)	37		-	1	1
COTCOT	Kalinainen Pertti	Kalinainen (2004)	137	273	1	2	3
		Schulze (2003)	70		1	-	1
	Heiskanen Ilkka	Heiskanen (1993)	66		1	-	1
CRECRE	Kalinainen Pertti	Kalinainen (2004)	343	553	7	1	8
		Schulze (2003)	141		1	-	1
	Heiskanen Ilkka	Heiskanen (1993)	69		-	1	1
GLAPAS	Kalinainen Pertti	Kalinainen (2004)	57	161	1	1	2
		Schulze (2003)	32		1	-	1
	Heiskanen Ilkka	Saurola (1995)	20		1	1	2
	Bruun Jan-Erik	Saurola (1995)	18		1	-	1
	Roche Jean	Roche (2001)	34		1	-	1
LOCFLU	Kalinainen Pertti	Kalinainen (2004)	328	1218	-	1	1
		Schulze (2003)	336		1	-	1
	Heiskanen Ilkka	Heiskanen (1993)	554		1	-	1
PICPIC	Kalinainen Pertti	Kalinainen (2004)	97	300	-	1	1
		Schulze (2003)	102		1	-	1
	"BirdLife"		3		1	-	1
	Christensen Jens		7		1	-	1
	Heiskanen Ilkka	Heiskanen(2004)	91		3	-	3
PORPOR	Kalinainen Pertti	Kalinainen (2004)	155	235	1	1	2
		Schulze (2003)	39		1	-	1
	Heiskanen Ilkka	Heiskanen (1993)	41		1	-	1

APPENDIX 2

Classification result of the 10 x 10 SOM network. The coloured areas illustrate the nodes of each species. The other nodes are outliers. The numbers indicate how many test sounds have fallen on each node.

ANAPLA

0	0	0	0	0	2	0	0	0	0
0	0	0	0	4	6	2	0	0	0
0	0	0	0	0	7	2	1	0	0
0	0	0	0	1	2	9	4	1	0
0	0	0	0	0	0	1	5	6	0
0	0	0	0	0	0	0	2	4	1
0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0

GLAPAS

0	0	2	11	0	0	0	0	0	0
0	0	5	4	0	0	0	0	0	0
0	0	2	13	1	0	0	0	0	0
1	4	3	2	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0

ANSANS

0	0	0	0	2	0	5	2	4	4
0	0	0	0	0	1	0	0	3	1
0	0	0	0	0	2	0	0	0	0
0	0	0	0	5	3	2	0	2	0
0	0	1	1	3	3	4	0	2	0
0	0	0	0	0	0	1	1	4	0
0	0	0	0	0	0	1	0	0	1
0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	1	0	0	0	0
0	0	0	0	0	0	0	0	0	0

LOCFLU

0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	3	0
0	0	0	0	0	0	0	141	103	0
0	0	0	0	0	0	0	1	47	0
0	0	0	0	0	0	0	33	0	0

COTCOT

0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0
2	0	0	0	0	0	0	0	0	0
1	0	2	0	0	0	0	0	0	0
2	1	0	0	0	0	0	0	0	0
4	0	0	0	6	1	0	0	0	0
7	8	9	8	31	1	0	0	0	0

PICPIC

0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0
0	0	0	0	1	0	0	0	0	1
0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	13	0	0	0	0
0	0	0	0	0	13	24	0	0	0
0	0	0	0	2	43	0	0	0	0

CRECRE

0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	18
0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	1	0	0	0	91

PORPOR

0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0
0	0	1	22	19	0	0	0	0	0
0	0	5	13	7	0	0	0	0	0
0	0	0	1	0	0	1	0	0	0
0	0	0	0	0	0	0	0	0	0