AUTOMATIC IDENTIFICATION OF BIRD SPECIES BASED ON SINUSOIDAL MODELING OF SYLLABLES

Aki Härmä

Helsinki University of Technology,
Laboratory of Acoustics and Audio Signal Processing
P.O. Box 3000, FIN-02015, Espoo, FINLAND
email: Aki.Harma@hut.fi

ABSTRACT

Syllables are elementary building blocks of bird song. In sounds of many songbirds a large class of syllables can be approximated as amplitude and frequency varying brief sinusoidal pulses. In this article we test how well bird species can be recognized by comparing simple sinusoidal representations of isolated syllables. Results are encouraging and show that with limited sets of bird species a recognizer based on this signal model may already be sufficient.

1. INTRODUCTION

Birds and their sounds are in many ways important for our culture. They can be heard even in big cities and most people can recognize at least a few most common species by their sounds. Bird song has also been an important source of inspiration for many composers, musicians, and writers. In this article we study automatic recognition of bird song. Technology for sound-based identification of bird species and even individuals would be a significant addition to the research methodology in taxonomy and monitoring of migration and population in biology. At a higher level it would also facilitate systematic research on vocal communications between birds and characterization of their sounds. There is also commercial potential for such systems because the number of active bird watchers is really large in many countries.

Sounds of birds are mainly produced by syrinx, which an organ unique to birds [1]. It is located in the intersection between the main bronchi of the lungs and the trachea, or in the trachea. There is a considerable variability in the anatomy of syrinx even in different families within the same order of birds. The function of syrinx resembles that of human vocal cords in many ways. But, it typically has much more complex structure and can produce a significantly larger variety of different sounds than glottis in mammals. While in human speech sounds are mainly produced by muscular control of the vocal tract, mouth, lips, tongue, and teeth, the syrinx is the main source of variability of sounds in birds. Only few bird species, mainly parrots, can use their tongue in a way which resembles speech production in humans [2].

Syrinx in birds can feature multiple simultaneous oscillation modes. In many cases, the sound production system is in a highly nonlinear or chaotic operation mode [3] which results in rapid changes in the operation of the organ. Consequently, the temporal variability in spectrum of bird song is typically orders of magnitude faster than in human sound production. The neural control of sound production is also significantly faster in birds than in human speech production or in any musical instrument. There are also recent experimental results showing that the temporal resolution of hearing is much better in birds than in humans [4]. Therefore, a high temporal resolution in the range of few milliseconds is needed in the analysis of bird song. The spectrum energy in song birds is typically concentrated on a very narrow area in the range of 1 to 6 kHz, and the sound is often composed of a single or a small number of sinusoidal components. Therefore, it is natural to use *sinusoidal* modeling [5] as a basic tool in representing bird sounds.

Bird song is typically divided into four hierarchical levels: notes, syllables, phrases, and song [6]. In many species there is high individual and regional variability in phrases and song patterns. Syllables can be seen as more elementary building blocks of bird vocalization [7] and may therefore be more suitable for automatic identification of bird species than song patterns. The duration of a syllable is in the range of a few to a few hundred milliseconds. Recognition of bird species directly from syllables would be technically more feasible approach than recognition by song in cases when there are, as usual, many birds singing simultaneously. In a continuous environmental recognition with multiple singing birds it is very difficult to segment a song of a bird. But, we may be able to isolate a number of individual syllables from a recording relatively easily. In addition, a syllable recognizer would be more invariant to regional variation in song patterns which is a common phenomenon with many species.

Relatively little has been done previously in the field. In a few studies the feasibility of automatic recognition of bird species [7, 8, 9] or even individual males of a given species [10, 11] using sound has been demonstrated. In this article we apply sinusoidal modeling to syllables of continuous bird song and use obtained parameters for recognition of a number of song bird species. This technique is proposed here as a baseline technique for bird sound identification and therefore the goal is to evaluate how well this fairly simple and low-complexity approach without any intelligent or context-aware processing works for a number of species. This helps us to anticipate which type of processing we may have to add in order to improve results in the future.

2. METHOD

Different digital representations of bird sounds have been presented earlier, e.g., in [9, 12, 13]. In [7, 8] template-matching of spectrograms was used to study learning of song patterns and syllables in two species. Their recordings were all made in a laboratory with cage birds. Matching of spectrograms is very sensitive to

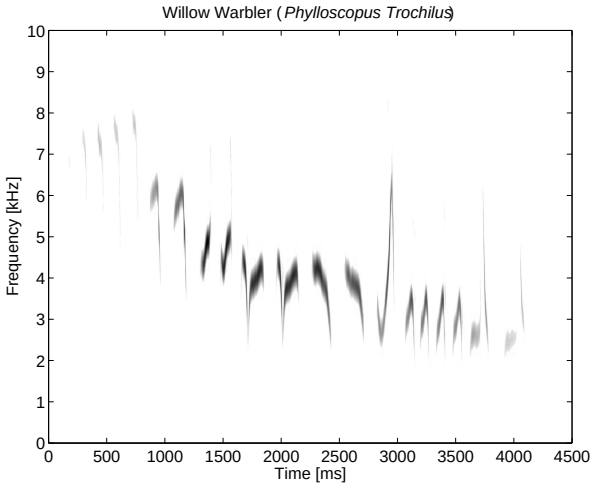

Fig. 1. Spectrogram of a typical song of Willow Warbler (*Phylloscopus trochilus*, PHYLUS). This song is composed of twenty syllables.
ă

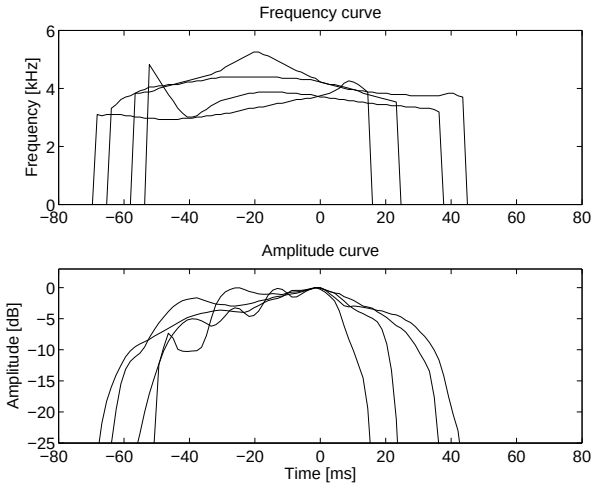

Fig. 2. Frequency and amplitude trajectories of a set of four syllables from the song of a Willow Warbler (*Phylloscopus trochilus*)

environmental noise of a field recording and is also highly demanding in data storage and computation. They also used hidden Markov models, adopted from speech recognition, to model transition propabilities between syllables in bird sing. In [9], the method was developed to identify song patterns based on parameters which represent average frequencies and durations of syllables and pauses in a bird song.

The basic methodology in the current article is to decompose a bird song recording to a set of brief frequency and amplitude modulated sinusoidal pulses. Each pulse represents one individual syllable and syllables are not overlapping in time or frequency. This is a highly simplified model for bird song, but, to our knowledge, it has not been tested previously with a large number of different species.

Short-time Fourier transform (STFT) was used to compute a spectrogram for a song segment. The spectrogram of a typical song pattern of *Willow Warbler* is illustrated in Fig. 1. In this example, the size of a *Kaiser* ($\alpha = 8$) window was 256, FFT size with zero-padding was 1024, and a spectrum vector was computed with 75 % overlap (64 sample steps) over a signal sampled at 44.1 kHz. The decomposition of a song to a set of N syllables runs as follows:

Algorithm 1

1. Compute a spectrogram of a song segment using FFT. We denote a spectrogram a matrix $S(f, t)$, where f represents frequency and t is time.
2. Repeat steps 3-7 for $n = 0, 1, \cdots, N-1$.
3. Find f_n and t_n such that $|S(f_n, t_n)|$ is the maximum value in the spectrogram. This position represents the maximum amplitude position of nth sinusoidal syllable.
4. Store frequency parameter $\omega_n(0) = f_n$ and amplitude $a_n(0) = 20 \log_{10} |S(f_n, t_n)|$ [dB].
5. Starting from $|S(f_n, t_n)|$, trace the maximum peak of $S(f, t)$ for $t > t_0$ and for $t < t_0$ until $a_n(t - t_0) < a_n(0) - T$dB, were the stopping criteria T is typically 30 dB. It will determine how the sinusoidal syllable starts and ends at times t_s and t_e, respectively around the amplitude maximum t_0.
6. Store obtained frequency and amplitude trajectories corresponding to the nth syllable in functions $\omega_n(\tau)$ and $a_n(\tau)$, where $\tau = t_0 - t_s, \cdots t_0 + t_e$
7. Set $S(f, [t_s, t_s + 1, \cdots, t_e]) = 0$ to delete the area of nth syllable.

The algorithm extracts N frequency and amplitude modulated sinusoidal pulses from a signal automatically starting with the one with highest peak amplitude. Some post-processing is typically necessary in order to remove clearly erroneous syllables, however, the method appears to be relatively insensitive to environmental noise and coloration both commonly found in typical field recordings. A set of four frequency and amplitude trajectories extracted from the song of Willow Warbler (see Fig. 1) is shown in Fig. 2.

The position of the maximum of the amplitude trajectory is the same in each syllable. This makes comparison between different syllables easy. In particular, in this article the distance criterion between two syllables is a weighted sum of mean differences between frequency and amplitude trajectories $\omega_n(\tau)$ and $a_n(\tau)$, respectively.

3. EXPERIMENTS

In this article we limit the test to a group of Passerine birds (*Passeriformes*) listed in Table 1. Many of these are common songbirds in all Northern Europe and are considered good singers. An expert listener can basically recognize all listed species by their song, although some pairs of birds in the selection may be difficult. Identification by isolated syllables only would be a very difficult task even for an expert. Median frequencies with upper an lower quartile values of sinusoidal syllables computed from the database are shown in Fig. 3. Most birds have a typical center frequency around 3-5 kHz and majority of their syllables are 40 to 400 ms whistles or chirps. Both territorial songs and isolated calls and warnings were used.

Lat. Abbr.	Common name	Recs.	Syllables
FICHYP	Pied Flycatcher	3	256
FRICOE	Common Chaffinch	6	365
PARATE	Coal Tit	4	402
PARMAJ	Great Tit	7	472
PHOPHO	Common Redstart	4	566
PHYBOR	Arctic Warbler	4	648
PHYCOL	Comm. Chiffchaff	9	774
PHYDES	Greenish Warbler	4	480
PHYLUS	Willow Warbler	10	1173
PHYSIB	Wood Warbler	6	751
SYLATR	Blackcap	5	783
SYLBOR	Garden Warbler	5	900
TURMER	Blackbird	5	673
TURVIS	Mistle Thrush	6	317

Table 1. Birds in the current study. The first column gives an abbreviation derived from the Latin name (a widely used convention), common English name, number of recordings from different birds, and the total number of syllables from each species. First three letters of the abbreviation indicate family of species.

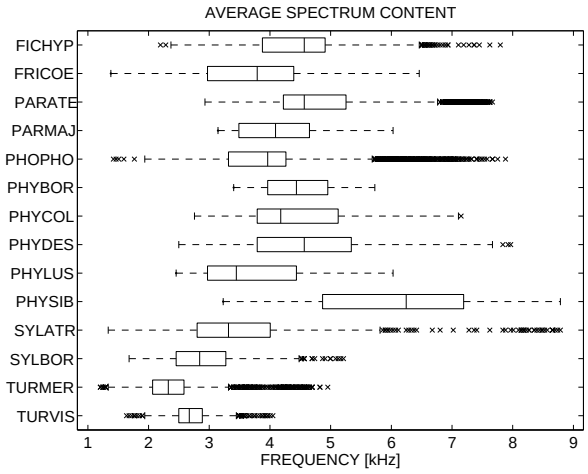

Fig. 3. Average frequency of syllables of 15 species listed in Table 1. A box indicates lower quartile, median, and upper quartile values.

Test material consists of a number of recordings from different birds at different sites mainly recorded in Finland. Most of recordings are raw field recordings with additional sounds of other birds and environment. Approximately 20 % of recordings were taken from commercially available CD-collections. The number of recordings (or birds) and the total number of syllables are in Table 1.

In a recognition experiment we first collected a number of sinusoidal representations of syllables from each recording in the database, see Table 1. Then differences between a *test syllable* from a recording and all syllables from all the other recordings were computed. Test syllable was then assigned a label representing a species which has a syllable with a smallest difference to the test syllable. This was repeated separately for all syllables in each recording. Finally, we computed a histogram of labels assigned to syllables of each species and computed recognition probabilities corresponding to all the species.

Results for two sets of species are shown in Table 2, where each column representing a species gives percentages that a single syllable is identified as a syllable of a species at different rows of the table. In Table 2A, syllables from the five birds in the genus of *Phylloscopus* get highest percentage for the right species (row). However, the difference is small in some cases. It should be noted that in some cases, e.g., in Greenish Warbler (PHYDES) and Willow Warbler (PHYLUS) the actual song pattern is very different and easy to identify by the ear. However, results in the table indicate that there is a significant risk that a syllable of PHYDES is identified as syllable of PHYLUS. However, the misclassification risk is much smaller for the syllables of PHYLUS, which is caused by the fact that the *vocabulary* of PHYDES is much smaller that that of PHYLUS.

In the right table of Table 2 a number of bird species from different families of songbirds was compared. In three cases the highest probability is obtained for misclassification of species. There is a clear trend that species within the same genus get most easily confused. For example, Thrushes (*Turdus*) TURMER and TURVIS, or Tits (*Parus*) like PARMAJ and PARATE get easily misclassified. However, this may partly reflect the fact that the average frequency content of Thrushes and Tits are different. Low percentage for the correct identification of Pied Flycatcher (FICHYP) may be caused by the fact that the number of recordings was low compared to many other species, see Table 1.

Finally, we made a full recognition experiment with all the species in the database. The results are shown in Table 3. Three species are clearly misclassified. For others, the percentage of correct identification of a syllable is highest. However, in many cases the percentage is only around 30%. For example, the correct identification of PHYBOR versus PHYDES could require more than 100 syllables which, for these species, corresponds to less than 20 seconds of continuous singing. It also turned out that in many cases there are significant differencies in the recognition accuracy for individual song segments or recordings within a species.

4. DISCUSSION

In this article we studied automatic sound-based identification of bird species. We started with a hypothesis that identification of species could be done by comparing sinusoidal representations isolated syllables of bird song. Possibility to identify species on the basis of isolated syllables instead of significantly longer song patterns would be beneficial for many reasons. First, regional variability of song patterns within the same species could be easily

A)

Ident. percentage %	PHYBOR	PHYCOL	PHYDES	PHYLUS	PHYSIB
PHYBOR	55	10	26	6	1
PHYCOL	15	51	14	16	2
PHYDES	21	11	27	10	11
PHYLUS	6	25	21	65	6
PHYSIB	1	3	12	4	80
Nobird	1	0	0	0	0

B)

Ident. percentage %	FICHYP	FRICOE	PARATE	PARMAJ	TURMER	TURVIS
FICHYP	15	14	7	11	4	2
FRICOE	9	43	7	5	4	4
PARATE	17	7	35	15	1	0
PARMAJ	46	15	48	64	6	2
TURMER	8	13	2	5	45	62
TURVIS	5	8	0	0	39	30
NoBird	0	0	0	0	0	0

Table 2. Identification results A) for five species from the family of Phylloscopus birds, and B) a set of other species. Columns give the percentage of syllables in a bird indicated in the top row being identified as a syllable of a species indicated in the leftmost column. The last row 'NoBird' represent the percentage of syllables where the difference to any other syllable is very large and therefore no recognition label was assigned.

Ident. percentage %	FICHYP	FRICOE	PARATE	PARMAJ	PHOPHO	PHYBOR	PHYCOL	PHYDES	PHYLUS	PHYSIB	SYLATR	SYLBOR	TURMER	TURVIS
FICHYP	5	8	1	1	2	0	2	1	1	0	3	1	2	0
FRICOE	1	58	0	2	2	0	0	0	1	0	1	5	2	2
PARATE	8	4	20	4	1	0	7	2	2	0	1	1	1	0
PARMAJ	29	2	27	55	15	0	6	1	12	4	14	3	4	0
PHOPHO	6	3	0	9	9	1	2	3	8	0	2	9	4	2
PHYBOR	0	2	0	0	5	56	5	18	3	1	0	1	0	0
PHYCOL	9	1	21	2	7	15	51	16	8	0	1	5	1	1
PHYDES	4	2	12	0	5	23	9	29	6	7	1	3	0	0
PHYLUS	9	6	16	15	29	4	14	16	47	1	8	16	3	9
PHYSIB	1	1	1	2	1	0	0	7	0	83	2	1	0	0
SYLATR	14	2	1	5	2	0	2	0	2	2	41	13	7	8
SYLBOR	11	11	1	0	16	1	2	7	8	2	15	25	18	16
TURMER	2	0	0	5	4	0	0	0	0	0	7	10	32	31
TURVIS	1	0	0	0	2	0	0	0	2	0	4	7	26	31

Table 3. Columns give the percentage of syllables in a bird indicated in the top row being identified as a syllable of a species indicated in the leftmost column.

neglected in identification. In addition, the case of a typical field recording with multiple birds could be handled without the need to separate songs of individual birds. Sinusoidal representation is a natural approach for recognition of songbirds since their calls and syllables are often clearly sinusoidal.

The presented results are very encouraging. They show that a model of one tone syllable with no song-level contextual information is already sufficient for the identification of many species in this selection of 14 species with relatively similar vocalizations. The original hypothesis that species could be identified by syllables only seems plausible. Test also verifies that the proposed signal model captures some essential properties of many bird sounds.

However, the risk of misclassification is high for some species. Many different effects are involved. For example, PHOPHO, SYLBOR and TURMER in Table 1, has a rich vocabulary which similar syllables to many other species. The signal model is probably too simplified for general recognition of syllables. This can be easily heard by synthesizing a song from estimated sinusoidal components. The song has the same melody line but the timbre of a synthesized song may be very different. Often the actual sound has a clear harmonic structure at least up to the second and third harmonic of the fundamental frequency. It is also very common, e.g., in FICHYP, that sounds contain frequency and amplitude modulated components which have a period smaller than the 256-point FFT window used in the current article. Finally, almost all species feature non-tonal sounds like clicks and rattles which cannot be modeled with a simple sinusoidal model.

There are basically two ways to improve results. First, the signal model needs to be refined by incorporating parametric representations of harmonic structure, modulation, and non-tonal sounds in syllables. Secondly, some song-level *statistical* and *structural* descriptions may need to be added to the current framework. Statistical description could contain, for example, a probabilities of different types of syllables in song of a species. Hidden Markov models could also be used since bird song often is composed of sequences of repeating phrases. However, a vast database of recordings would be needed for training HMM or neural network models for a recognizer. One useful structural description is the rate at which syllables follow each other in a typical song.

5. ACKNOWLEDGMENTS

This work has been supported by the Academy of Finland and the Graduate School GETA. The author is grateful to Mikko Ojanen and Juha Tanttu for insightful discussions and help in writing this article.

6. REFERENCES

[1] A. S. King and J. McLelland, eds., *Form and Function in Birds*, vol. 4. Academic Press, 1989.

[2] D. K. Patterson and I. M. Pepperberg, "A comparative study of human and parrot phonation: Acoustic and articulatory correlates of vowels," *J. Acoust. Soc. Am.*, vol. 96, pp. 634–648, August 1994.

[3] M. S. Fee, B. Shraiman, B. Pesaran, and P. P. Mitra, "The role of nonlinear dynamics of the syrinx in the vocalizations of a songbird," *J. Acoust. Soc. Am.*, vol. 95, pp. 67–71, September 1998.

[4] R. J. Dooling, M. R. Leek, O. Gleich, and M. L. Dent, "Auditory temporal resolution in birds: Discrimination of harmonic complexes," *J. Acoust. Soc. Am.*, vol. 112, pp. 748–759, August 2002.

[5] R. J. McAulay and T. F. Quatieri, "Speech analysis/synthesis based on a sinusoidal speech model," *IEEE Trans. Acoust., Speech, and Signal Proc.*, vol. 34, pp. 744–754, August 1986.

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

[7] 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, pp. 1209–1219, August 1996.

[8] 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, pp. 2185–2196, April 1998.

[9] A. L. McIlraith and H. C. Card, "Birdsong recognition using backpropagation and multivariate statistics," *IEEE Trans. Signal Processing*, vol. 45, pp. 2740–2748, November 1997.

[10] P. Galeotti and G. Pavan, "Individual recognition of male Tawny owls (Strix aluco) using spectrograms of their territorial calls," *Ethology, Ecology & Evolution*, vol. 3, no. 2, pp. 113–126, 1991.

[11] K. Ito, K. Mori, and S. Iwasaki, "Application of dynamic programming matching to classification of budgerigar contact calls," *J. Acoust. Soc. Am.*, vol. 100, pp. 3947–3956, December 1996.

[12] C. Rogers, "High resolution analysis of bird sounds," in *IEEE Int. Conf. Acoust. Speech and Signal Processing*, pp. 3011–3014, 1995.

[13] A. Härmä and M. Juntunen, "A method for parametrization of time-varying sounds," *IEEE Signal Processing Letters*, vol. 9, pp. 151–153, May 2002.