PARAMETRIZATION OF INHARMONIC BIRD SOUNDS FOR AUTOMATIC RECOGNITION

Seppo Fagerlund *

Laboratory of Acoustics and
Audio Signal Processing
Helsinki University of Technology (HUT)
Otakaari 5a, 02015, Espoo, Finland
phone: + (358) 9-451 6029, fax: + (358) 9-460 224
email: Seppo.Fagerlund@hut.fi

Aki Härmä

Philips Research
Prof. Holstlaan 4, WO-090, 5656 AA,
Eindhoven, The Netherlands

ABSTRACT

We have earlier found that the sinusoidal modeling and related parametrization is a promising technique for the automatic analysis and recognition of typical sounds produced by songbirds. In this article we study techniques that can be used to characterize sounds that cannot be efficiently parameterized using the sinusoidal model. Most familiar examples of such sounds are creaky sounds of Crows and many of the sounds produced, e.g., by Mallards. Often those sounds feature irregular pitch pattern. We introduce a method for feature reduction and optimal feature selection for recognition of bird species.

1. INTRODUCTION

The long-term objective in the current work is to develop feature extraction and classification methods for a system that could automatically recognize bird species by their sounds in field conditions. It has been demonstrated earlier that sounds of many songbirds are clearly tonal and can be efficiently modeled by one or a small number of time-varying sinusoidal components [1]. Nevertheless, songbirds regularly produce also sounds which have a complex spectrum and temporal envelope. In other than songbirds such cases are even more common. For example, the Common Raven rarely produces anything that fits to the sinusoidal signal model. In the current article we develop a more appropriate set of descriptive parameters for those sounds.

Bird sounds are divided by the function into songs and calls, which are further divided into hierarchical levels, which are phrase, syllable, and element or note [2]. Elements are smallest separable units of bird vocalization. In the simplest case syllable is constructed from one element but more complex syllables may include several elements. Phrase is a series of syllables that occur in a particular pattern. A phrase is often, but not always, a sequence of similar syllables.

Relatively little have been done previously to find efficient parametrization of bird sounds for recognition. For example, in [3, 4] bird sounds were represented by spectrograms of syllables or elements. Most of the earlier work on automatic recognition of birds have been related to the recognition of songs of birds [5, 6] or some restricted set of predefined sounds from one species [4]. In this work we test recognition bird species based on individual syllables. Nelson [6] noted that different species used different cues to recognize their own species. In this work we introduce a method to measure features importance for classification and try to find species-specific feature sets.

The recognition experiments in the current article are based on the bird song database collected in the Avesound project [7] at HUT. Audio files in the database contain songs, calls or series of calls mainly recorded in Finland. Individual *syllables* are extracted from songs using a segmentation algorithm based on the short-time signal energy and an adaptive estimate of the background noise. Feature vectors are then formed from various signal measures introduced below. Finally, syllables are then classified based on those representations.

In this article we first try to characterize what types of non-tonal sounds are common in avian vocalization. Secondly, we study the performance of several different computational measures that could be used as features in an automatic recognizer. We use low-level signal parameters such as the spectral centroid and signal bandwidth. These parameters have been used previously, for example, in general audio context classification [8], music genre classification [9], but, to our knowledge, have not been tested for bird sounds previously. For comparison we also test recognition with Mel-frequency cepstral coefficient (MFCC) representation of syllables. MFCC-model have been popular parametrization method in different types of audio recognition tasks, e.g. in automatic speech recognition [10].

2. THE CLASS OF INHARMONIC SOUNDS IN BIRDS

In [1], harmonic bird sounds were divided into four classes by the observed harmonic structure. Classes I and II were for pure sinusoidal and pure harmonic signals, respectively. Class III syllable has a harmonic structure such that the fundamental frequency component (F0) is heavily attenuated and, in the Class IV both F0 and F1 are weaker than F2. It was found in [1] that syllables that are not harmonic usually fell outside of the four classes or went to the harmonic class IV. In these cases likelihood of a syllable to belong to the pure sinusoidal class (class I) was also very small. In this article this observation has been turned into a criterion for selecting sounds that do not fit into the sinusoidal signal model. In particular, if the likelihood to belong to pure sinusoidal class is less than 60%, syllable is labelled to the class of *inharmonic* sounds. Note that the set of inharmonic sounds defined this way will contain many different types of sounds and some of those can also be considered harmonic.

Hooded Crow (*Corvus corone cornix*) is a good example

*Supported by the Academy of Finland (AveSound project) .

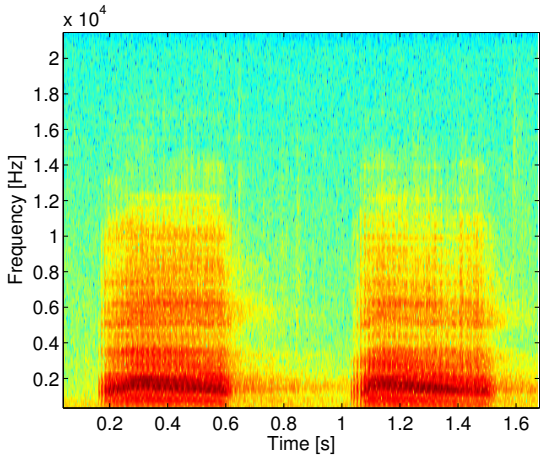

Figure 1: Sequence of two syllables of Hooded Crow (*Corvus corone cornix*, CORNIX).

Lat. Abbr.	Common name	Recs.	Inharmonic	Syllables
CORRAX	Common Raven	7	96%	91
CORNIX	Hooded Crow	8	98%	160
PICPIC	Magpie	7	99%	312
GARGLA	Eurasian Jay	9	99%	99
ACRSCH	Sedge Warbler	6	65%	331
ACRRIS	Marsh Warbler	8	34%	277

Table 1: Birds in the current work. Column are a widely used abbreviation derived from the Latin name, common English name, the number of recordings from different species, percentage of amount of inharmonic syllables and total number of inharmonic syllables.

of a bird species that produces inharmonic sounds. Based on to the criterion above 96% of syllables of Hooded Crow are labeled to this class and usually likelihood to belong to the pure sinusoidal class was only a few percents. The spectrogram of a typical song of Hooded Crow is shown in Figure 1.

Table 1 shows the set of species studied in this article. The vocalization of these species commonly contain different types of inharmonic sounds. The first four species are close relatives of Crow (*Corvidae* family) and the last two species belong to the *Acrocephalus* family of songbirds.

3. METHODS

The system for automatic recognition of syllables consists of three components. First, a recording containing bird sounds is segmented into syllables using the segmentation algorithm introduced in [1]. Then a set of parametric representations is computed from each syllable. The obtained feature vectors are divided into training and testing data sets from where the former set is used as models of the syllables in the classifier and the latter is used to test recognition ability of the classifier. In the current article, we compare two different ways of representing the sounds to the classifier.

3.1 Low-level descriptive parameters

In the first method each syllable is represented by 11 low-level acoustical parameters. Seven features are calculated on the frame basis. These provide a short-time description of the

Spectral features		
Feature	Abbreviation	Frame feat.
Spectral centroid	mSC, vSC	*
Signal bandwidth	mBW, vBW	*
Spectral roll-off frequency	mSRF, vSRF	*
Spectral flux	mSF, vSF	*
Spectral flatness	mSFM, vSFM	*
Frequency range	range1, range2	
Temporal features		
Zero crossing rate	mZCR, vZCR	*
Short time energy	mEN, vEN	*
Syllable duration	T	
Modulation spectrum	MSm, MSf	

Table 2: Descriptive parameters used in current study columns are the name and the abbreviations of the feature. Asterisk (*) in last column indicates that the feature is calculated on the frame basis.

syllable. Mean and variance values of these features are used as actual features of the classification system, thus we have 14 actual features calculated on frame basis and five more features that are calculated from the entire duration of a syllable. Features are divided into spectral (frequency domain) and temporal (time domain) features, according to their calculation domain.

For frame basis features, syllables are first divided into overlapping frames. In this work we use the frame size of 256 samples (6*ms*) with 50% overlap, thus step size when going from frame to another was 128 samples. Features are calculated for each windowed frame. Hanning window was used for the spectral features and rectangular window for the temporal features. Spectral features were calculated from Fourier-transformed signal frames. The final measures are the mean and variance values of the feature trajectories computed over each syllable, which are used as actual features in the classification.

Descriptive parameters in current study, grouped by their calculation domain, are listed in the Table 2. Detailed description of these features is provided in [11]. Frequency range and duration of the syllable defines spectral and temporal boundaries of the syllable. Modulation spectrum is not purely temporal domain feature, because it is a spectrum of signal envelope. Signal envelope is given by magnitude of discrete-time analytic signal, which is formed via the Hilbert transformation [12]. Measures MSm and MSf are related respectively to modulation index and dominating frequency of the amplitude modulation.

3.2 Mel-frequency cepstral coefficients

In the second set of parameters the syllables are presented with 12 first Mel-frequency cepstral coefficients excluding the zero coefficient. The syllables are divided into overlapping frames of 256 samples with 50% overlap of adjacent frames. Each frame is transformed into mel-frequency scale by using filter bank of 32 triangular filters. The *i*th MFCC coefficient is calculated as

$$MFCC_i = \sum_{k=1}^{K} X_k \cos\left[i\left(k-\frac{1}{2}\right)\frac{\pi}{K}\right] \qquad (1)$$

where X_k is logarithm of kth mel-spectrum and K is total number of mel-spectrum bands, which in this work was 32. MFC coefficients of the syllable frames are averaged over the syllable and are used as actual features in the classifier.

3.3 Discriminative power of individual features

Classification ability of individual signal parameter feature was tested using the linear discriminant analysis (LDA) [13]. The goal is to make the classifier more efficient and improve its generalization properties by reducing dimensionality of the feature space. In the current work we test the discrimination ability of individual features for all the species in Table 1 together but also for individual species. In LDA method classification power of individual features or feature sets is evaluated by their within-class and between-class scattering matrices. Within-class scattering matrix is defined for M species by

$$S_w = \sum_{i=1}^{M} P_i S_i \tag{2}$$

where S_i is correlation matrix of features for species i and P_i is *a priori* probability of the species. *A priori* probability is defined here as $P_i = n_i/N$, where n_i is number of samples (syllables) for the species i out of total number on samples N. The between-class scatter matrix is defined as

$$S_b = \sum_{i=1}^{M} P_i(\mu_i - \mu_0)(\mu_i - \mu_0)^T \tag{3}$$

where μ_i is the mean vector of the feature vectors of species i and μ_0 is mean of all feature vectors among all species.

The criterion

$$J = \frac{\det(S_b)}{\det(S_w)} \tag{4}$$

gives large values when different classes are well separated. When testing individual species we set species of interest to one class and all others species to the another. Therefore we have a two-class problem where we can use the criterion defined in (4). The discriminative powers of individual features, defined in (4), in general and species specific cases are presented in Table 3.

Individual features were selected for classification using the scalar feature selection method [14]. In this method features are treated individually and a subset of features for recognition is selected based on the class separability measure of the individual features (Table 3).

3.4 Classification

In this article the classification is based on the k-Nearest-Neighbor (kNN) method. The Nearest Neighbour of a test vector is a vector in the training data set with the minimum distance to the test vector. In kNN method test vector is assigned to the class, which is most often represented in k-nearest neighbour. Recognition performance of the system was tested with different numbers of neighbors. Both Euclidean and Mahalanobis distance measures were used to calculate the distances between the feature vectors. With Euclidean distance measure features were first scaled to the same dynamic range. in order to obtain equal significance for different features. With Mahalanobis distance measure this is done automatically.

The *leave-k-out method* was used in splitting the segmented syllables from the database to the training and testing data sets. With this method we can use basically all available data for training and still maintaining the individual independence between training and testing data sets. Syllables left out from the training data set were selected so that those were never compared with syllables from the same recording. Syllables from same individual are likely correlated and including those in training and testing data sets would have resulted optimistic true error probability.

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

Table 3: Discriminative power of individual features. First column gives discriminative power of individual features for all species together. Latter columns gives species-specific discriminative power of features. Features are identified by their abbreviation. Lower case m and v is related to the mean and variance of the feature calculated on the frame basis.

a)

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

b)

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

Table 4: recognition results for species in current study using a) Euclidean distance measure and b) Mahalanobis distance measure and the Nearest Neighbour classifier. Columns tells the percentage of the syllables of the species on the top row being recognized as a syllables of the species on the leftmost column.

4. RESULTS

The recognition rate was measured for the species described in Table 1 using different numbers of neighbors and two distance measure. The overall species recognition rate with all signal parameter features was 49% using the nearest neighbor classifier and Euclidean distance measure. Mahalanobis distance measure improved recognition results significantly among all species and overall recognition rate improved to 71%. Confusion matrices for six species using Euclidean and Mahalanobis distance measures are presented respectively in Tables 4 a) and b). Altering the number of the neighbours in the classifier had only small effect to the overall recognition rate.

Decreasing the dimension of the feature space by removing the features features with a low classification power had

a)

recog. rate	CORRAX	CORNIX	PICPIC	GARGLA	ACRSCH	ACRRIS
CORRAX	87	3	3	2	0	0
CORNIX	1	62	7	8	7	1
PICPIC	4	19	74	9	5	2
GARGLA	7	8	5	70	11	2
ACRSCH	1	9	4	9	61	8
ACRRIS	0	1	6	2	16	86

b)

recog. rate	CORRAX	CORNIX	PICPIC	GARGLA	ACRSCH	ACRRIS
CORRAX	92	4	5	2	0	0
CORNIX	2	66	22	8	3	1
PICPIC	4	16	63	9	5	0
GARGLA	1	8	7	80	11	0
ACRSCH	0	4	1	1	57	12
ACRRIS	0	2	2	0	24	86

Table 5: recognition results for species in current study using MFCC-coefficients and a) Euclidean distance measure and b) Mahalanobis distance measure. Recognition rates are as in the Table 4

only a small effect to the recognition result. Dimension of the feature space could be decreased down to six features without significant effect to the recognition rate. However, using less than six features recognition rate was heavily decreased. The result was the similar when the general and species-specific classification powers were used for feature selection. However the the change in recognition accuracy of individual species was higher when different feature sets were used for different species.

Recognition results with MFCC representation of syllables are presented in Tables 5 a) and b) using Euclidean and Mahalanobis distance measures, respectively. Average recognition rate was higher compared to low level signal parameter representation of syllables with both distance measures. The average recognition rate with Euclidean distance measure was 73% and with Mahalanobis distance measure 74%.

5. CONCLUSIONS

In this paper we have studied methods for parametrization of *inharmonic* bird sounds, that is, sounds that are not clearly tonal or harmonic. We have used several different computational measures to characterize properties of syllables for a recognizer. We also introduced a method for comparing discriminative power of individual features. This method is used for dimension reduction of feature vectors. Classification power of individual features shows that features related to the frequency band of the sound, such as SC, SRF and frequency range, provides good classification power within inharmonic sounds. Also modulation measures give good classification power in some species.

It was found that the average values of feature trajectories computed over a syllable are much more useful in classification than the variances of those trajectories. This may reflect the fact that many of the sounds in the current study are relatively stationary over the duration of the syllable. Low classification power in mEN and mSF supports this assumption.

Recognition results suggests that, on the average, MFC coefficients provide more accurate representation of inharmonic sounds of birds than descriptive signal parameters. Small difference in recognition accuracy between two distance measures in MFCC representation is probably due to fact that MFC coefficients are already decorrelated by the discrete cosine transform. With descriptive signal parameters and Euclidean distance measure features are not decorrelated.

Recognition results suggests that the classification power of features can be used efficiently for the reduction of the dimensionality of a feature vector. Clearly features have different importance for classification, but also the results suggests that the features with low classification power does not disturb the recognition task. However classifiers complexity increases and generalization properties are weaker when features with low classification power are with the representation of syllables.

REFERENCES

[1] A. Härmä and P. Somervuo, "Classification of the harmonic structure in bird vocalication," in *IEEE Int. Conf. Acoust. Speech and Signal Processing*, Montreal,Canada, May 2004.

[2] C. K. Catchpole and P. J. B. Slater, *Bird Song: Biological Themes and Variations*, Cambridge University Press, Cambridge, UK, 1995.

[3] S. E. Anderson, A. S. Dave, and D. Margoliash, "Template-based automatic recognition of birdsong syllables from continuous recordings," *J. Acoust. Soc. Am.*, vol. 100, no. 2, pp. 1209–1219, August 1996.

[4] J. A. Kogan and D. Margoliash, "Automated recognition of bird song elements from continuous recordings using dynamic time warping and hidden Markov models: A comparative study," *J. Acoust. Soc. Am.*, vol. 103, no. 4, pp. 2185–2196, April 1998.

[5] A. L. McIlraith and H. C. Card, "Birdsong recognition using backpropagation and multivariate statistics," *IEEE Trans. Signal Processing*, vol. 45, no. 11, pp. 2740–2748, November 1997.

[6] D. A. Nelson, "The importance of invariant and distinctive features in species recognition of bird song," *Condor*, vol. 91, pp. 120–130, 1989.

[7] S. Fagerlund, "Avesound - automatic recognition of bird species by their sounds," http://www.acoustics.hut.fi/˜sfagerlu/project/avesound.html, 2005, Avesound project web-site.

[8] D. Li, I. K. Sethi, N. Dimitrova, and T. McGee, "Classification of general audio data for content-based retrieval," *Pattern Recognition Letters*, vol. 22, pp. 533–544, 2001.

[9] M. F. McKinney and J. Breebaart, "Features for audio and music classification," in *Int. Conf. on Music Information Retrieval*, 2003.

[10] S. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 28, no. 4, pp. 357–366, 1980.

[11] S. Fagerlund, "Automatic recognition of bird species by their sounds," M.S. thesis, Helsinki University of Technology, 2004.

[12] W. M. Hartmann, *Signals, Sound, and Sensation*, AIP Press, Woodbury, New York, USA, 1 edition, 1997.

[13] K. Fukunaga, *Introduction to Statistical Pattern Recognition*, Academic Press, San Diego, California, USA, 1990.

[14] S. Theodoridis and K. Koutroumbas, *Pattern Recognition*, Academic Press, San Diego, California, USA, 1 edition, 1998.