CLASSIFICATION OF THE HARMONIC STRUCTURE IN BIRD VOCALIZATION

A. Härmä[*]

Laboratory of Acoustics and Audio Signal Processing, Helsinki University of Technology Espoo, Finland

P. Somervuo[†]

Neural Networks Research Center Helsinki University of Technology Espoo, Finland

ABSTRACT

This article is related to the development of techniques for automatic recognition of bird species by their sounds. It has been demonstrated earlier that a simple model of one time-varying sinusoid is very useful in classification and recognition of typical bird sounds. However, a large class of bird sounds are not pure sinusoids but have a clear harmonic spectrum structure. In this article, we introduce a way to classify bird syllables into four classes by their harmonic structure.

1. INTRODUCTION

Automatic classification and recognition of sound sources and generic audio material differs from speech recognition in many ways. In the most generic case we cannot specify a source model which would aid in finding efficient parametric representations for sound events. In speech recognition a certain source model assumed and we may expect the signal to obey the laws of a specific spoken language with a vocabulary and a grammar. In other than speech signals, such as music, environmental sounds, or animal sounds, this is not always clear. Nevertheless, there are classes of other than speech sounds which probably have a *vocabulary* which in automatic recognition can be characterized using a set of descriptive parameters.

Bird songs is a good example of a class of natural sounds where we can expect to find a vocabulary. In bird vocalization we also have a pretty good understanding on the physics of sound production, see, e.g., [1], for references. Automatic recognition of bird species and even individuals by their sounds is a potential new tool for biological sciences. There are also extensive international initiatives on building biological multimedia databases on all living organisms (such as the *Global Biodiversity Information Facility* (GBIF) by the OECD countries). So far, little has been done to incorporate animal sounds into those, but quite obviously that will be considered in the future.

In [1] we studied automatic recognition of fourteen bird species common in all Northern Europe. The working hypothesis was that it would be possible to recognize bird species directly from *syllables* which are elementary building blocks of bird song [2]. Typically the duration of a syllable ranges from few to few hundred milliseconds. If this should be possible, the recognition of species could be performed even from brief clean periods in a noisy environmental recording. The alternative approach of recognizing song melodies is difficult in some species due to high regional variability and imitation of the song of other species, which is a common phenomenon. A majority of earlier work on automatic recognition of bird species have focused on recognition of melodies, see [3] for review.

In [1] the parametrization was based on sinusoidal modeling of syllables. Recognition results were encouraging even if the signal model was clearly oversimplified: each syllable was represented by frequency and amplitude trajectories of a single time-varying sinusoid. A time-varying model is significantly better in recognition of syllables than just a center frequency of a syllable, which has been used, for example, in recognition of song melodies in a pioneering work by McIlraith [4]. Our recent result in Fig. 1 also show that the use of a time-varying sinusoidal model instead of a center frequency of a syllable increases the accuracy of song recognition rate by 10-30 %.

In the current article, we study how harmonic structure of sounds can be efficiently estimated and parametrized to further improve the accuracy in recognition of bird species.

2. METHODS

In the current article, syllables of bird vocalization are modeled using a parametric line spectrum estimation method which is often called Analysis-By-Synthesis/Overlap-Add (ABS/OLA) when referring to an efficient frequency-domain algorithm proposed by George and Smith [5]. This technique is better than the one we used in [1] because it guarantees that the removal of a new sinusoid in a frame will

[*]Dr. Härmä was supported by the Academy of Finland.

[†]Dr. Somervuo was supported by the Academy of Finland, project no. 44886 New information processing principles (Finnish Centre of Excellence Programme 2000-2005)

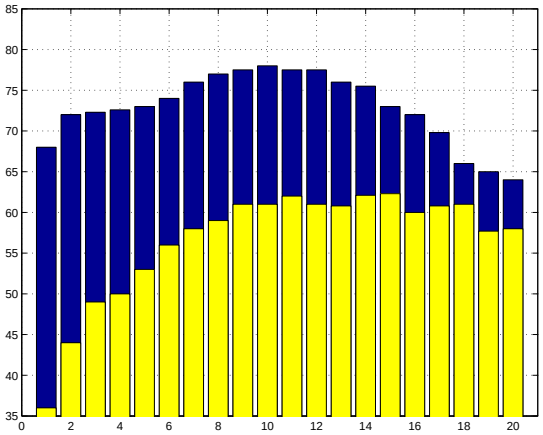

Fig. 1. The song recognition percentage as a function of the number of consecutive syllables. Upper plot: DTW-based distance function between syllables represented as a time-varying sinusoid. Lower plot: Euclidean distance between syllables represented by the average frequency over its duration. Previously unpublished results left out of [3].

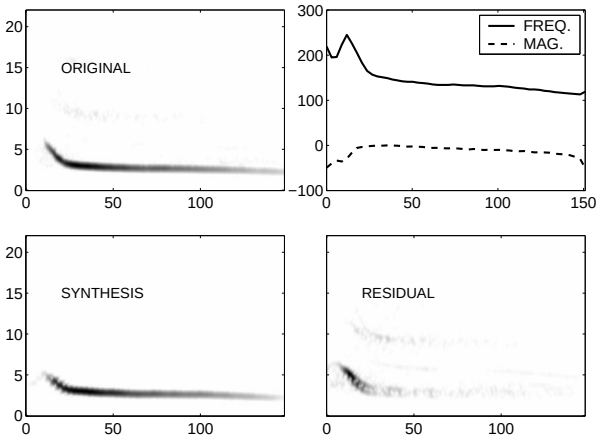

Fig. 2. The top left panel shows a spectrogram of a typical syllable from Willow Warbler. The top right panel shows frequency and amplitude trajectories of the one-sinusoid model (in FFT-bins and decibels). The two lower panels show spectrograms of a synthesised signal and the residual after substracting the sinusoid from the original signal. The y-axis represents frequency in kHz and the x-axis is time in milliseconds.

always decrease the energy of the residual signal.

The segmentation of a recording to individual syllables is done using an iterative time-domain algorithm. First, we compute a smooth energy envelope of the signal and find the global maximum M_{dB}. Next, we initialize a threshold D_{dB} to a suitable value (e.g., 20 dB). Then we apply the following algorithm sequentially until it converges such that the estimate of the level of the background noise N_{dB} becomes sufficiently stable.

Algorithm 1

1. Find maximum points and regions which are within D dB below the global maximum of the envelope.
2. Estimate N_{dB} from gaps between high energy regions.
3. Update the threshold, e.g., $D_{\mathrm{dB}} = (M_{\mathrm{dB}} - N_{\mathrm{dB}})/2$ and return to step 1.

In the current article the implementation of ABS/OLA is such that we first use it to find a single time-varying sinusoidal component over a syllable. To get a smooth sinusoidal representation we start at the energy maximum of a syllable ad proceed forward and backward in time so that the maximum frequency difference between peaks in consecutive frames is not allowed to exceed a certain limit. With the step size of 127 samples (with a 256-sample Hanning window and FFT-size of 1024) at the sampling rate of 44.1 kHz we allow a 5% deviation (in relation to the center frequency) in going from one frame to another. The frequency trajectory of a sinusoid is terminated when the amplitude of the estimated sinusoid falls below D_{dB} (see above).

In a frame n we first find the frequency of the maximum ω_n (within the allowed frequency range) and utilize the frequency domain algorithm proposed in [5] to find phase ϕ_n and magnitude m_n corresponding to an optimal sinusoidal pulse. Conceptually, we may write a function call

$$[m_n, \phi_n, \hat{s}_n] = \mathrm{absola}(\hat{x}_n, \omega_n), \qquad (1)$$

where $\hat{x}_n$ is a windowed signal segment corresponding to the nth frame of the original signal x, and $\hat{s}_n$ is a sinusoidal signal which can be used to synthesize a sinusoidal representation of the signal, denoted s_I, in the overlap-add sense. We can also compute a modeling error signal by $e_I = x - s_I$.

As was demonstrated in [1], a single sinusoidal model is often enough. However, syllables with a clear harmonic structure are common. In this article we divide sounds into four classes by their harmonic structure.

Class I representation is the one-sinusoid model. For example, a syllable from the Willow Warbler (*Phylloscopus trochilus*) illustrated in Fig. 2 is a good example of a pure sinusoidal syllable. In Class II representation the single sinusoid is a fundamental of a harmonic series. For example, the top left spectrogram of Fig. 3 representing a syllable from Blackbird (*Turdus merula*) has the first and the second harmonic of the estimated sinusoidal component clearly visible.

Fig. 4 represents a typical Class III syllable from Icterine Warbler (*Hippolais icterina*). In this class the fundamental component is weak and the sinusoidal component with the highest amplitude is the first harmonic of the series. In Fig. 5, from March Warbler (*Acrocephalus palustris*), the

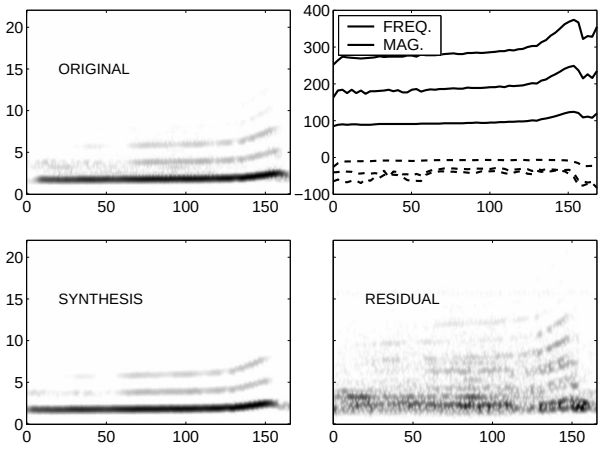

Fig. 3. Representations of a syllable from Blackbird with two harmonic components of the fundamental sinusoidal. Panels are similar to Fig. 2.

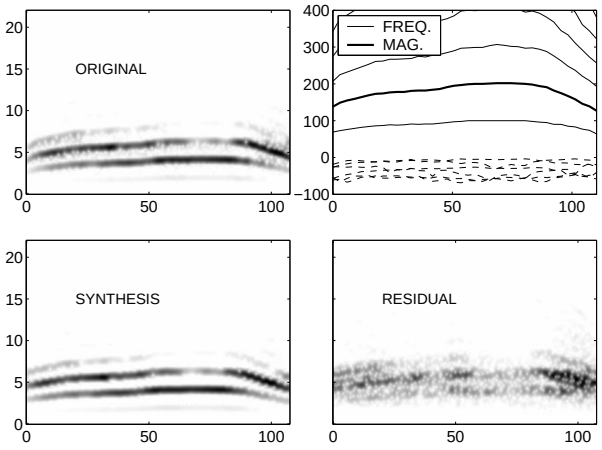

Fig. 4. Representations of a syllable from Icterine Warbler the strongest sinusoidal is actually the first harmonic. Panels are similar to Fig. 2.

sinusoid with the highest amplitude is the second harmonic of the series which is characteristic for Class IV syllables. In our database of bird recordings cases of harmonic sounds where the dominant sinusoidal component would be higher than the second harmonic are rare.

The estimation of the levels of harmonic components can be done using the following procedure. First, we fit one time-varying sinusoidal signal over a syllable to find time-varying parameters ω_n, ϕ_n, m_n, for $n = 0, ..., N-1$, synthesized signal s_I and residual e_I. After that we may use frequency estimates corresponding to the dominant sinusoid ω_n to estimate the level of its kth harmonic. Using the notation from (1), estimation of parameters for a harmonically related component of ω_k is given by

$$[m_{nk}, \phi_{nk}, \hat{s}_{nk}] = \text{absola}(\hat{e}_n, k\omega_n). \quad (2)$$

For example, frequency curves in Fig. 3 corresponding to Class II were computed using values $k = k_{\text{II}} = 1, 2, 3$. In Class III, $k = \bar{k}_{\text{III}} = \frac{1}{2}, 1, \frac{3}{2}, ...$, and finally Class IV is represented by a harmonic series formed by $k = \bar{k}_{\text{IV}} = \frac{1}{3}, \frac{2}{3}, 1, \frac{4}{3}, ...$ Synthetic signals s_C whose spectrograms has been illustrated in bottom left panels in Figs. 2-5 were created using overlap-add synthesis of a sum of corresponding synthesized components $\hat{s}_{nC}$, where C denotes a class. Moreover, the residual signal in bottom right panels was given by $e_C = x - s_C$.

Next, we note that $\bar{k}_{\text{III}}$ and $\bar{k}_{\text{IV}}$ intersect with k_{II}. Therefore, in the following we define Class III and IV harmonic series k_{III} and k_{IV} so that multipliers 2, 3, and 4 have been removed from sets. In addition, we define a new group A of harmonics which is obtained as an union of all other classes. That is, $k_{\text{A}} = k_{\text{II}} \cup k_{\text{III}} \cup k_{\text{IV}}$.

In this article, we compute a modeling gain corresponding to a class C in the following way:

Sample	H_{I}	H_{II}	H_{III}	H_{IV}	R
Fig. 2	**0.933**	0.140	0.324	0.43	0.53
Fig. 3	0.226	**0.472**	0.042	0.27	7.02
Fig. 4	0.000	0.005	**0.736**	0.0240	22.66
Fig. 5	0.002	0.001	0.006	**0.951**	18.10

Table 1. Harmonic parameters corresponding to syllables in Figs. 2-5

$$G_{\text{C}} = 20 \log_{10} \left(\frac{E[x^2]}{E[e_C^2]} \right), \quad (3)$$

where $E[.]$ denotes expectation. For ABS/OLA it holds that for any signal $G_{\text{I}} < G_{\text{A}}$, and all other class estimates will fall in $[G_{\text{I}}, G_{\text{A}}]$. Therefore, we define a range measure $R = G_{\text{A}} - G_{\text{I}}$ which gives the difference in modeling gain between the cases where only one sinusoid has been modeled (Class I) and where the dominant sinusoid and all its harmonics and sub-harmonics have been modeled. Finally, we define a test which gives a likelihood that a certain syllable is from Class C={II, III, IV}:

$$H_{\text{C}} = (G_{\text{C}} - G_{\text{I}})/R, \text{ with } 0 < H_{\text{C}} < 1. \quad (4)$$

In order to determine if a signal belongs to Class I, we introduce a heuristic measure given by

$$H_{\text{I}} = ((1+\exp(0.6R-3))(1+exp(-0.2G_{\text{I}}-2)))^{-1} \quad (5)$$

which gives a value close to one when the modeling gain of G_{I} is large but R is small. In case of a high noise level or signals which do not match any of the proposed models, G_{A} gives a small value. Measured H_{C} values corresponding to

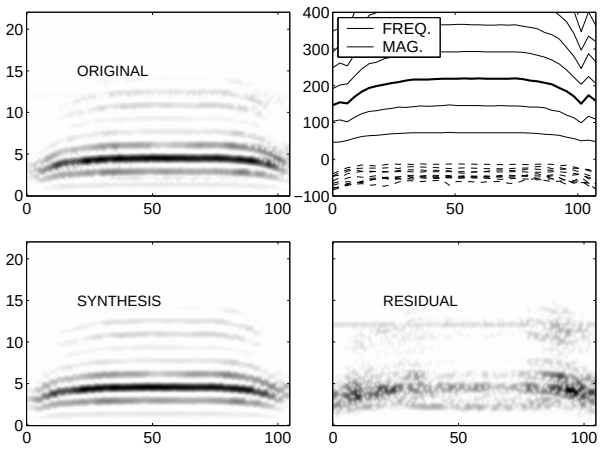

Fig. 5. Representations of a syllable from Marsh Warbler where the strongest sinusoidal component is the second harmonic of a fundamental. Panels are similar to Fig. 2.

the signals in Figs. 2-5 are shown in Table 1. All signals are classified as expected (the highest value is bolded).

The method can be made very efficient because the computation of all e_C can be implemented directly in the FFT-domain due to the properties of ABS/OLA and the Parseval's theorem.

3. RESULTS

The current XML-based bird song recording database collected at HUT/Acoustics has nearly 2000 recordings from almost 150 bird species. The total number of syllables in the database is more than 30000. However, for a majority of species the number of bird individuals is not sufficiently large for reliable species recognition experiments. Classification results over the passerine birds in the database shows that almost 60% of syllables are classified as pure sinusoidal sounds (Class I), and 14 % are in Class IV which is the second largest class. But, only 7% of the syllables can be considered noise because of a low modeling gain G_{A} (1–6 dB).

Statistics of four species are shown in Table 2. The results in Table 2 for the first four species seem reasonable. For example, Willow Warbler's syllables are typically clean sinusoidal chirps while others have more variability in timbre. The last species, Hooded Crow, is clearly an outlier. Visual inspection of a spectrogram of a typical crow's caw shows very little harmonicity. However, the current classification finds a high number of Class IV syllables.

4. CONCLUSIONS

In this article we introduced a computationally efficient technique to classify sinusoidal representations of brief segments of bird song, syllables, to four classes by their harmonic structure. It was demonstrated that the method is probably useful in developing system for automatic recognition of bird species. It also seems that the proposed signal models match with the spectral structure in 93 % of syllables in our database. Actual recognition results have not been reported in the current article. This is work in progress and a snapshot of recognition results is held up-to-date at our web-site [6]. In many bird species the use of the proposed harmonic class information improves recognition results by 5-20 %. However, in some species improvements are small and suggest that the recognition cannot be done on the basis of isolated syllables only, but it may be necessary to take into account also song-level structural information.

Species	I	II	III	IV	Noise
Willow Warbler	83	11	4	0	1
Comm. Chaffinch	56	5	33	0	6
Blackbird	46	13	38	3	1
Marsh Warbler	41	5	20	26	8
Hooded Crow	0	9	9	81	0

Table 2. Percentages of syllables belonging to different classes for a selection of species.

5. REFERENCES

[1] A. Härmä, "Automatic recognition of bird species based on sinusoidal modeling of syllables," in *IEEE Int. Conf. Acoustics, Speech, Signal Processing (ICASSP 2003)*, Hong Kong, April 2003.

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

[3] P. Somervuo and A. Härmä, "Bird song recognition based on syllable pair histograms," in *Submitted to IEEE ICASSP 2004 (this conference)*, Canada, May 2004.

[4] 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.

[5] B. George and M. J. T. Smith, "Speech analysis/synthesis and modification using an analysis-by-synthesis/overlap-add sinusoidal model," *IEEE Trans. Speech Audio Processing*, vol. 5, no. 5, pp. 389–406, September 1997.

[6] A. Härmä, "Avesound project web-site," http://www.acoustics.hut.fi/˜aqi/projects/avesound.html, 2003.