HFCC BASED RECOGNITION OF BIRD SPECIES

Robert Wielgat [1], Tomasz P. Zieliński [2], Tomasz Potempa [1], Agnieszka Lisowska-Lis [1], Daniel Król [1]

[1] *Department of Technology, Higher State Vocational School in Tarnów, Tarnów, Poland*
[2] *Department of Telecommunications, AGH University of Science and Technology, Krakow, Poland*

ABSTRACT

Results from preliminary research on recognition of Polish birds' species are presented in the paper. Bird voices were recorded in a highly noised municipal environment. High 96 kHz sampling frequency has been used in order to record birds' voices. As a feature set standard mel-frequency cepstral coefficients (MFCC) and recently proposed human-factor cepstral coefficients (HFCC) parameters were selected. Superior performance of the HFCC features over MFCC ones has been observed. Proper limiting of the maximal frequency during HFCC feature extraction results in increasing accuracy of birds' species recognition. Good initial results are very promising for practical application of the methods described in the paper in monitoring of protected birds' area.

1. INTRODUCTION

An acoustical analysis of bird voices is important and dynamically developed aspect of the ornithological research. Identification of bird species can serve as an example. It is often used for educational and pedagogical purpose [3] but can also be applicable for biological and agricultural research and studies [4]. Specific application of the identification of bird species is detection of bird species in the monitored area. Such detection would be money- saving, and especially helping in bird monitoring actions that are work and time- consuming. In bird species monitoring and counting actions, that are usually organised by the national institutions (National Parks, others) and different organisations (national or private), there are engaged usually up to hundreds of people (necessary to be effective). They are undertaken in hardly- accessible areas and need intensive work of specialist educated and skilled to recognise birds by only vocalisation.

The frequency spectra of bird voices, besides sounds laying within human frequency range, include also some frequency components being infra or ultrasounds, which are not perceived by the human beings [5, 6]. The performed frequency analysis should be capable to deal with very fast signals and wide signal modulation. Some birds sometimes mix signals (voices) coming from both vocal cords (the unique anatomical construction for birds) when singing. These conditions make bird voices analysis difficult to some extent in application to bird species recognition.

Acoustical identification of bird species is usually based on methods used for human speech recognition [1, 2, 15]. As feature vectors typically Mel frequency cepstral coefficients (MFCC) are used [1, 15] while Hidden Markov Models (HMM) and Dynamic Time Warping (DTW) are the most popular classifiers[2, 15]. Recently, feature extraction methods based on spectral peaks tracking [2] and syllable pair histograms [16] have been introduced.

In the research described in the paper Human Factor Cepstral Coefficients (HFCC), proposed recently in the context of human speech [11], have been used for bird species recognition. The experiments were carried out on exemplary voices of several Polish wild birds. Some consideration concerning selection of sampling frequency and A/D converters are also presented. These problems are usually omitted in the literature. A performance comparison of two microphone types is presented as well.

2. PROBLEM STATEMENT

There are many technical problem encountered in bird species recognition. They concern mainly acquisition and feature extraction stage in the whole process of bird species recognition.

2.1 Acquisition of bird voices

The first problem is an acquisition of acoustic bird's signal. Sampling frequency is a very important parameter of signal acquisition that strongly determines further signal processing.

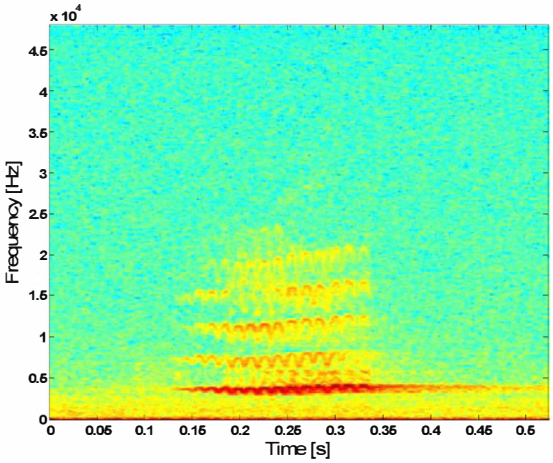

Figure 1 – Spectrogram of the exemplary Chaffinch (*Fringilla coelebs*) voice.

Many researchers carrying out the experiments with bird voices use sampling frequencies equal or less than 48 kHz [4, 17] assuming that spectrum of bird voices lays within human audibility range. Sampling frequency of 48 kHz is also justified by frequency response of the commonly used

microphones. Their frequency response strongly decreases above 20 kHz limit. Research presented in this paper takes another assumption that bird audibility range differs from the human one (usually is wider) and therefore higher sampling frequency is required. In our research 96 kHz sampling frequency has been chosen. The microphone used for recordings should have wider band as well, however in the described research hypercardioidal and cardioidal microphones having 20 kHz upper frequency limit have been used. A microphone with the broader band was not available for the authors.

The above statements are confirmed on figure 1, where *Fringilla coelebs* voice is presented. There is barely visible high 29 kHz component of the spectrum. Weak visibility is caused probably by microphone attenuation.

These problems indicate the need of using broadband microphones and higher sampling frequencies laying beyond 48 kHz limit during bird voices recording.

Another technical problem concerns analog-to-digital converters (ADC). In most audio recording systems sigma-delta ADCs are used. Sigma Delta conversion technology is based on oversampling, noise shaping and decimation filtering [7, 8, 9]. Advantages of sigma delta ADC are: high resolution (up to 24 bits) and high signal-to-noise ratio (SNR). However delta sigma technology of ADC makes possible a sampling frequency only upto 96kHz. In addition, frequency response of sigma delta converters is not flat and must be corrected by the pre-emphasis filters. Therefore their bits resolution is variable.

For capturing frequencies over 40 kHz, ADC with successive approximation register (SAR) is highly recommended. The bits resolution of SAR technology is up to 18 bits [10] and is constant in all bandwidth.

2.2 Feature extraction

Subsequent difficult problems resulting from acoustical environment are high degree of inherent noise in the signal and big diversity among identified bird species. At this point choosing appropriate signal features, robust to above specific conditions, is crucial for all the process of bird species recognition. In the presented research two types of feature vectors have been used (both are borrowed from speech recognition methods):

- Mel-Frequency Cepstral Coefficients (MFCC)
- Human-Factor Cepstral Coeffidients (HFCC)

2.1.1. Mel-frequency cepstral coefficients (MFCC)

The *mel-frequency cepstral coefficients* (MFCC), originated from modeling of acoustic signal processing performed in cochlea, were calculated in our experiments as follows:

1) blocking signal into frames and windowing by Hamming window;
2) application of the Fast Fourier Transform to the windowed frames;
3) packing the FFT power into the uniform, overlapping mel frequency bands with equally spaced center mel frequencies using triangular weighting in mel-scale (number of bands is a parameter of the algorithm); conversion from linear-frequency scale to the mel-frequency scale and vice versa is given by the equations:

$$f_{mel} = 2595\log_{10}(1 + f_{Hz}/700) \tag{1}$$

$$f_{Hz} = 700 \cdot \left(10^{f_{mel}/2595} - 1\right) \tag{2}$$

3) calculation of log spectral power coefficients in mel bands;
4) performing DCT on coefficients vectors ($n = 0, 1,..., q\text{-}1$):

$$X(n) = c(n)\sum_{k=0}^{K-1} \ln(S_k)\cos\left(\frac{\pi(2k+1)n}{2K}\right)$$
$$c(0) = \sqrt{\frac{1}{K}}, \quad c(n) = \sqrt{\frac{2}{K}} \quad \text{for } n = 1,..., q\text{-}1 \tag{3}$$

where: S_k – log spectral power coefficient in the k-th frequency band; K – number of frequency bands; q – number of MFCC coefficients;
5) calculating first and second derivatives of the MFCC coefficients in respect to time, so-called *delta* and *delta-delta* coefficients respectively.

2.1.2. Human-factor cepstral coefficients (HFCC)

The novel *human factor cepstral coefficients* (HFCC) approach to speech features extraction has been proposed and described in details in [11]. The method and its algorithmic implementation are very similar to the MFCC method described above. The only but crucial difference between these two methods is that now filter bandwidth is decoupled from filter spacing. In HFCC filter center frequencies are equally spaced in mel frequency scale (1), as in the MFCC method, but filter bandwidth is a design parameter, measured in equivalent rectangular bandwidth (ERB):

$$ERB = 6.23 f_c^2 + 93.39 f_c + 28.52 \quad \text{Hz} \tag{4}$$

where filter center frequency f_c is expressed in kHz. When wider filter bandwidth than ERB is exploited (ERB scaled by some factor > 1) then the HFCC-based speech recognition can be under some circumstances more resistant to noise.

It should be noted that both MFCC or HFCC, that have been used in the presented research, operate in the frequency range 0 ÷ 48 kHz which is much broader in comparison with speech recognition applications. This range however can be limited what gives some improvements in bird voice recognition accuracy (see section Results).

2.3 Classification Method

Dynamic time warping (DTW), as good speech classifier, is well known for many years [12]. It was used in reported research for its simplicity of implementation and analysis as well as for relatively high recognition accuracy comparable with HMM method.

The main idea in DTW algorithm used in speech recognition is finding an optimal path with minimal cost

from lower left corner to upper right corner of the local distance array (see fig. 2). A single element d_{mn} of the array equals to distance between m-th feature vector (MFCC or HFCC) of recognized bird voice and n-th feature vector (MFCC or HFCC) of the reference pattern. An Euclidean distance was used as a distance measure in the reported research.

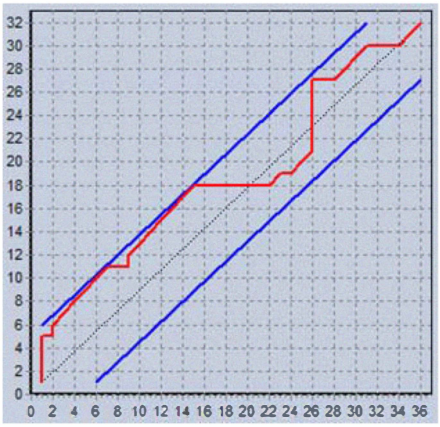

Figure 2 - An optimal search path restricted by two parallel lines in DTW algorithm

An accumulated distance at each point of the search path was calculated according to recursive procedure given by the equations:

$$g(i,j) = \min \begin{bmatrix} g(i-2,j-1)+d(i,j) \\ g(i-1,j-1)+d(i,j) \\ g(i-1,j-2)+d(i,j) \end{bmatrix} \quad (5)$$

In order to normalize obtained result the accumulated cost was divided by factor D:

$$D = \sqrt{N_w^2 + N_S^2} \quad (6)$$

where: N_W – number of feature vectors of the reference pattern, N_S – number of feature vectors of the word being recognized. The search path was limited by two parallel lines shifted by the coefficient:

$$Q = round(w \cdot \max(N_S, N_W)) \quad (7)$$

where: w – path width coefficient (equals 0.2 in the reported research). A detailed description of DTW algorithm can be found in [12], [13], [14]. An example of the DTW application to bird voices recognition can be found in [2], [15].

3. EXPERIMENTS

As an analyzed material recorded samples of sounds emitted by the followings birds have been taken:

- House Sparrow - *Passer domesticus*
- Common Swift - *Apus apus* (flush)
- Great Tit - *Parus major*
- Rock-dove - *Columba livia*
- Chaffinch - *Fringilla coelebs*

The birds were attracted (annoyed) with vocalization of previously recorded sounds of the analyzed bird species. The experimental recordings were accompanied by simultaneous camera recording. This visual material was of help to interpret the species and aspect of vocal soundings. Records were done in Tarnow, in the center of town (city park: *Parus major, Columba livia, Fringilla coelebes*, and high buildings: *Apus apus* (sounding of flush), *Passer domesticus*), in August 2006. This month is off-sounding-season, and the vocal emission examples as recorded, are rather poor and simple (not so much modulation, twitting, etc.). In the other way, they are better for the electro-acoustic analyze. The species that were of interest of recordings were exemplary of synantopic species.

A 24bit/96kHz sigma delta analog-to-digital converter and two capacitive microphones were used in experiments. The first microphone was of unidirectional hyper-cardioid type. The second one was cardioidal microphone. A unidirectional microphone is sensitive to sounds from only one direction. The most common unidirectional mike is a cardioid microphone, so named because the sensitivity pattern is heart-shaped. A hyper-cardioid microphone is similar but with a tighter area of front sensitivity and a tiny lobe of rear sensitivity.

Five birds species were recorded and analyzed in performed experiments. There were 10 examples of bird voice per one species in the training set recorded both by microphone of hipercardioidal type as well as cardioidal one. Because there was limited time of experiments and limited number of particular members of bird species, the number of birds voices per one species in testing set is different. The structure of the testing set is shown in Table 1.

Bird voice segments were extracted by standard signal detection based on energy threshold. Afterward detected signal segments were manually corrected. It should be noted that signal detection accuracy is not crucial for birds monitoring application.

Table 1. Structure of the testing set

Bird species	**Number of examples**
House Sparrow *(Passer domesticus)*	17
Common Swift *(Apus apus)*	64
Great Tit *(Parus major)*	25
Rock dove *(Columba livia)*	63
Chaffinch *(Fringilla coelebes)*	95

After the acquisition, signal was pre-emphasized with pre-emphasis coefficient value 0,9375. 30 ms length of frame signal and 20 ms overlapping have been chosen. Windowing by Hamming window has been applied to the framed signal. As feature vectors the MFCC and HFCC coefficients have been used. The values of the parameters of these features are presented in Table 2. Besides of standard MFCC and HFCC parameters a *maximal frequency* parameter is also added. This parameter denotes the final frequency of the last triangle filter in HFCC.

As a classification method the DTW procedure has been used. The experiments were carried out on the closed set of bird species. The recordings being disturbances (eg. Human speech, sound of bells, cars etc.) were not included in the training set as the separate class in order not to obscure this preliminary problem analysis. The recordings are however strongly noised by the sounds coming from environment.

Table 2. Parameters of feature extraction

Parameter	HFCC	MFCC
DFT length	4096, 8192	4096, 8192
Number of filters	32	32
Number of coeffs	15	15
ERB factor	2	-
Delta coeffs	yes	yes
Delta size	1	-
Delta-delta coeffs	no	no
Maximal frequency	20 kHz ÷ 48 kHz	48 kHz

4. RESULTS

In the first experiment 5 bird species have been recognized from the closed set by the MFCC and HFCC method using both hipercardioidal or unidirectional microphone. 4096-point DFT was used. Obtained results are presented in Tables 3÷6.

Looking at the achieved results it should be stated that the HFCC method gives higher recognition accuracy than the MFCC one. It stays with correlation with the former authors' experience with human speech recognition [18]. It is however interesting that HFCC being the features motivated by human audition physiology with frequency range up to 20 kHz can also give good results after stretching the frequency range to 48 kHz limit.

The overall bird species recognition accuracy is presented in table 7.

The problem of frequency range has been further investigated using only HFCC as signal features (since they were observed to be more efficient than MFCC). This time the final frequency of the last triangle filter in HFCC (*maximal frequency* parameter in table 2) was changed from 4 kHz to 48 kHz. Obtained results are shown on Fig.3. In this case DFT length was increased to 8192 points.

Table 3 – Confusion matrix of the bird species recognition for MFCC features (32 bands, 15 MFCC coefficients) and hipercardioidal microphone

	Rock-dove	Common Swift (flush)	Great Tit	House Sparrow	Chaf-finch
Rock-dove	63				
Common Swift (flush)	1	56		7	
Great Tit	1		23	1	
House Sparrow		8		9	
Chaffinch	2	4	3	1	85
Accuracy	**94.0%**	**82.4%**	**88.5%**	**50%**	**100%**

Table 4 – Confusion matrix of the bird species recognition for MFCC features (32 bands, 15 MFCC coefficients) and cardioidal microphone.

	Rock-dove	Common Swift (flush)	Great Tit	House Sparrow	Chaf-finch
Rock-dove	63				
Common Swift (flush)	1	57	1	5	
Great Tit	1		24		
House Sparrow		2		15	
Chaffinch	2	2	6		85
Accuracy	**94.0%**	**93.4%**	**77.4%**	**75%**	**100%**

Table 5 – Confusion matrix of the bird species recognition for HFCC features (32 bands, 15 HFCC coefficients) and hipercardioidal microphone.

	Rock-dove	Common Swift (flush)	Great Tit	House Spar-row	Chaf-finch
Rock-dove	63				
Common Swift (flush)	1	60		3	
Great Tit	1	1	23		
House Sparrow	1	4		12	
Chaffinch	3	3			85
Accuracy	**91.3%**	**88.2%**	**100%**	**80%**	**100%**

Table 6 – Confusion matrix of the bird species recognition for HFCC features (32 bands, 15 HFCC coefficients) and cardioidal microphone.

	Rock-dove	Common Swift (flush)	Great Tit	House Spar-row	Chaf-finch
Rock-dove	63				
Common Swift (flush)		55	2	7	
Great Tit	1		24		
House Sparrow		1		16	
Chaffinch	2		4		89
Accuracy	**95.5%**	**98.2%**	**80%**	**69.6%**	**100%**

Table 7 – Overall accuracy of the bird species recognition by different microphones and signal features

	Hipercardioidal microphone	Cardioidal microphone
MFCC	**82.97 %**	**87.98 %**
HFCC	**91.91 %**	**88.65 %**

As one can see from Fig. 3, decreasing the maximal frequency has led to increasing accuracy to the point close to the end of microphone pass-band. After this point recognition accuracy decreases again. Explanation of this phenomena can be following: decreasing maximal frequency limits the energy of noises of spectrum above upper boundary of the microphone frequency response. Because this noises bear nearly no useful information and cause disturbances in the classification process, their limitation causes increasing of the recognition accuracy. On the other hand, decreasing maximal frequency below the upper boundary of the microphone causes cutting off important information laying in the bird voices spectrum thereby decreasing overall recognition accuracy.

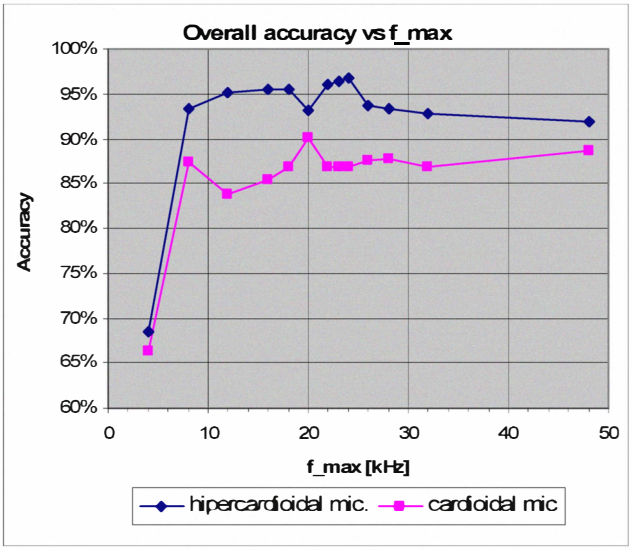

Figure 3 – Overall accuracy versus maximal frequency for two types of microphone used in the research.

Presented recognition results are relatively good but they were obtained in the closed set experiment without another real world sounds. For the open set experiment, worse results can be expected, but recommendation to the choice of the feature extraction and acquisition stages of the recognition process will be the same like in the research described above.

5. CONCLUSIONS AND FUTURE RESEARCH

The performed research can be concluded as follows.

- Sampling frequency above 48 kHz and broadband hipercardioidal microphones should be used for birds' voices recordings in order to analyse the signal properly. It seems to be contradictory to the results depicted on fig.3, where limiting the frequency band increases the recognition accuracy. It should be noted however that frequency range above 20 kHz is strongly attenuated by the frequency response of used microphones and for this range nearly no useful information is present in the bird voice signal. It could be precious to carry out the experiments using sampling frequencies and microphones which do not limit the frequency spectrum of bird voice signal. Such experiment conditions can be only achieved by using sampling frequencies above 48 kHz (for some bird species) and broadband microphones of the upper frequency boundary laying above 20 kHz limit (most common boundary for the majority of microphones).
- HFCC based bird species recognition gives better results in comparison with the MFCC method. Of course this statement is not statistically relevant and concerns only the testing set we examined in our research.
- Proper limiting of the frequency band in HFCC signal modelling can result in higher recognition accuracy.

Obtained results are very promising for building a computer-based bird monitoring system. Future research will include:

- performing experiments in open set with the presence of noise on the more numerous acoustical data,
- using more relevant signal filtering,
- using Hidden Markov models as a classification method.

REFERENCES

[1] S. Fagerlund, A. Harma, "Parametrization Of Inharmonic Bird Sounds For Automatic Recognition", *in Proc. EUSIPCO 2005*, Antalya, Turkey, September 4-8 2005.

[2] Z. Chen, R. C. Maher, "Semi-automatic classification of bird vocalizations using spectral peak tracs*", J. Acoust. Soc. Am.*, Vol. 120, No. 5, November 2006

[3] T. Oba, "Application of automated bioacoustic identification in environmental education and assessment", *Annals of the Brazilian Academy of Sciences*, (2004) 76(2), pp. 445-451.

[4] D. Chesmore, "Automated bioacoustic identification of species", *Annals of the Brazilian Academy of Sciences*, (2004) 76(2), pp. 435-440.

[5] A. Trepka, "Rekordy zwierząt – ptaki", *R.A.F Racibórz*, 1997, in Polish.

[6] K. Shmidt-Nielsen, „Fizjologia zwierząt – adaptacja do środowiska", *WN PWN*, Warszawa 1992, in Polish.

[7]R.M. Gray "Quantization noise in DeltaSigma A/D converters," Chapter 2 of Delta-Sigma Data Converters, edited by S. Norsworthy, R. Schreier, and G. Temes, *IEEE Press, 1997*, pp. 44-74.

[8] N. T. Thao, "Overview on a new approach to one-bit n-th order sigma-delta modulation," in *Proc. ISCAS, 2001*.

[9] A. Y. Kwentus, Z. Jiang, and A. N. Wilson, Jr. Application of filtersharpening to cascaded ingrator-comb decimation filters. *IEEE Transactions on Signal Processing*, 45(2):457-467, 1997.

[10] Analog Devices "18-Bit 500 kSPS PulSAR Unipolar ADC with Reference"*Analog Devices Data Sheet* 2003.

[11] M.D. Skowronski, J.G. Harris "Exploiting independent filter band-width of human factor cepstral coefficients in automatic speech recognition", *J. Acoust. Soc. Am.,* 116(3): 1774-1780, 2004.

[12] H. Sakoe, S. Chiba, "Dynamic Programming Algorithm Optimization for Spoken Word Recognition", *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-26: 43-49, Feb. 1978.

[13] L.R. Rabiner, A. Rosenberg, S. Levinson, "Considerations in Dynamic Time Warping Algorithms for Discrete Word Recognition", *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-26: 575-582, Dec. 1978.
[14] M.H. Kuhn, H.H. Tomaschewski, "Improvements in Isolated Word Recognition", *IEEE Trans. Acoust., Speech, Signal Processing*, 31(1): 157-167, 1983.
[15] J. A. Kogan, D. Margoliash, "Automated bird song recognition elements from continuous recordings using dynamic time warping and hidden Markov models: A comparative study", *J. Acoust. Soc. Am*. 103 (4), April 1998
[16] P. Somervuo, A. Härmä, "Bird Song Recognition Based on Syllable Pair Histograms", *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'04)* Montreal, Canada, May 17-21, 2004.
[17] A. Härmä, "Automatic identification of bird species based on sinusoidal modeling of syllables", *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, Vol. 5, pp. 545-548, 2003
[18] R. Wielgat, T. Zieliński, Ł. Hołda, D. Król, T. Woźniak, S. Grabias, " HFCC Based Pathological Speech Recognition", *Advances in Quantitative Laryngology*, Gronningen, Netherlands, Oct. 2006.