ZÁPADOČESKÁ
UNIVERZITA
V PLZNI

Faculty of Applied Sciences
Department of Cybernetics

Birds Individual Automatic Recognition

PhD THESIS REPORT

Ing. Ladislav Ptáček

advisor: Doc. Ing. Luděk Muller, Ph.D.

Pilsen, 2012

Contents

List of Figures

List of tables

List of Abbreviations

ANN	Artificial Neural Network
CI	call independent
CMD	Window Command Line
EM	Expectation Maximization
FE	Feature Extraction
FI	Fisher Information
FA	False accept
FR	False reject
GMM	Gaussian Mixture Model
GWF	Greenwood warping function
HMM	Hidden Markov Model
LLR	Log-Probability Ratio
LPC	Linear Predictive Coding
LPCC	Linear Prediction Cepstral Coefficients
MAP	Maximum A-posteriory Probability
MFCC	Mel Frequence Cepstral Coefficients
ML	Maximum Probability
MLLR	Maximum Probability Linear Regression
MLP	Multilayer Perceptron
NN	Neural Networks
PLP	Perceptual Linear Prediction
PLPCC	Perceptual Linear Prediction Cepstral Coefficients
PNN	Probabilistic Neural Networks
PVM	Tool for automatic speaker verification
SD	Speaker Dependent
SI	Speaker Independent
SOM	Self organizing Map
SR	Speaker Recognition
SVM	Support Vector Machine
UBM	Universal Background Model
VT	Vocal Tract

1 Introduction

1.1 Overview

Understanding animals is an ancient theme of many novels, sci-fi as well as children books. Automatic recognition of animal sound represents a very interesting area with great potential. In the animal world there is a great number of species for which vocalization plays an important role. In addition, their vocal tract anatomy is similar to the human vocal tract. Animals make sounds for many purposes: communication, defending territory, danger warning, courtship, fear, satisfaction, expressing emotions, etc. People do not understand animals. Moreover, it is a logical assumption that we will never be able to accurately interpret the meanings of animal sounds. Creating "interpretative" dictionary appears to be an unrealistic task. However, it should be possible to get other crucial information from sound: *Identification of the individual*. This task is similar to the problem of speaker recognition. Identification of people has been used for a long time and the background knowledge is well mapped.

This work deals with the use of automatic recognition of bird individuals Chiffchaff (*Phylloscopus collybita (Lat.), Budníček menší (Czech*) and Willow Warbler (*Phylloscopus trochilus (Lat.), Budníček větší (Czech*). Chiffchaff and Warbler are abundant in the Czech Republic and the whole of Europe, which makes recording, monitoring and mapping of origin relatively easy. The work is based on the well-known method of speaker text-independent verification on humans using the GMM. Any "linguistic" content (transmitted information) is ignored and instead we focus on the characteristics and properties of voice of each individual.

1.2 Motivation and goals

The major challenge in ornithology is the impossibility to differentiate individuals from each other. The only solution is bird ringing. This procedure brings some negatives:

- It is necessary to capture the bird.
- The bird gets ringed for life.

Firstly, the capture is a very stressful event. If the ornithologist does not wear gloves, the bird is exposed to human contact. The bird can be caught in a net for several hours, till the zoologist arrives. It happens especially when night birds are caught. Secondly, a bird receives a ring on the body, which changes its appearance, increases its weight and sometimes hinders its movement. Furthermore, the ring may not only bother the bird itself, but there is a question whether its colour and appearance does not distract the partners or other individuals from its community.

The author of this work cooperates with some ornithologists of the Faculty of science of University of South Bohemia, who have observed that *Chiffchaffs* which were caught, do not return to the same place very often. Estimation of return of the ringed birds is about 15% of the original number. Moreover, it is impossible to repeatedly catch the birds once caught. Ornithologists agree that *captured birds alter their behaviour.* It follows that fundamental question ornithologists have to answer is: *To what extent this is due to ringing?* A new approach is the only way to get an answer. It is necessary to find a solution that will allow researchers to move further. Based on these facts, the use of automatic recognition of individuals opens up entirely new possibilities to solve the problem of *contactless identification* without the risky ringing.

Main goals of this thesis are

- To use bird verification records made continuously in terrain under real circumstances, i.e. noisy, many birds sing in time, variable distance between ornithologist and birds changing during recording.
- To use the real data which were recorded by ornithologist in real circumstances without any editing of the records (cut, glue, purify).
- To find a solution to automatic recognition of Chiffchaff individuals for months even years, i.e. speaker/bird verification task.
- To develop techniques and define parameters for the best result achievement.
- To determine the probability to which the recognition system is capable to identify Chiffchaff individuals.
- To specify the influences and obstacles that would affect the recognition ability.
- To enable simultaneous detection of more than one individual extending the type of performed tasks to both speaker verification and speaker identification.

1.3 State-of-the-art

The use of method GMM for bird individual recognition was first described by Cheng et al. [CHE10]. Cheng used MFCCs parameters and GMM classification. Objects of research were birds of Chinese warbler, Humes warbler, Gansu leaf warbler and Chinese bulbul species. For data they used single songs of specified length, which was cut from song records. Recording was performed outdoors in noisy background. The achieved accuracy of recognition was between 89.1% and 92.5%.

Next, Kuntoro et al. experimented with both song-type classification and individual identity clustering [KUN10]. For song-type classification HMM was used, with achieved accuracy of the song-type between 50% and 98.8%. For bird individual identity the clustering error rate achieved was from 2.9% to 50%, which was evaluated by the author as unusable. Data from 2000 was used as training data, while data of 2001 was used for evaluation.

Fox describes call independent identification in birds [FOX08], where the records were cut into parts and then some parts were used for training and some for identification. The length of the parts varies, but is about 10 [s]. MFCC was used for parameters, while the classifier used was ANN MLP implemented in the NN toolbox in Matlab. The network had one hidden layer with 16 neurons. Identification accuracy for willie wagtails was 72.9% and 97.1% for non-trained and trained ANN respectively, 54.3% and 98.6% for canaries and 75.7% and 96.5% for singing honeyeaters. The techniques used for signal enhancement and for removal of noise from recordings of passerine are discussed. It was demonstrated that the accuracy greatly depends on noise because after the noise reduction procedure the accuracy rate increased.

Clemins deals with classification of animal vocalization using MFCC and PLP parameters and HMM classifier implemented in HTK [CLE05]. In the first part call-type identification is solved, the second speaker identification. It was recommended to use Greenwood warping function (GWF) instead of Mel bank filter. The GVF is better suited as the filter to animal species auditory. Influence of parameters to accuracy is discussed in detail and experimentally proved. Achieved results for call type recognition were between 51%

and 90%. The results highly depend on type of used parameters and classifier. Tested species were croaks, elephants, and beluga whales. For every species a particular GWF was computed.

Selin focused on bird sound classification using wavelets [SEL05]. For automated classification of acoustic signals ANN was used. MLP and self-organizing map (SOM) were used as classifiers. Eight bird species were tested. Accuracy of sound classification was 96% and 93.8% for MLP and SOM respectively.

Some authors recognize the sounds of animals in order to identify (interpret) the meaning. Molnar et al collected more than 6,000 barks in an attempt to recognize the meaning of dog barking [MOL08]. He could distinguish five kinds of barking, identifying their meaning as stranger, fight, alone, ball, play. Classification efficiency was 43% to 52%.

In my knowledge there is no article which uses raw data in addition over the years. Most of the articles describe one-time experiments. Repeating the experiments seems to be very difficult if not impossible.

1.4 Outline

The first part deals with hearing and vocalization of both humans and birds. It describes main differences between bird and man and introduces models of their vocal tracts. Next, chapter 3 describes signal analysis and cepstral parameters extraction. Chapter 4 outlines basic definition of the speaker recognition problem, whereas chapter 5 is focused on main phases of speaker verification task. It also introduces GMM-UBM system and EM algorithm. Chapters 6 and 7 deal with core of this thesis, where the first one describes speaker verification system „PVM“, which will be used for birds recognition experiments. The second describes used data, evaluation methodology and achievement of preliminary experiments which have not been performed yet. Last chapters summarize plans of future work.

2 Human and bird, voice and song

From the anatomical point of view the vocal tract of passerine is similar to humans. The fundamental difference is that birds have a syrinx which is equivalent to the human voice box or larynx. Like the larynx, the syrinx contains special membranes which vibrate and generate sound waves when air from the lungs is forced through them [36_bird songs]. It allows the birds to generate two independent audio signals simultaneously. In practice, however, there are only a few „two-tone singers“.

A significant feature of a birdsong is its duration. It is common to hear a bird singing continuously tens of seconds without interruption. It is considered that this is achieved thanks to the anatomy of the bird vocal tract mentioned above, where one of the tubes drives the singing while the second performs micro-breathing.

2.1 Human voice model

The model of the vocal system is shown in Figure 2.1. The air flows from the lungs. During breathing the glottis is opened, while during speech production the glottis is opened and closed. The air flows through causing oscillation and producing vocalization. The vocal basic tone F_0 is based on this vibrations. When creating the *voiced* vowels the glottis is nearly closed. When voiced consonants are produced the glottis is not closed so tight causing the sound of not periodical (tonal, pure) character. When creating *unvoiced* sounds the vocal cords are almost opened and the sound is created by modification of the air stream in cavities.

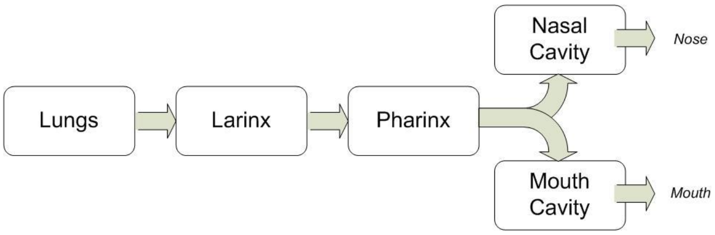

Figure 2.1: Human vocal tract

As an equivalent for circuit diagram of the vocal tract the connection of two generators is used, which are alternately connected to the circuit according to processed sound, see Figure 2.2. Pulse generator is dedicated for voiced sounds and white noise source for unvoiced.

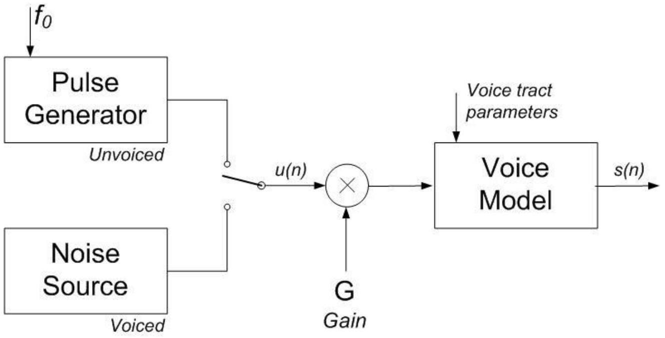

Figure 2.2: Vocal tract, equivalent circuit diagram

After simplifying, the whole process can be replaced by the source signal *u(t)* passing through the system with impulse response *h (t),* as shown in Figure 2.3

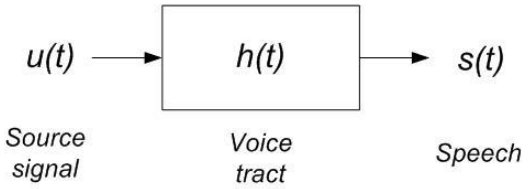

Figure 2.3: Vocal tract, simplification

The human speech is therefore modelled by convolution of the excitation signal *u(t)* and the vocal tract with impulse response *h(t).* This is used for speech synthesis purposes as well as for finding the speaker voice characteristics. In this thesis we use the same approach for getting the birds vocal characteristics. In the timeline convolution is described as

$$u(t) * h(t) = s(t), \tag{2.1}$$

For discrete signal

$$u(n) * h(n) = s(n). \tag{2.2}$$

In frequency domain we get

$$S(f) = X(f) \cdot H(f)_{,} \tag{2.3}$$

and if we use Z-transform

$$U(z) \cdot H(z) = S(z). \tag{2.4}$$

To obtain vocal characteristics it is necessary to perform deconvolution, Figure 2.4. Detail description how to compute the cepstral coefficients using the deconvolution is described in chapter 3.1.

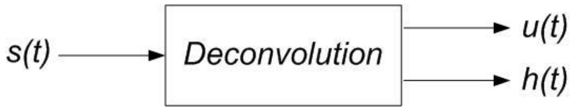

Figure 2.4: Deconvolution

2.2 Birds and human hearing

Human and bird ears are variously sensitive to different frequencies. The Figure 2.5 shows the dependence of both human and birds hearing on the frequency.

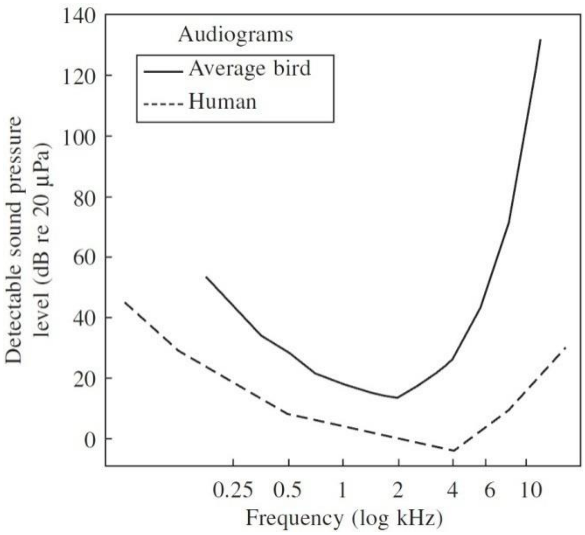

Figure 2.5: Average audibility curves human and bird [CAT08].

In humans, this dependence is described by Fletcher-Munson curves. Birds are less sensitive to lower frequencies to the contrary of better hearing at higher frequencies. It probably relates to the higher frequency bands of birdsong, their communication running between 0.5 kHz and 6 kHz in average.

2.3 Birds vocal

For human frequency associated with vocal tract dimensions is fundamental. Thus, it is lower for men, and highest for children. In animals much greater variability of vocal box can be found. Figure 2.6 shows the dependence on animal body mass and emphasized frequencies of vocalization. With a suitable choice of a scale a line with a slope of *-1* can be added to the graph, describing this dependency. Small animals use high frequencies while larger animals lower frequency.

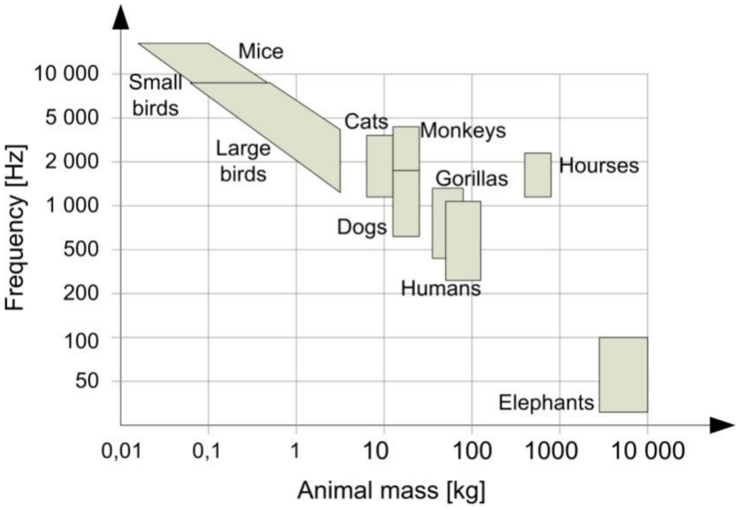

Figure 2.6: Animal body mass and frequencies of vocalization.[CAT08]

The dependence for both humans and animals is related to the basic relationship between wavelength and frequency

$$\lambda = \frac{c}{f} \tag{2.5}$$

where c is the speed of sound. Dry air is approximated by the relation

$$c = (\mathbf{331.57} + \mathbf{0.607} \cdot t)\ [ms^{-1}]. \tag{2.6}$$

A syrinx is the principal organ of birdsong creation. Figure 2.7 shows divided structure with two sound generators.

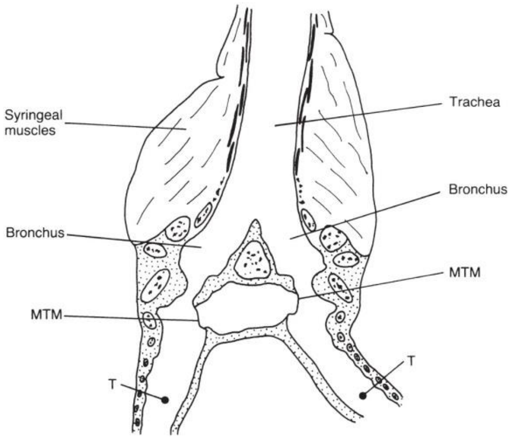

Figure 2.7: Section through the syrinx of a brown thrasher.
(T) thermistors, (MTM) medial tympaniform membranes (Suthers 1990).

Unlike humans, animals are usually equipped with less noise harmonics. Some animals may produce purely sinusoidal (singers) and pure noise character (small rodents). Just as the human vocal tract, the principal of sound generating in birds will be approximated by convolution both generating signal *x(n)* and impulse response *h(n)* of the vocal tract:

$$s(n) = x(n) * h(n), \tag{2.7}$$

where *s(n)* is song (speech signal), *x(n)* is an excitation (signal source) and *h(n)* is impulse response of vocal tract (vocal tract filter).

In general, the sound generated (vocal tract) in many species of animals is similar to humans: Monkey, some singers, cetaceans. Some birds may also sing in two-tone (lark, nightingale, thrush). Some types of sound production are completely different and operate without the use of vocal tract. For example, the oscillation of the wings (mosquito), using a special membrane (cicada), rubbing the wings together (cricket), in rodents banging his head against the wall hole (Lesser Bamboo Rat).

It is impossible or at least very difficult for animals to build something like a dictionary or lexicon of the speech corpus. However humans often "understand" the meaning of animal vocalization. For some species a lexicon can be made, though with difficulty and with obvious reservations about the interpretation of such imperfection. Main incompleteness and ambiguity originate when a sound has more than one meaning, for example. [MOL08].

For animals whose vocal tract resembles human it can be theoretically assumed that a better equipped brain would produce sound similar to humans. On the contrary, some singers who can imitate human speech (parakeet, cockatiel, starling) clearly have less powerful brain than other species. It is certain that the ability to imitate human speech is related to the more developed brain parts used for vocal tract control. Birds do not understand the meaning and content of the spoken words. Spoken (sung) words are only an "interesting sound" developed because of good musical memory (imprint of human words). The exact mechanism of speech is unknown and it is questionable whether it will ever be understood.

2.3.1 Bird song

The basic bird song stands between calls and songs. The calls are short squawks emitted by birds as an emergency or warning sound. The song is natural vocalization of the passerines. It consists of phrases and syllables (see Figure 2.8).

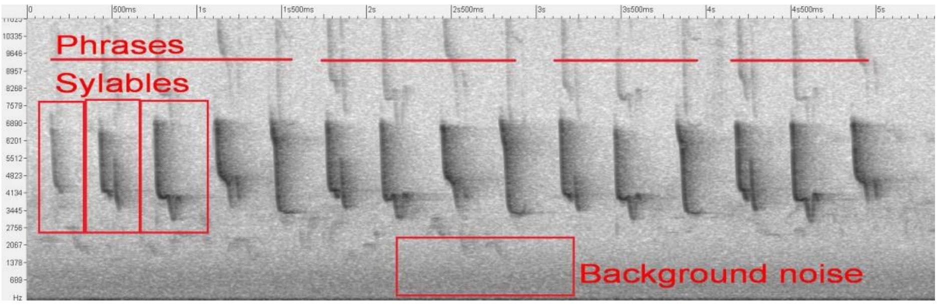

Figure 2.8: One song of *Chiffchaff*

The elements are then divided into further elements, see Figure 2.9

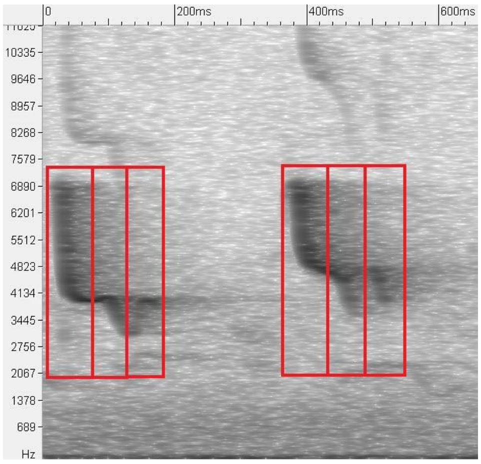

Figure 2.9: Two syllables divided into elements.

3 Signal analysis

3.1 Cepstrum

The real cepstrum is defined as the inverse Fourier transform (IFT) of the logarithm of the amplitude signals spectrum

$$c_n = IDFT\left\{\ln\left|S\left[k\right]\right|\right\} = \sum_{k=0}^{N-1}\ln\left|S\left[k\right]\right|\exp\left\{j\frac{2\pi kn}{N}\right\} \tag{3.1}$$

where *c(n)* are the individual cepstral coefficients, *X [k]* signal spectrum. Sound (speech signal) *s(t)* is formed by convolution of source signal *x(n)* and impulse response of vocal tract *h(n)* so for discrete signals is

$$s(n) = x(n) * h(n). \tag{3.2}$$

For automatic recognition it is necessary to obtain parameters *x(n)* and *h(n)* separately. Then they can be used to build speakers models. Discrete spectrum of the signal is given as

$$S(k) = DFT\left\{s(n)\right\} \tag{3.3}$$

$$S(k) = \sum_{n=0}^{N-1}s(n)\exp\left\{-j\frac{2\pi k}{N}n\right\}. \tag{3.4}$$

If DFT is applied on the discrete signal *s(n)* the convolution formula is transformed to the:

$$DFT\{s(n)\} = DFT\{x(n) * h(n)\} = DFT\{x(n)\} \cdot DFT\{h(n)\}. \tag{3.5}$$

In frequency domain the equation changes to

$$S(k) = X(k) \cdot H(k). \tag{3.6}$$

Applying the natural logarithm changes the multiplication to the sum

$$ln[|S(k)|] = \mathbf{ln}[|X(k)|] + ln[|H(k)|]. \tag{3.7}$$

To obtain the cepstral coefficients the IDFT was applied to the previous formula

$$IDFT\left\{\ln\left(|S(k)|\right)\right\} = IDFT\left\{\ln\left(|X(k)|\right)\right\} + IDFT\left\{\ln\left(|H(k)|\right)\right\}. \qquad (3.8)$$

Because we assume that the speech signal is given by the convolution of two input signals, then for the real cepstra coefficients is valid

$$c(n) = Re\{IDFT[ln|S(k)|]\} = c_x(n) + c_h(n), \qquad (3.9)$$

And each cepstral coefficients can then be expressed as

$$c_s(n) = \mathrm{Re}\left\{IDFT\left\{\ln\left(|S(k)|\right)\right\}\right\} \qquad (3.10)$$

$$c_x(n) = \mathrm{Re}\left\{IDFT\left\{\ln\left(|X(k)|\right)\right\}\right\} \qquad (3.11)$$

$$c_h(n) = \mathrm{Re}\left\{IDFT\left\{\ln\left(|H(k)|\right)\right\}\right\}. \qquad (3.12)$$

The equations show that the correlation of coefficients transforms into the sum. In practice, separation from each other is achieved by, so called, liftering. The lower coefficients represent the spectral envelope, i.e. the vocal tract, the higher are the excitation coefficients. Typically, for Speaker recognition tasks about 20 cepstral coefficients would be used. Notice that the Speaker recognition system process of Mel-Cepstral uses Discrete cosine transform DCT instead of the Inverse Fourier transformation.

3.2 Linear prediction cepstral coefficients

Linear prediction coding (LPC) predicts speaker parameters directly from a speech signal. The main stages of LPC calculation are shown in the Figure 3.1.

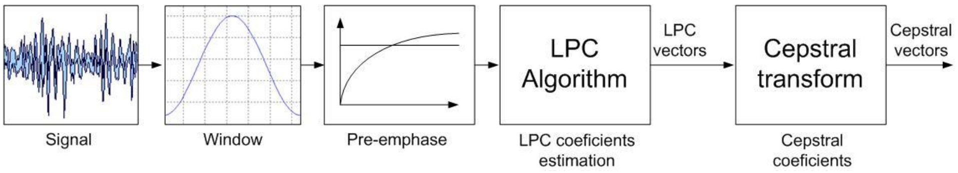

Figure 3.1: LPC coefficients calculation stages [BIM]

The principle of LPC is a computing the s(n) sample of a voice as a linear combination of a previous samples with excitation *u(k)* enhanced excitation *G*, so

$$s(k) = -\sum_{i=1}^{Q} a_i s(k-1) + Gu(k) \tag{3.13}$$

where G is the gain coefficient and Q is the order of the model. Transfer function H (z) can then be written as

$$H(z) = \frac{S(z)}{U(z)} = \frac{G}{A(z)} \tag{3.14}$$

where H (z) is defined as

$$H(z) = \frac{G}{1 + \sum_{i=1}^{G} a_i z^{-1}} \tag{3.15}$$

.

Figure 3.2 shows block diagram of cepstral coefficient extraction. The signal has to be weighted by a short enough window to be considered approximately stationary. Than can be used to determine the parameters a_i and *G* using the method of least squares.

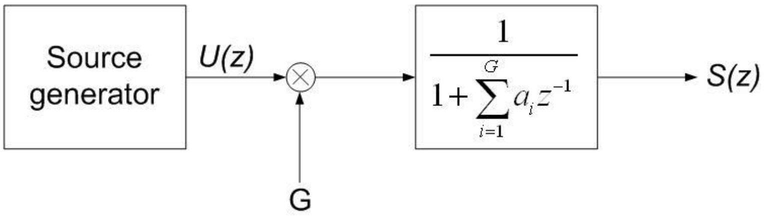

Figure 3.2: LPC, cepstral coefficients

3.3 Mel frequency cepstral coefficients

For the purpose of our work are used Mel Frequency Cepstral Coefficients (MFCC). First a sliding window is used to divide the speech into short segments. Each window is than pre-emphased to straighten frequency balances changed by a vocal tract. Next the Mel-frequency filter is applied to better adapt signal to human hearing. Due to logarithmic calculation the multiplication of spectrum changes to the sum. Finally by application of Discrete cosine transform (DCT) cepstral coefficients are obtained. The cepstrum is so-called the *Mel-frequency cepstrum* because Mel-frequency filter is used within the process.

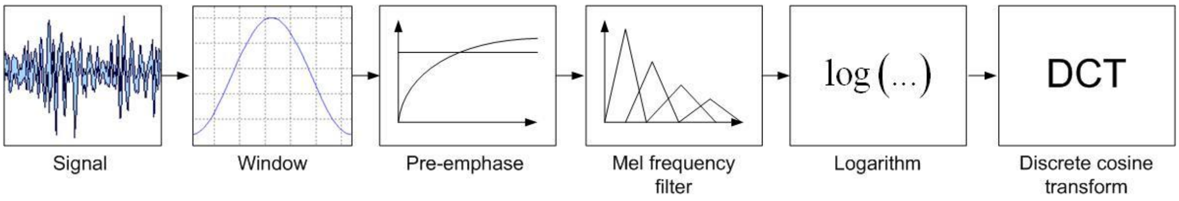

Figure 3.3: Mel-frequency cepstral coefficients computing, data diagram

For the purpose of our work we use cepstral coefficients. It is still unclear what sort of frequency filter will be used. The Mel-filter bank is adapted to human hearing so different filter adapted to birds should be more useful. In the current experiments the Mel-filtering filters were switched off.

3.4 Hamming window

Length of the window is chosen so that the period should be considered as a quasi-stationary. Step of sliding frame is chosen so that the overlap supresses the side effect of the window. The most commonly used is Hamming or Hanning window.

The Hamming window reduces leakage in the spectrum due to its side-effect. It is defined as

$$w[n] = 0{,}54 - 0{,}46\cos\frac{2\pi n}{N}, \quad where\ 0 \leq n \leq (N-1), \tag{3.16}$$

the ***N*** indicates the number of samples in length windows.

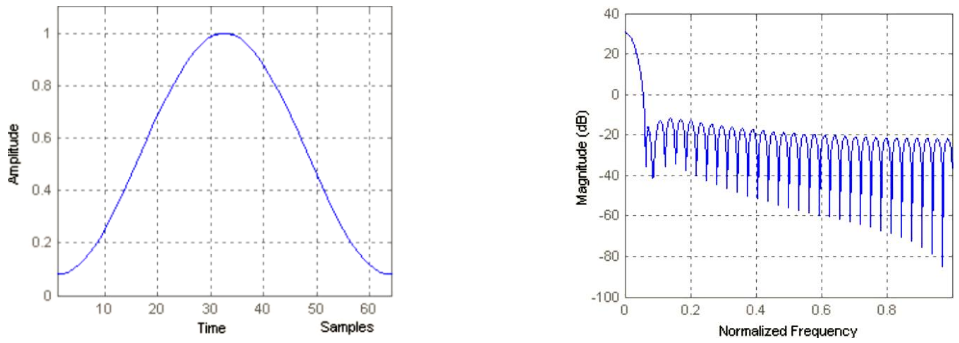

Figure 3.4: Hamming window a) Time domain, b) Frequency

For the speech signal Hamming window of length 20 [ms] with step 10 [ms] is usually used. For the purpose of our work the length 30 [ms] with 15 [ms] step has been empirically determined.

3.5 Pre-emphasis

During the progression of sound through an articulation mechanism higher frequencies are suppressed. Pre-emphasis is compensated by application of a first-order filter, which amplifies the higher frequency components. For modulating the filter shape

$$x_p(t) = x(t) - a \cdot x(t-1), \tag{3.17}$$

in discrete domain then

$$s_p[n] = s[n] - a \cdot s[n-1]. \tag{3.18}$$

Pre-emphasis coefficients are usually chosen in the interval [0.95 ÷ 0.99]. It is not yet finally determined what values should be used in our work. Notice some authors switch the pre-emphases off.

3.6 Mel filter bank

It was found empirically that the human ear perceives sound intensity signal unevenly depending on the frequency. Therefore, in applications of automatic speech recognition it is desirable to adjust the signal so that its distribution is near to the hearing. Mel filter banks are used for this correction. It converts the frequency *f* [Hz] into so-called frequency f_{MEL} [mel] which is based on human hearing. The conversion between f and f_{MEL} is defined as the relationship

$$f_{MEL} = 2595 \; log_{10}\left(1 + \frac{f}{700}\right), \tag{3.19}$$

Following figure shows the behaviour of the function

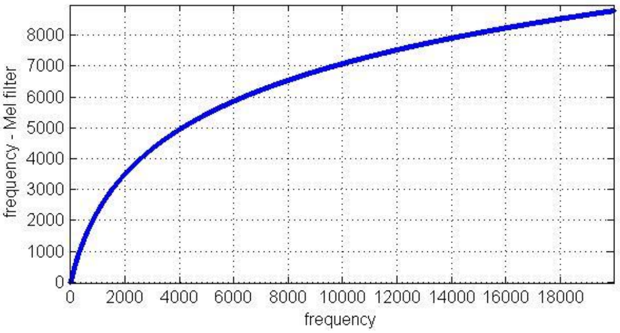

Figure 3.5: Characteristic Mel-frequency [mel] and frequency [f] domains

For the reversed conversion following relationship is valid

$$f = 700\left(10^{\frac{f_{MEL}}{2595}} - 1\right). \tag{3.20}$$

Mel filter banks are realized by a set of ***M*** bands. For instance, if bandwidth is 4 kHz 20 banks are usually used. These filters have a linear scale in a triangular shape with overlapping bands by half as seen in Figure 3.6.

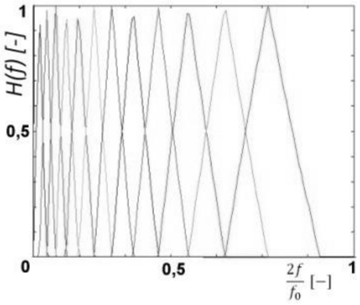

Figure 3.6: Mel filter bank

Using triangular overlapping filters helps to modify a magnitude of the spectrum with respect to human hearing.

4 Speaker recognition

4.1 Task definition

In our work we use theoretical foundations created for the purpose of speaker recognition. This can generally be divided into two separate tasks: Speaker Identification (SI) and Speaker verification (SV), each further divided into two sub-problems, for detail see Figure 4.1.

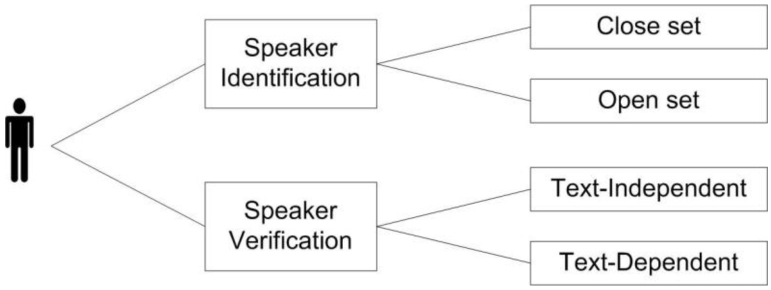

Figure 4.1: Speaker recognition task classification

4.2 Speaker identification

The task is to assign a specific speaker a speech record from a database of speakers. The aim is to determine which speaker or speech has been identified as most similar. A typical example using of SI is authentication of a person entering a building. Moreover it is important to define whether the number of people to be identified is closed or open.

For the closed-set case the models of all persons will be created. The goal is then to identify a person by using speaker models selected from a limited set

$$\Lambda(X)=\{\lambda_1,...,\lambda_L\}. \tag{4.1}$$

In the open-case set an unknown person may appears in addition to known persons. Then the set of models has to contain a model of unknown persons

$$\Lambda(X)=\{\lambda_1,...,\lambda_L\}\bigcup\Lambda_{UNKNOWN} \tag{4.2}$$

where the set of unknown speakers can theoretically be infinite

$$\Lambda_{UNKNOWN}=\{\lambda_1',...,\lambda_\infty'\}. \tag{4.3}$$

4.3 Speaker verification

The task is to *confirm* or *deny* whether the speech record belongs to a particular speaker. The system has to infer an identity which the speaker claims. An example of SV tasks is to authenticate a user when logging into a system. There are some applications where the speech is only biometric parameter that can be used, in a phone conversation for instance. Verification depends on what the speaker says. This task is further divided into two cases:

- text-independent
  speaker says any word or phrase

- text-dependent (text-constrained)
  speaker pronounces a pre-specified word or phrase, such as a digit or a code word.

Since we do not understand the bird language and we cannot even make the bird sing exactly what we order (except cases of extreme trained singers), the chosen approach for the automatic recognition system of a bird is S*peaker verification text-independent* task.

5 Speaker verification system

The process of speaker verification runs in three steps (see Figure 5.1) [NAIM]. First, a feature extraction is performed, see chapter 3.3. In the second step a training module is applied (Chapter 5.1), creating a speaker model. For the purpose of our work GMM-UBM model will be used. The last stage of the process is testing (Chapter 5.2) and making a decision based on likehood scoring.

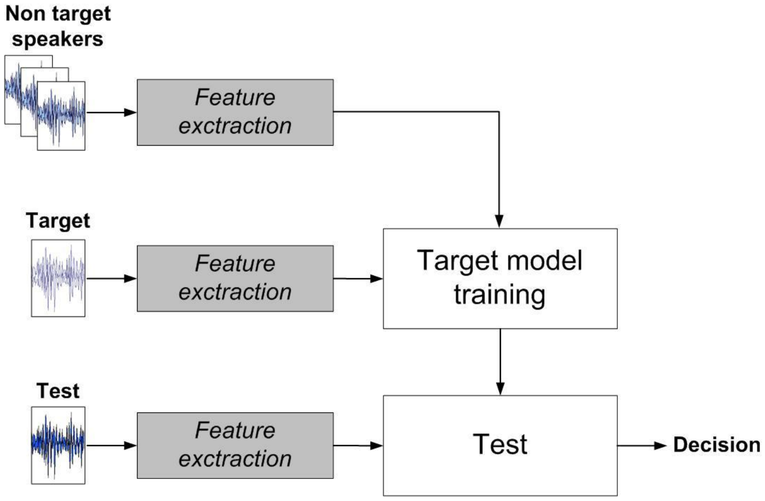

Figure 5.1: Speaker verification system

According to [BIM04] the speaker verification system process can generally be described in two phases only:

- Training phase.
- Test phase.

5.1 Training phase

Main stages of a training phase are shown in Figure 5.2. In the first step the speech parameters are extracted from the recording. Speech parameterization module follows, creating feature vectors. Last, a statistical model of the speaker is created. Using the same procedure, models of other speakers as well as the UBM model are created.

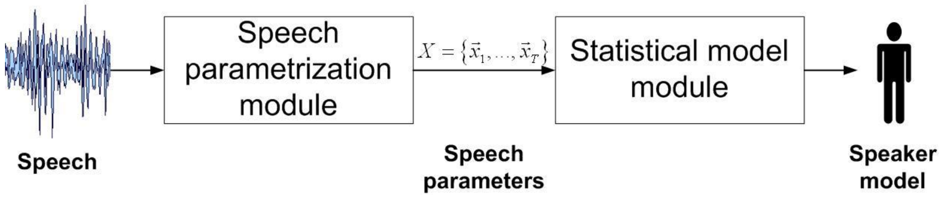
Speech
Speech parametrization module
$X = \{\vec{x}_1, \ldots, \vec{x}_T\}$
Speech parameters
Statistical model module
Speaker model

Figure 5.2: Training phase, basic stages

5.2 Test phase

Main stages of the test phase are shown in Figure 5.3 . The main block of this phase is the normalization decision stage. It has three inputs. The first is the parameterization of Speech module, which produces feature vectors from the input speech signal. Its function is the same as in the training phase, but the identity of the speaker is not known. The other two inputs are statistical models of both UBM and speaker, whose identity should be verified. These models are based on the parameters calculated during the training phase.

Scoring normalization decision stage then follows, computing the *decision scores*, normalizing them, and making a decision: *Accept* or *reject* that claimed identity belongs to the tested speaker.

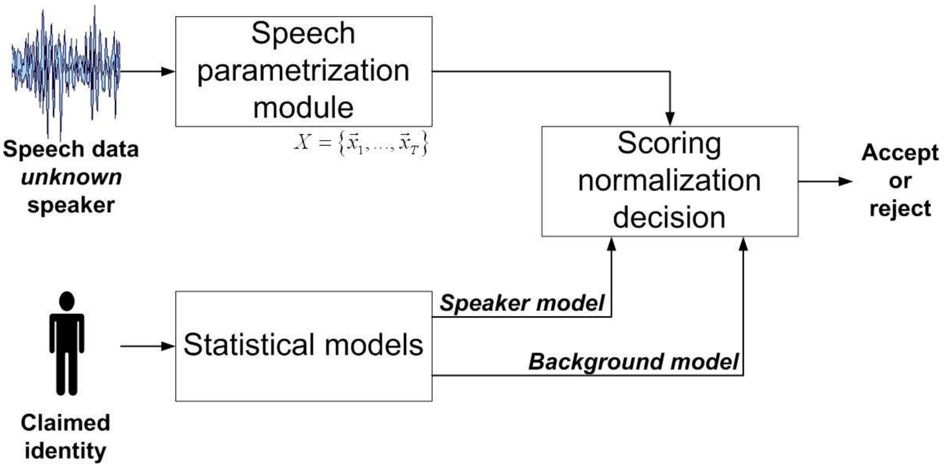
Speech parametrization module
$X = \{\vec{x}_1, \ldots, \vec{x}_T\}$
Speech data *unknown* speaker
Scoring normalization decision
Accept or reject
Speaker model
Statistical models
Background model
Claimed identity

Figure 5.3: Test phase, basic stages

5.3 Speaker verification system GMM-UBM

Using GMM-UBM system for Speaker verification task is described in [REY00], [REY95], [NAJ09]. The Gaussian mixture models (GMMs) become dominant for Speaker verification applications based on probability models. The used system is referred to as the *Gaussian Mixture Model-Universal Background Model* speaker verification system (GMM-UBM) [REY95].

The speaker verification task definition is *to determine if speech* Y *was spoken by speaker* S. If we suppose that *Y* contains speech of only one speaker then we name the task as single-speaker verification. If not, the task becomes to multi-speaker detection. With regard to the objectives of our work we define task of bird individual automatic recognition: *To determine if song* Y *was sung by a bird individual* S. Used data in our work contains song of only one bird at one time, so it fulfils the condition of single bird verification.

If we define two possible conclusions of a single-speaker verification

H_0…Y was spoken/sung by a *S*
H_1…Y was not spoken/sung by a *S*

then the goal of a task is to determine both probabilities

$$p(Y|H_0) \tag{5.1}$$

$$p(Y|H_1). \tag{5.2}$$

These are called a probability of a hypotheses H_O and H_1 respectively. Finalization of a process is then defined by authors as decision:

Probability ratio	Threshold	Decision
$\frac{p(Y\|H_O)}{p(Y\|H_1)}$	$\geq \theta$	H_0*...accept* Y *was spoken/sung by a* S
	$< \theta$	H_0*...reject* Y *was not spoken /sung by a* S

Table 5.1: Decision based on threshold

The basic goal of a system is to determine techniques to compute values for the two probabilities [REY00]. Basic stages of a speaker verification system are shown in Figure 5.4.

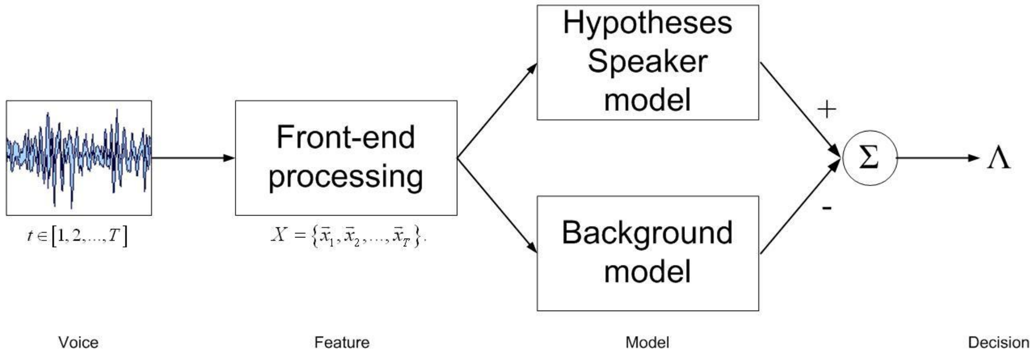

Figure 5.4: Probability ratio-based speaker detection system [REY00]

Input stage contains a speech (song) of length *t*. Front-end processing stage extracts feature vectors which contain speaker-dependant information

$$X = \{\vec{x}_1, \vec{x}_2, ..., \vec{x}_T\}. \tag{5.3}$$

T is the number of feature vectors. It depends on the length of a speech/song and the length of a frame which divides the speech into segments. Computing the probabilities H_0 and H_1 follows, based on the feature vectors.

Probability H_0 is mathematically represented by a model λ of the speaker. In a different way these probabilities represent the hypotheses

Hypothesis	Model	Denote
H_0	λ_{hyp}	*Hypothesized speaker* S *is in the feature space of* x
H_1	$\lambda_{\overline{hyp}}$	*Alternative to* H_0

Table 5.2: Two models of speakers

The basic assumption is that the Gaussian distribution *best represents* the distribution of the feature vectors for H_0. The model would be than be denoted by a mean vector and a covariance matrix of the Gaussian distribution [REY00]. Probability ratio statistics of both models can be expressed as

$$\frac{p(X|\lambda_{hyp})}{p(X|\lambda_{\overline{hyp}})} \tag{5.4}$$

With the logarithm of this statistically given log-probability ratio during the last stage is

$$\Lambda(X) = \log p(X|\lambda_{hyp}) - \log p(X|\lambda_{\overline{hyp}}). \tag{5.5}$$

The model λ_{hyp} of H_0 hypothesis will be well established using training speech from *S*. On the contrary, model $\lambda_{\overline{hyp}}$ must include all of the possible alternative hypotheses covered by H_1. To estimate model $\lambda_{\overline{hyp}}$ two approaches come into consideration:

- Use models of all others speakers in order to cover all alternative hypotheses. The approach is known as *background speakers (BS).*
- Merging speeches from several speakers to train a single model. The approach is known as *universal background model (UBM).*

Research of *UBM* shows the advantage of the only one λ_{ubm} model that has to be trained for all tasks with particular background. In contrast, *BS* needs to use speaker-specific background model i.e. one individual for every speaker.

5.4 Probability model methods

The probability approach is based on creating models λ of all speakers and UBM. These models replace the original parameters by specific functions. Then these models are used to compare with model of an unknown speaker instead of comparing the parameters i.e. feature vectors. By comparison between models, search for the one which corresponds to the highest probability of the model of an unknown speaker, is conducted.

Creating the best usable models λ is crucial for the probability function $p(X|\lambda)$.

For the *text-dependent* tasks where prior knowledge of what the speaker says is used as the basic probability function of the hidden Markov model (HMMs).

For *text-independent* speaker recognition tasks i.e. no prior knowledge of what a speaker says (no matter what he or she says, no matter what a bird sings) uses Gaussian mixture models as the most successful probability function.

5.4.1 Hidden Markov models

Hidden Markov models (HMMs) can incorporate additional temporal principles as pre-build lexicon, orthography, language rules, etc. Basic assumption is that the system can be situated just in one state for every particular time segment. The status then changes step by step in time into other well-defined states. Transitions between states are described by probabilities a_{ij}. Assuming that we exactly know both the number of states and transitions between them, we can define Markov model as

$$\lambda^s = \{\boldsymbol{A}^s, \boldsymbol{b}^s\} \tag{5.6}$$

where $\boldsymbol{A}^s = [a^s_{ij}]$ is a probability matrix of transitions for speaker *s*, and $\boldsymbol{b}^s = [b^s_j(.)]$ is a vector of speaker *s* output probabilities [PSU06].

For text-dependent systems left-right models are used as the transitions are dependent on the linguistic content. The number of states is usually between 3 and 6. In the text-independent systems Markov ergodic models are used instead, where the number of states usually is 3 or 5.

One uses forward-backward algorithm or Viterby algorithm to calculate probability of passage through all states.

5.4.2 Gaussian mixture models

Let's suppose a *D*-dimensional feature vector, ***x***. Then the probability function $p(\boldsymbol{x}|\lambda)$ is defined as a linear combination of weights w_i and unimodal Gaussian densities $p_i(\boldsymbol{x})$

$$p(x|\lambda) = \sum_{i=1}^{M} w_i p_i(x) \tag{5.7}$$

where weights satisfy

$$\sum_{i=1}^{M} w_i = 1 \tag{5.7}$$

The *M* is the number of mixture densities, $p_i(x)$, defined as

$$p_i(x) = \frac{1}{(2\pi)^{\frac{D}{2}} |\Sigma_i|^{\frac{1}{2}}} \exp\left\{ -\frac{(x-\mu_i)'}{\Sigma_i}(x-\mu_i) \right\} \tag{5.9}$$

and each density is parametrized by mean μ_i … *D x 1* vector and Σ_i covariance matrix … *D x D* matrix, or rather diagonal covariance matrix is used. Model parameters of a particular speaker ***S*** are denoted as

$$\lambda_S = \{w_i, \mu_i, \Sigma_i\}, \quad i = 1,..,M. \tag{5.10}$$

where *M* denotes the number of parameters. For successful use of GMM determination of an appropriate number of Gaussian mixtures as well as to estimate their positioning is of key importance.

When the speach is divided into time segements *t={1,..,T}*, the feature vector is defined as

$$X = \{\vec{x}_1, \vec{x}_2, ..., \vec{x}_T\}. \tag{5.11}$$

Then *M* values are extracted for each segment

$$\begin{aligned}\vec{x}_1 &= \{x_{11}, x_{12}, ..., x_{1M}\} \\ \vec{x}_2 &= \{x_{21}, x_{22}, ..., x_{2M}\} \\ &... \\ \vec{x}_T &= \{x_{T1}, x_{T2}, .., x_{TM}\}\end{aligned} \tag{5.12}$$

The process of feature vectors extraction is drawn in Figure 5.5. From the speech, particular segments are selected by windowing. *M* coefficients are calculated for each segment. If the speech is divided into *T* segments, each feature vector consists of *T* vectors **x**, where each has the dimension of *M*.

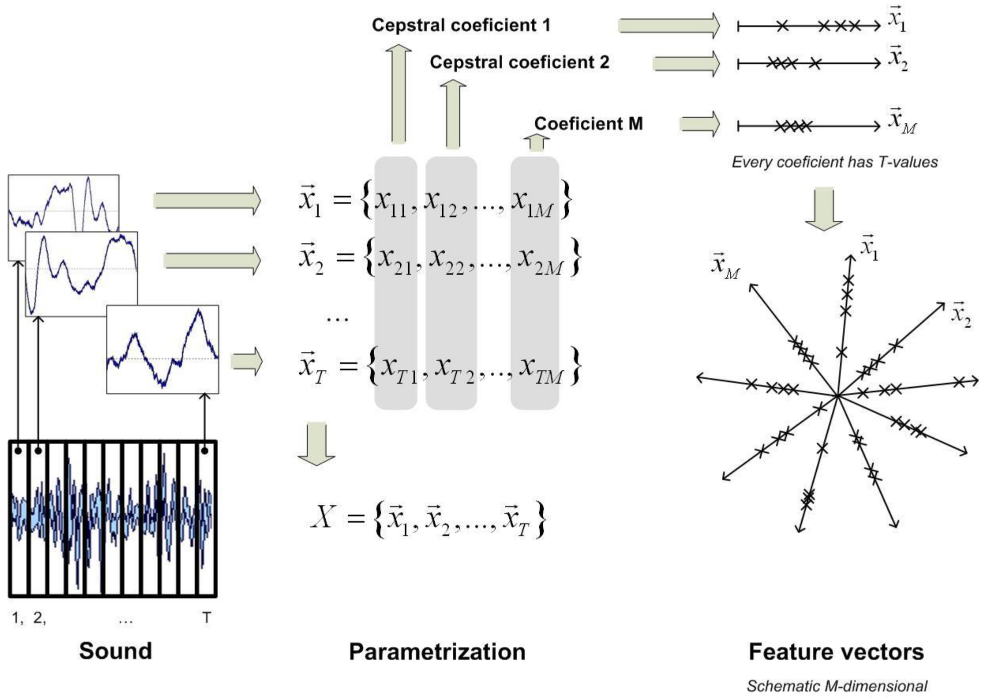
Cepstral coeficient 1
Cepstral coeficient 2
Coeficient M
$\vec{x}_1$
$\vec{x}_2$
$\vec{x}_M$
Every coeficient has T-values
$\vec{x}_1 = \{x_{11}, x_{12}, \ldots, x_{1M}\}$
$\vec{x}_2 = \{x_{21}, x_{22}, \ldots, x_{2M}\}$
...
$\vec{x}_T = \{x_{T1}, x_{T2}, \ldots, x_{TM}\}$
$X = \{\vec{x}_1, \vec{x}_2, \ldots, \vec{x}_T\}$
1, 2, ... T
Sound
Parametrization
Feature vectors
Schematic M-dimensional

Figure 5.5: M-dimensional Gaussian mixture model, schematic

5.5 Expectation-maximization EM

GMM parameters are better estimated using the expectation-maximization (EM) algorithm. It iteratively recalculates GMM parameters so that for every (*k+1*) step of iterations, following statement is valid:

$$p(X|\lambda^{(k+1)}) \geq p(X|\lambda^{(k)}) . \tag{5.13}$$

The overall probability model is the product of individual probabilities for individual frames for the time segments *t=0,..,T*

$$P(X|\lambda) = \prod_{i=1}^{T} P(x_i|\lambda) \tag{5.14}$$

EM algorithm maximizes the function for the optimized model $\hat{\lambda}$, which is given by

$$\hat{\lambda} = \arg\max_{\lambda} \prod_{i=1}^{T} P(x_i|\lambda) = \arg\max_{\lambda} \sum_{i=1}^{T} \log P(x_i|\lambda) \tag{5.15}$$

Five steps are defined as the sufficient number of iterations [REY00]. Notice that calculation of UBM parameters is similar to the parameters for Speaker only with additional use of Bayesian. For GMM training it is not necessary to use full number of Gaussian mixtures at the beginning, but the number can be increased stepwise. The process of adapting the EM model to a certain speaker is called adaptation.

5.6 Parameter extraction

The input signal is divided into individual segments by sliding window. Feature vector is calculated for every particular segment. For the purpose of our work the Hamming window is used. Although the Figure 5.5 simplifies that there is no overlapping between windows, in practice, overlapping is used because of better suppressed influence of the side bands. As appropriate parameters for the detection of *Chiffchaff* we use 30 [ms] length windows, with a step of 15 [ms]. Feature vectors for particular segments form

$$X = \{x_1, x_2, ..., x_T\}, \tag{5.16}$$

where $t = \{0,1,\dots,T\}$.

Every feature vector consists of *M* parameters

$$x_1 = \{x_{11}, x_{12}, ..., x_{1M}\} \tag{5.17}$$

where *M* denotes the total number of parameters. In addition to MFCC parameters several others are usually caltulated, such as energy parameters, etc. For the purpose of our work we use M ≥ 20 that consist of MFCC, Delta and energy parameters. Over time *t*, P is the total number of parameters resulting from

$$\mathbf{P} = M \times T. \tag{5.18}$$

Most important methods used for speech parameterisation are Mel Frequency Cepstral Coefficients (MFCC) and Linear Prediction Coding (LPC) [NAI09]. Both have good computable efficiency and give short-term parameters calculated from a quasi-static signal extracted from a sliding window.

6 Speaker verification system PVM

6.1 Overview

For purposes of this work we use tool called *PVM*. The system was developed by Aleš Padrta and Jan Vaňek from the *Department of Cybernetics*, *Faculty of Applied Sciences* in Pilsen. Currently, the system is being further developed and maintained by Lukas Machlica. The software was written in C++ and is used for solving the speaker verification tasks.

6.2 Process flow

Verification task is divided into four stages:

1. PRM. Parameterization of all input files.
2. UBM. Creation of an UBM model.
3. GMM. Adaptation of UBM/GMM models.
4. VERIFY. Test phase of a speaker verification task.

Flow diagram of PVM processes is shown in Figure 6.1.

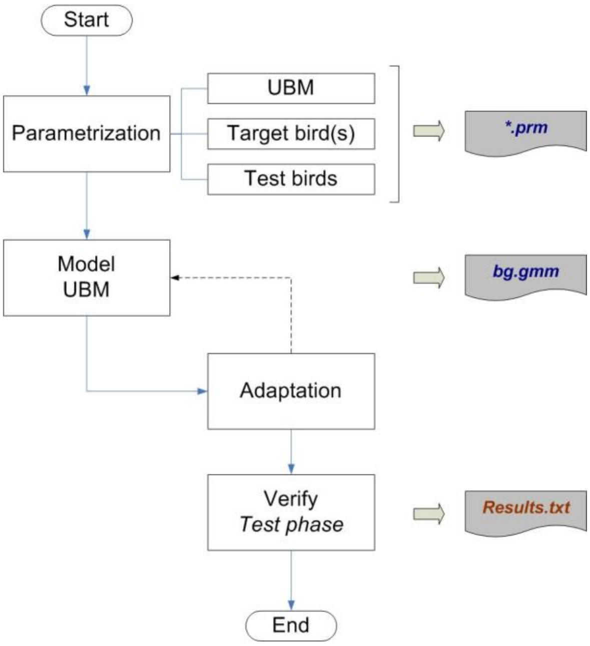

Figure 6.1: Function of PVM, flow diagram

In the first stage of *parameterization* corresponding parameters are extracted for every input **.wav* file. These parameters are saved into **.prm* files. In the next stage *ModelUBM*, a model of UBM is computed, and saved as *bg.gmm* file. Then follows an UBM/GMM model adaptation, based on incoming data. At the last stage, named *verification*, defined pairs of songs are tested against each other. The probability of match for every couple is computed and the results are recorded into *result.txt*.

6.3 Input and output data

Table 6.1 summarizes input and output data of all process stages. The recorded *.wav files, listed in configuration files, are entered into the *parameterization* PRM stage. Further stages use data created here, and finally the last stage, *verify*, writes results into the *results.txt* file.

Process	Input directory	Input files	Output directory	Output files
PRM	WAV\	*.wav	UBM_DIR\ GMM_DIR\ TEST_DIR\	*.prm
UBM	PRM\UBM_DIR\	*.prm	UBM\	bg.gmm
GMM	PRM\GMM_DIR\	*.prm	models\	*.gmm
VERIFY	PRM\TEST_DIR\	*.prm	VERIFY\	**result.txt**

Table 6.1: Speaker verification system PVM, inputs and outputs

6.4 PVM results

The PVM computes the probability of song/speech pair similarity. The probabilities are written into the file **results.txt**. Table 6.2 shows PVM output data. In the first column *result.txt* probabilities are copied from the result.txt file. In the second column *tested couple* couples of tested songs are placed, where the letter represents a bird individual while the number labels particular bird record. Last column describes the decision made by the supposed threshold $\Theta=0$, see chapter 7.1. If the value of threshold is lower than 0, the result is *rejected* and vice versa.

content of file **result.txt**	tested couple	result description
-2.177158	A01-B01	**reject**
0.49978	A01-A21	**accept**
2.836717	A02-A07	**accept**
1.461095	A02-A22	**accept**
-2.14189	A01-B06	**reject**
-0.012654	A01-A22	**reject** Error: False reject
-3.909118	A02-B08	**reject**
0.328327	A02-D03	**accept** Error: False accept
-4.295674	A01-C07	**reject**
4.007644	A01-A04	**accept**

Table 6.2: Example of content of output file *results.txt*.

In Table 6.2 for the column „tested couple, the letter represents bird individual, the number nr. of a particular bird's record, for instance:

A01… bird individual A, record number #01

A02… bird individual A, record number #02

C22… bird individual C, record number #22

6.5 Using the PVM

The PVM doesn't require an installation. It is ready for both 32-bit and 64-bit OS. The system is running on Windows Vista and Windows 7. Before starting the application it is necessary to set up configuration parameters and to determine the input files. For details see Table 6.3. The PVM runs in CMD (Window Command Line). Programme continuously prompts status as well as progress of current operation.

File	Content
Param_KW.ini	Set up the parameters of parameter vectors extraction. *For instance length of window, overlapping, number of MFCC parameters, switch on/off the pre-emphasis, etc.*
filelist_test	List of *.wav files to be tested. *The files will be compared with trained speaker/bird (Target) during Verify/Test stage.*
filelist_train	List of *.wav files to be system trained for. The Target speaker/bird. *The files will be compared with trained speaker/bird (Target) during Verify/Test stage.*
filelist_ubm	List of *.wav files from which UBM's model will be created.
Model_KWGMM.ini	Set up the parameters for an UBM model creating process. *For instance Number of Gaussians, etc.*
Model_ADAPT.ini	Set up the parameters for a target speaker/bird model creating process. *For instance type of adapting (MAP, MLLR,…) , etc.*
Verify.ini	Set up the parameters of Verification (Test) process. *For instance Threshold, format of results written in* results.txt*, etc.*
Trials.ndx	List of pairs of files which will be compared together. *For every couple is computed final probability of similarity in to* results.txt *file.*
Results.txt	List of computed probabilities for every tested pairs.

Table 6.3: PVM, configuration of the speaker/bird verification

7 Experiments

Only the *first* set (year 2011) of records has been used so far. DET curve (Chapter 7.1) was selected as the measurement of successful methodology of the system accuracy. Achieved results are encouraging. If suitable parameters are found, ERR could be depressed below 12%. The aim of the experiments is to achieve automatic recognition of birds by using the PVM system. All data sets will be used in the future. The main goal is to identify avian individuals using records of *chiffchaff* not just over months but over years.

First the evaluation methodology is described, followed by methodology of data recording and organization into three sets description. Last subchapter deals with some examples of achieved results.

7.1 Evaluation methodology

Four different situations may occur during the Speaker verification task, see Figure 7.1.

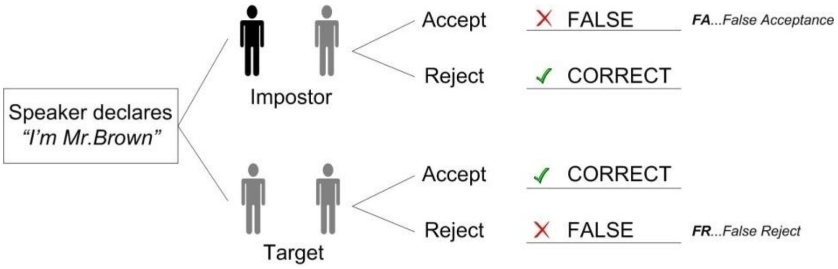

Figure 7.1: Speaker verification task: False and correct decision

For successful description of the system a special type of errors was introduced. The first is *False acceptation* $R_{FA}(\Theta)$ which expresses an average number of false acceptations. The second is called *False rejection* $R_{FR}(\Theta)$. It counts for an average number of false rejections [PSU06].

Incorrect acceptance error $R_{FA}(\Theta)$ is defined as

$$R_{FA}(\Theta) = \frac{n_{FA}(\Theta)}{n_{IM}}, \tag{7.1}$$

where n_{FA} is the number of cases when the system incorrectly accepts impostor and n_{IM} is the total number of cases where an impostor has been tested.

Incorrect rejection error $R_{FR}(\Theta)$ is defined as

$$R_{FR}(\Theta) = \frac{n_{FR}(\Theta)}{n_{TRGT}}, \tag{7.2}$$

where n_{FR} is the number of cases when the system incorrectly rejected the Target (right speaker/bird) and n_{TRGT} is the total number of cases when the target was incorrectly rejected.

Setting the threshold Θ affects the total number of R_{FA} and R_{FR}. Increasing the threshold reduces the false acceptance error rate *FA*, but it simultaneously increases the false rejection *FR* error. This happens because the system requires a higher probability of similarity. In contrast, if the threshold is lower, the *FR* error decreases, but the *FA* increases as the system needs lower probability of similarity to get accepted. This leverage effect is summarized in Table 7.1. Both errors are called *operating point* [PSU06].

Highest Θ	$R_{FA}(\Theta)$	*decrease*
	$R_{FR}(\Theta)$	*increase*
Lowest Θ	$R_{FA}(\Theta)$	*increase*
	$R_{FR}(\Theta)$	*decrease*

Table 7.1: Level of threshold Θ value and error rates

Equal Error Rate (*EER*) is used for single number evaluation of the system, which indicates the threshold value Θ_{EER} at which are both equal. It is defined as

$$R_{EER} = R_{FR}\left(\Theta_{EER}\right) = R_{FA}\left(\Theta_{EER}\right) . \tag{7.3}$$

In real experiments, however, a threshold Θ must be set first, where after the decisions and R_{FA} and R_{FR} errors can be calculated. Finding the threshold Θ_{EER} can therefore be very difficult.

For quoting the system success rate by one number the curve DET (*Detection Error Trade-off Curve*) is used. Error rates are here plotted as a function of the threshold [BIM04]. The advantage of the DET curve is a good readability especially for low differences between the errors. Another advantage is good distinction when more curves are plotted at once. This is useful when different system parameters are set and optimal values are sought.

7.2 Recording

The bird recording starts at about 5AM. The ornithologist takes a place at a suitable distance from the bird. He or she turns the microphone at an avian and starts up the recording. The recording then runs without interruption for several minutes. Typical recording time is between 3 and 15 continuous minutes. It is usual that during recording which takes a longer time a bird flies away to another location. The ornithologist then has to walk (as slowly and quietly as possible) with a microphone following the bird to stay closer.

Used records were made at a clearing on the border of a county town České Budějovice. Time of recording was usually between 5AM and 8AM.

Because of all that facts the main features of the recording are:

- *Noise*. The records contain many noises. The sound of traffic, wail of engines, occasionally a sound of a military jet.
- *Sound of forest background.* Although recorded near a town the background of record sounds like forest. The records contain sounds of other birds, insects, woodpecker’s pecking etc.

- *Not only one bird may sing at the same time*. Because the recording takes place in a clearing where many nests of chiffchaff are located, the records may sometimes contain two or more birds singing at one time. However, when more birds sing at precisely the same time the vocal of the target bird is usually much louder than of unwanted birds.
- *Variable level of a bird song record*. When the bird changes location the level of the recorded song varies.
- *Unusual noises.* Sometimes an unusual noise is recorded. For instance a wood crackling when ornithologist moves to remain near a bird. Sometimes it is an ornithologist speech when recording geographic data by voice.

7.3 Data sets

The experiment uses data set of bird individuals Chiffchaff (*Phylloscopus collybita (Lat.), Budníček menší (Czech*) and Willow Warbler (*Phylloscopus trochilus (Lat.), Budníček větší (Czech*). The records were made by ornithologists of the Faculty of science of University of South Bohemia, led by Pavel Linhart.

The data/records are divided into a three data sets which were named as follows:

1. Chiffchaff_2011
   Contains records from spring 2011. The data has very good quality despite background noise.
2. Chiffchaff_08-10
   Contains records from between years 2008 and 2010. The data has not very good quality because of heavy background noise.
3. Chiffchaff_2012
   Data not recorded yet. Expected data quality similar to set Chiffchaff_2011

Detail parameters of recorded songs collected in data sets 1 and 2 are listed in Table 7.2. and Table 7.3.

Sampling frequency	originals were recorded at 44,1 [kHz] for PVM system are used records oversampled at 22.050 [kHz]
Distance during recording	5 to 20 [m]
Background noise	30 to 70 [dB] *high level of noise occur occasionally only (acceleration of trolleybus, jet,)*
Number of records per a male	4 to 22. *In average 9 records per individual*
Length of record	30 [s] to 15 [min] *in average 6 [min]*
Number of songs per one record	5 to 30
Number of records in total	132x
Data size	2,7 GB (22,5 kHz)
Overall quality of the records in data set (clarity and distinctiveness of the songs)	**Very good**

Table 7.2: Data set a „*Chiffchaff_2011*“

Detail parameters of recorded songs collected in data set „Chiffchaff_2001“ are listed in Table 7.2.

Sampling frequency	originals were recorded at 44,1 [kHz] for PVM system are used records oversampled at 22.050 [kHz]
Distance during recording	5 to 20 [m]
Background noise	30 to 70 [dB] *high level of noise occur occasionally only (acceleration of trolleybus, jet,)*
Number of records per a male	4 to 22. *In average 9 records per individual*
Length of record	1 [s] to 15 [min] *in average 6 [min]*
Number of songs per one record	5 to 30
Number of records in total	132x
Data size	2,7 GB (22,5 kHz)
Overall quality of the records in data set (clarity and distinctiveness of the songs)	Good (worse than in Set 1)

Table 7.3: Data set 2 „*Chiffchaff_08-10*“

Both tables describe basic data of the records transferred by ornithologist to PVM. No editing was performed.

All data was recorded in terrain continuously under real circumstances, it is noisy, more than one bird may sings in time, distance between ornithologist and birds is changing during recording.

We can use the data in a three ways:

1. *Continuous record*. Whole recording is used in PVM with no cuts, noise cancelation , etc., see Figure 7.2

2. *Single songs*. Single songs are cut out from the recordings, see Figure 7.3.

3. *Combination of both*, single songs and continuous records.

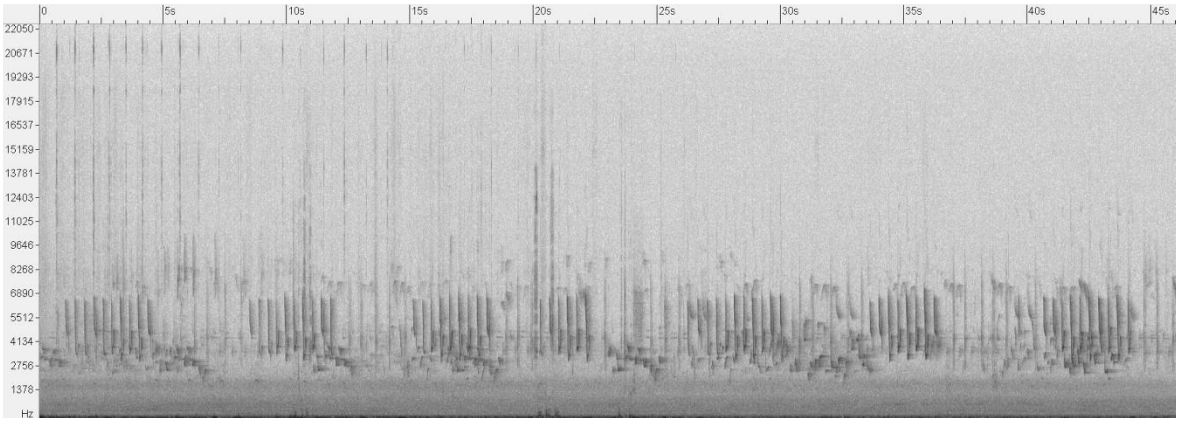

Figure 7.2: Continuous record, *Chiffchaff*

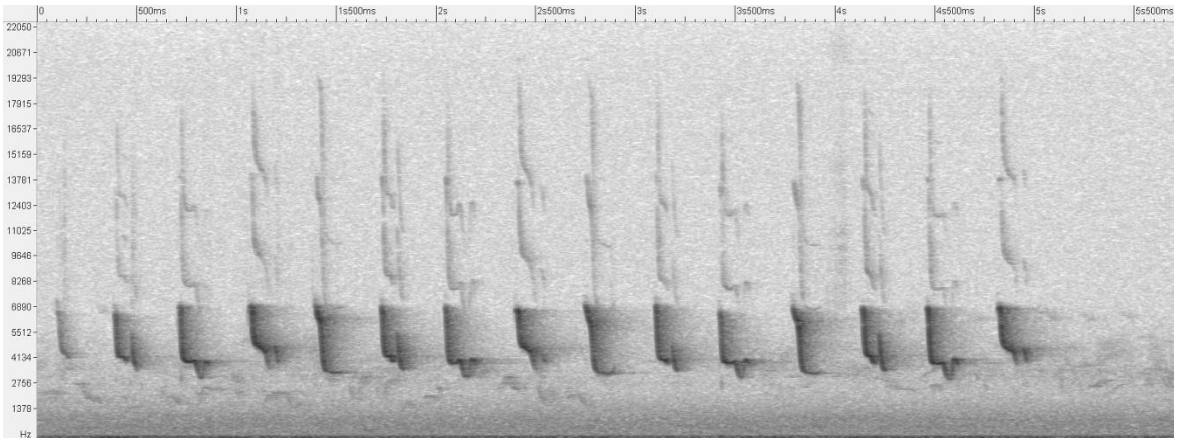

Figure 7.3: Single song, *Chiffchaff*

7.4 Initial results

For the first experimental phase there is a first set of data *Chiffchaf_2011* only. Moreover we used the continuous recording, with no edit (cut, split, clear etc.). Speaker/bird verification was performed. At the beginning only one bird was selected as Target one. Half of the targeted bird records were used for training the system. The rest serves for the verification/testing phase. One third of the total number of records was used for UBM model set-up. Remaining records were used to test the system. Duration of a task bird/speaker verification varied according to the number of records and their length. On average, the duration of one task was from five minutes to an hour. Preparation of the experiments took minutes to tens of minutes.

Final evaluations of an experiment are made on Excell spread-sheet and Matlab. In Excel there are ready-to-fill structures where data are inserted. Subsequently, the spread-sheet automatically determine the right decision (accept, reject) by file names. Correct results are compared with Result.txt, both FR errors and FA are calculated in Excel. The DET curve is plotted in Matlab.

For the future the plan is to automate the preparation of experiments, especially the creation of pairs, which is very time consuming. Also, the evaluation of experiments should be more automatic. Advantage could be taken from the appropriate linking of Excel and MATLAB for example.

7.4.1 Experiment example 1

Set up of the basic parameters :

Parameter	Value
Length of Hamming window [ms]	20
Shift of window [ms]	20
Number of filters	5

Table 7.4: Example 1, values of basic parameters

Number of compared pairs was 876. Achieved result EER=24,24 %

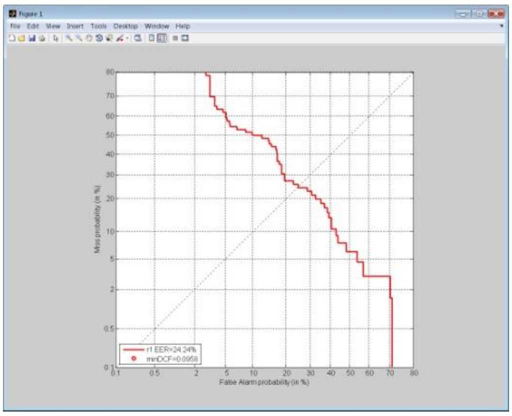

Figure 7.4: DET curve, example 2

Comment:

Length of a window is set up efficiently. However the number of filters is very low.

7.4.2 Experiment example 2

Set up of the basic parameters :

Parameter	Value
Length of Hamming window [ms]	30
Shift of window [ms]	30
Number of filters	20

Table 7.5: Example 1, values of basic parameters

Number of compared pairs was 876. Achieved result EER=13,64 %

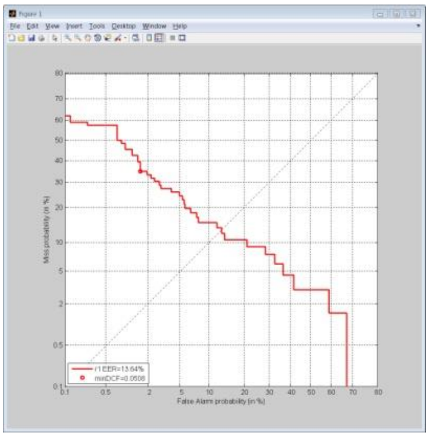

Figure 7.5: DET curve, example 1

Comment:

Length of a window is too long. Number of filters is set up better then in previous example. Achieved ERR of the PVM system is good for practical use.

7.4.3 Experiment example 3

Set up of the basic parameters :

Parameter	Value
Length of Hamming window [ms]	20
Shift of window [ms]	20
Number of filters	20

Table 7.6: Example 1, values of basic parameters

Number of compared pairs was 876. Achieved result EER=10,61 %

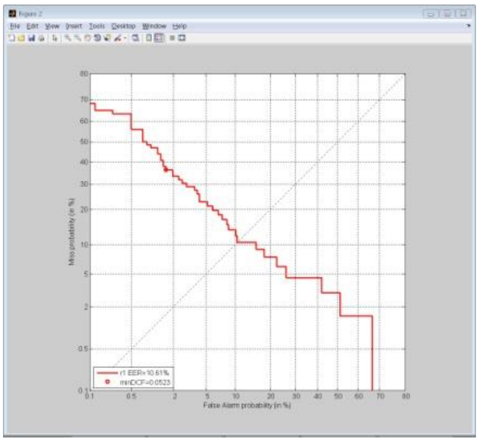

Figure 7.6: DET curve, example 3

Comment:

Both length of the window and the number of filters is set up correctly. The value EER of 10% is very good for practical use.

8 Conclusions

This thesis report outlines theory and framework for future work, dealing with research of using GMM/UBM for individual verification and identification of birds. The system PVM of *Department of Cybernetics, Faculty of Applied Sciences* in Pilsen is used as a tool. Three main goals can be underlined:

First, the practical experiments use real data which are recorded in the open air under common circumstances. The data is used raw as recorded. No editing process is used, for instance noise cancelation, cutting, gluing, equalization. Verification uses both continuous records and single songs.

Second, the experiments focus on recognizing bird individuals through months or even years.

Third, only a limited database of records exists for our disposal. In contrary to speaker verification a number of birds record is always limited. While creating a database of human voices is theoretically unlimited, the researcher needs only „time and money“, building up a database of songs of a particular bird is strictly limited. Recording depends on season, weather, bird mood and condition, accessibility of a nest, and on many more influences, at least on random.

Based on our knowledge this work is the first for bird verification and identification using real raw data. Tool PVM is used incorporated into the GMM/UBM system. Future work will develop methods outlined in this thesis report. The goal is to prove the possibility of bird recognition and to find a suitable method to supersede ringing by non-contact identification. Desired value of EER is 15% and lower.

Automatic bird identification offers a wide spectrum of application, for instance:

- Territory survey
  Researcher installs automatic record machines (start up when level exceed a limit) near nests. The recorders don't need an operator. He or she downloads records after a while then uses them for automatic recognition. At present, similar systems are used for night birds recording. However, humans are not able to recognize an individual from the records. The data serves only for confirmation if there is an owl for instance living in the place.
- Migration birds mapping
  Ornithologists from different countries could share the data. From these records a database of individual birds could be established after a precise bird GMM model register. Than every user of this register may verify if recorded bird is included in the register.
- Inaccessible breeding grounds observation
  After installation of an automatic recorder, data can be collected automatically. After a while ornithologist takes the records down.

Bibliography

[BIM04] F. Bimbot, J.F. Bonastre, C. Fredouille, G. Gravier, M.I. Chagnolleau, S. Meignier, T. Merlin, O.J. Garcia, P. Delacretaz, and D.A. Reynolds. A tutorial on text-independent speaker verification. EURASIP Journal on Applied Signal Processing, 4:430–451, 2004.

[BIS06] Bishop, Ch., M.: Pattern Recognition and Machine Learning, Springer, 2006

[CAT08] Catchpole, C., K., Slater, P., J., B.: Bird Song. Biological Themes and Variations, 2nd edition, Cambridge Press University, 2008

[CER01] Černocký, J.: Cepstrální analýza řeči a její aplikace, FIT VUT Brno, 2005

[CER03] Černocký, J.: Temporal processing for feature extraction in speech recognition, FIT VUT Brno, 2003

[CER04] Černocký, J.: Úvod do číslicového zpracování signálů, FIT VUT Brno, 2004

[CLE05] Clemins, P.,J.: Automatically classification of animal vocalizations, ProQuest, Ph.D. thesis, 2005

[FAG04] Fagerlund, S.: Acoustics and physical models of bird sounds , HUT, Laboratory of Acoustics and Audio Signal Processing, 2004

[FAG04b] Fagerlund, S.: Automatic Recognition of Bird Species by Their Sounds, Helsinki University of Technology, 2004

[FAG07] Fagerlund, S.: Bird species recognition using support vector machina, Helsinki University of Technology, 2007

[FOX08] Fox, E.,J.,S.: Call-independent identification in birds, School of Animal Biology, School of Computer Science and Software Engineering, University of Western Australia, Ph.D. Thesis, 2008

[GLE11] Ondrej Glembek, Lukas Burget, Pavel Matejka, Martin Karafiat, Patrick Kenny: Simplification and optimization of i-Vector extraction, ICASSP, Prague, Brno Universisty of Technology, 2011

[HAR03] Harma, A.: Automatic identification of birds species based on sinusoidal modeling of syllables, Helsinki University of Technology, 2003

[CHE10] Cheng, J., Yuehua, S., Liqiang, J.:A call-independent and automatic acoustic system for the individual recognition of animals: A novel model using four passerine, Laboratory of Animal Ecology and Conservation Biology, Institute of Zoology, Chinese Academy of Sciences, Beijing, China, 2010

[KUN10] Kuntoro, A., Johnson, M., Osiejuk, T.,: Acoustic censuing using automatic vocalization classification and identity recognition, J.Acoustic. Soc. Am, 2010

[MCH06] Machlica, L.: Ověřování totožnosti člověka z nahrávek jeho hlasu, Diplomová práce, Fakulta aplikovaných věd, KKY, Plzeň, 2006

[MOL08] Molnár, c., Kaplan, F., el al.: Classification of dog barks: a machine learning approach, Springer, 2008

[NAJ09] Najim, D.: Discriminative and generative approaches for long- and short-term speaker characteristics modeling: Application to speaker verification, Quebec University, Montreal, 2009

[OSI03] Osiejuk, T., S., Ratynska, K., Cygan, J., P., Dále, S.: Song structure and repertoire variation in ortolan bunting from izolated Norwegian population, Finnish Zoological and Botanical Publishing Board, 2003

[OSI05] Osiejuk, T., Trawicky, M., B., Johnson, M., T.: Automatic song-type classification and speaker identification of Norwegian Ortolan bunting, Marquette University, Milwaukee, 2005

[PSU06] Psutka, J., Muller, L., Matousek, J., Radova, V.: Mluvíme s počítačem česky, Academia, 2006

[REY00] Reynolds, D., A., Quatieri, T., F., Dunn, R., B.: Speaker Verification Using Adapted Gaussian Mixture Models, MIT Lincoln Laboratory, 2000

[REY99] Douglas A. Reynolds, Thomas F. Quatieri, and Robert B. Dunn: Speaker Verification Using Adapted Gaussian Mixture Models. M.I.T. Lincoln Laboratory, Massachusetts, 1999

[REY95] Reynolds, D. A. and Rose, R. C., Robust text-independent speaker identification using Gaussian mixture speaker models, IEEE Trans. Speech Audio Process. 3 (1995), 72–83

[SEL05] Selin, A.: Bird sound classification using wavelets, Tampere University of Technology, 2005

[SOM03] Somervuo, P., Harma, A.: Analyzing bird song syllables on the Self-Organizing Map, Helsinki University of Technology, 2003

[SOM04] Somervuo, P., Harma, A.: Bird song recognition based on syllable pair histograms, Helsinki University of Technology, 2004

[TAN03] Tanttu, T., Turunen, J., Ojanen, M.: Automatic Classification of Flight Calls of Crossbill Species (Loxia spp.) , Tampere University of Technology, 2003

[TEJ10] Tejkal, M.: Využití akustických metod při monitoringu sov (Strigiformes) v lesních ekosystémech, ČZU v Praze, 2010

[TRI08] Trifa et al.: Automated species recognition of antbirds in a mexican rainforest using hidden markov models, Helsinki University of Technology, 2008

[TUR05] Turunen, J., Selin, A., Tanttu, J., T., Lipping, T.: De-noising aspects in the context of feature extraction in automatic bird call recognition, Tampere University of Technology, 2005

[UHL07] Uhlíř, J., Sovka, P., Pollák, P., Hanžl, V., Čmejla, R.: Technologie hlasových komunikací, ČVUT, 2007

[VIL06] Vilches, E., Escobar, I., A., Vallejo, E.,E.: Data mining applied to acoustic bird species recognition, Campus Estado de México, 2006