INTERSPEECH 2016
September 8–12, 2016, San Francisco, USA

Bird Song Synthesis Based on Hidden Markov Models

Jordi Bonada[1], *Robert Lachlan*[2], *Merlijn Blaauw*[1]

[1]Music Technology Group, Universitat Pompeu Fabra, Spain
[2]School of Biological and Chemical Sciences, Queen Mary University of London, UK
jordi.bonada@upf.edu, r.f.lachlan@qmul.ac.uk, merlijn.blaauw@upf.edu

Abstract

This paper focuses on the synthesis of bird songs using Hidden Markov Models (HMM). This technique has been widely used for speech modeling and synthesis. However, features and contextual factors typically used for human speech are not appropriate for modeling bird songs. Moreover, while for speech we can easily control the content of the recordings, this is not the case for bird songs, where we have to rely on the spontaneous singing of the animal. In this work we briefly overview the characteristics of bird songs, compare them to speech, and propose strategies for adapting the widely-used HTS (HMM-based Speech Synthesis System) framework to model and synthesize bird songs. In particular, we focus on Chaffinch species and a database of recordings of several song bouts of one male bird. At the end we discuss the synthesis results obtained.

Index Terms: context-dependent HMM, HMM based synthesis, parametric synthesis, bird song synthesis

1. Introduction

While current efforts in realistic sound synthesis focus on imitating, by means of computational models, sounds produced by objects, musical instruments or the human voice, the synthesis of realistic non-human animal acoustic vocalizations lags significantly behind, unable to meet the demands in as varied areas as virtual reality, animation, robotics, animal assisted therapy, applied biology or psychology.

We might contend that what is really necessary for a general-purpose synthesizer is a good probabilistic model of the relevant time-varying sound characteristics. For the case of human speech, given a probabilistic model with the appropriate contextual description, it has been shown that a linear source-filter decomposition together with specific signal models (e.g. sinusoids plus noise) for the excitation component are sufficient for producing intelligible speech. Starting from this basic framework, several refinements have been proposed over the last decade that greatly improve the quality and naturalness of the synthetic sound, as well as address the expression of emotion in speech.

This paper focuses on bird songs and we propose to model them with Hidden Markov Models (HMM), a well established method in speech synthesis [1]. What has made this method work really well for speech is not just the statistical approach itself, but to a great extent the acoustic parameterization and the specific signal models resulting from decades of focused research on speech. For our purpose, it is not sufficient to adapt standard methods used in speech synthesis. On one hand, bird songs exhibit strong and rapid frequency modulations that require non-stationary analysis methods to accurately compute acoustic features. On the other hand, contextual description used for speech is not applicable to bird vocalizations, since the syntactic organization is very different.

In fact, human language uses much more complex structures than animal communication systems. From a limited set of speech sounds we can create an infinite number of meaningful sentences using grammar. Many animal vocalizations are organized in simpler structures best described as a phonological syntax [2]. Vocalizations consist of a set of vocal units that are combined to create strings which are in turn organized in different patterns. Despite their enormous variation among animal vocalizations, up to date there is no evidence of vocal syntactic structures extending beyond that of a probabilistic finite-state grammar [3]. In the case of birds, for instance, first order Hidden Markov Models (HMM) can already explain complex sequencing rules of birdsong [4].

Visual and acoustic behavioral observations allow for better understanding the behavior of individuals within a population or species. The description and knowledge of a species' vocal repertoire allow searching for meaningful information in different sounds. Ethological observations are used to understand the context-specificity of vocal signals exchanged between conspecifics and playback experiments are used to clarify the function and referential use of the emitted sounds (mother calls to young, alert messages, aggressive alarms). Scientists working on animal communication mostly employ simple synthetic calls and/or rudimentary audio transformations for playback experiments (i.e. reproducing pre-recorded animal vocalizations through a loudspeaker). Recent research shows an increasing interest in re-synthesis of vocalizations from a reduced set of controls of a physical model, especially for bird songs [5, 6]. Resynthesized signals are used in experiments where HVC neuronal activity is evaluated to assess the model performance.

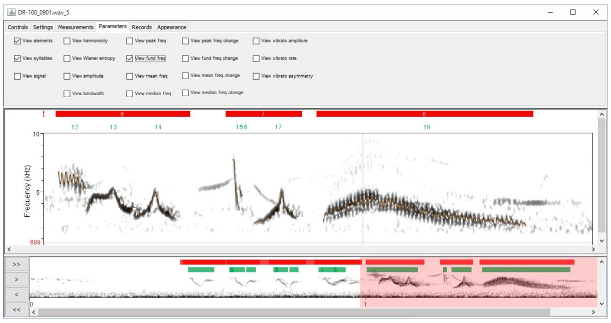
This papers proposes a method for bird song synthesis based on Hidden Markov Models. In section 2, we describe our system, based on three main steps: database creation, feature extraction, and context and HMM models definition. In section 3, we present and discuss the synthesis results and in section 4 we summarize our main contributions and propose future refinements.

2. Methodology

2.1. Sound database

We use the Chaffinch (Fringilla coelebs) dataset provided by coauthor R. Lachlan. It consists of recordings of over 700 individuals, carefully annotated specifying the specimen, recording location, date and equipment used. Most recordings have been performed with directive microphones, which greatly help to reduce the presence of undesired acoustic sources. The dataset has been annotated with Luscinia [7], an open source software

http://dx.doi.org/10.21437/Interspeech.2016-1110


DR-100_0901.wav_5

Controls | Settings | Measurements | Parameters | Records | Appearance

☑ View elements
☑ View syllables
☐ View signal
☐ View harmonicity
☐ View Wiener entropy
☐ View amplitude
☐ View bandwidth
☐ View peak freq
☑ View fund freq
☐ View mean freq
☐ View median freq
☐ View peak freq change
☐ View fund freq change
☐ View mean freq change
☐ View median freq change
☐ View vibrato ampliture
☐ View vibrato rate
☐ View vibrato asymmetry

Figure 1: *Screenshot of the spectrogram and segmentation in Luscinia software. The estimated f0 contour is plotted in brown color on top of the spectrogram. Syllables and elements are respectively drawn in red and green colors.*

for bioacoustics archiving, measurement and analysis. Many of the bird songs in the dataset have been segmented hierarchically by experts into song, syllable and element segments. Luscinia offers a spectrogram view of the bird songs, where users can control several parameters to enhance the visualization of the vocalization (e.g. dynamic range). In addition, one can indicate the regions over the spectrogram of a bird song where the vocalization signal is present, what is really useful when multiple sources are present in the recording. This procedure is used for performing a robust semi-supervised analysis, obtaining fundamental frequency and energy estimations among other features. Analysis data supervised by experts is provided for many of the bird songs in the dataset. Figure 1 shows a screenshot of the Luscinia interface showing the spectrogram and the segmentation of a chaffinch recording.

For our experiments we have considered several recordings of the same individual realized on the same day. Those recordings are segmented into 27 songs that correspond to 5 different song types, as specified in Table 1. Chaffinches typically have a repertoire of 1 to 7 distinct songs.

song	#
s1	6
s2	5
s3	4
s4	10
s5	2

Table 1: *Number of recordings for each different song type.*

2.2. Feature extraction

In general, there are several constraints we have to consider for obtaining a good acoustic representation: (1) It must be possible to resynthesize a vocalization from its acoustic representation with high quality; (2) A low-dimensional and locally stationary representation is preferable, since it has the advantage of facilitating the statistical modeling as well as requiring fewer amounts of data for training. Significant amplitude and frequency modulations should be extracted and parameterized; (3) It is usually preferable to have semantically meaningful features, easy to interpret; (4) The representation should cover some excitation characteristics of the target species particularly difficult to model, such as biphonation or strong modulations in certain bird species [8].

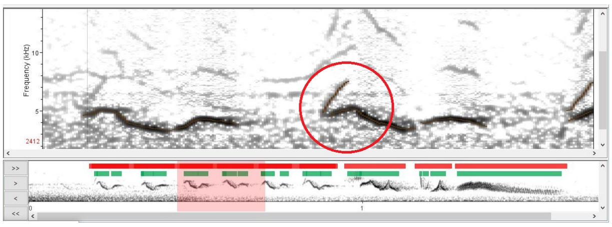


Figure 2: *Screenshot of the spectrogram and segmentation in Luscinia software. Two different frequency trajectories (in brown) overlap in the marked area*

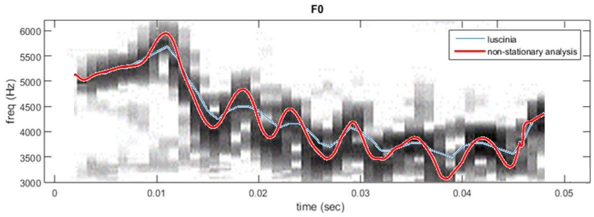


Figure 3: *Non-stationary analysis of a Chaffinch vibrato compared to Luscinia features.*

One expected issue with animal vocalization recordings is that they are often affected by reverberation (in our case mostly caused by reflections from trees). In addition, we find rapid and strong frequency and amplitude modulations for many species, causing standard analysis methods to over-smooth the extracted features. Non-stationary analysis and de-reverberation techniques are then essential for improving the accuracy of the acoustic measures. Figure 3 illustrates this fact, showing the frequency estimations obtained for an excerpt of a Chaffinch bird song which exhibits a deep frequency modulation of a roughly 200 Hz rate. The estimations computed by Luscinia (in blue) are obtained by an algorithm based on the harmonic summation model and applied to a de-reverberated spectrogram (in gray). Those estimations are further refined using a recently proposed non-stationary analysis method (in red) [9].

Considering the mentioned requirements, our acoustic representation for bird songs is based on 6 different features , detailed in Table 2. In particular we propose a simple model with a single sinusoid, where features are computed each 0.5 ms. We used the energy and frequency features resulting from the Luscinia semi-supervised analysis. Vibratos segments were manually labeled. For each marked vibrato segment, frequency contour was automatically decomposed into a baseline contour (free of modulations) and a residual. This baseline contour was estimated by interpolating the frequency points of maximum absolute slope. The slope was computed by the convolution of the estimated frequency contour with a linear decreasing kernel (e.g. $[L, L-1, ..., 0, ...-L+1, -L]$). In our experiments, the kernel had a length of 15 ms. We can see in Figure 4 one example of the vibrato analysis of an excerpt of a bird song, including the estimated depth and rate. The sinusoidal amplitude also may present significant periodic or aperiodic modulations. We model those modulations with two features: tremolo depth and resonance frequency. Figure 5 illustrates this approach showing the frequency modulation and the estimated resonance frequency. Note that the amplitude peaks occur around the crossing times between the resonance frequency

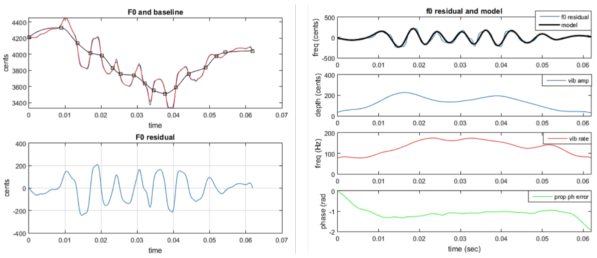


Figure 4: *Vibrato analysis.*

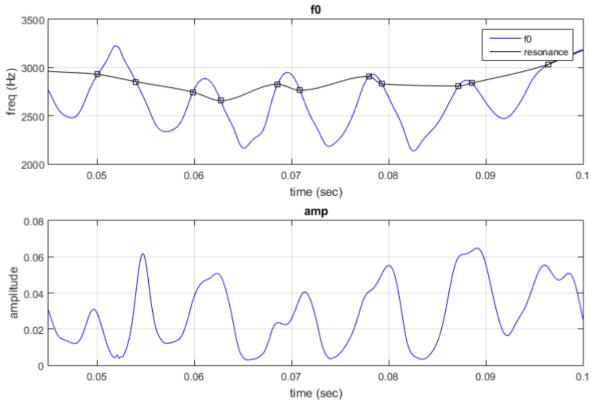


Figure 5: *Amplitude modulation analysis. The top figure shows the estimated sinusoidal frequency and resonance frequency. The bottom figure shows the amplitude feature.*

and the frequency feature.

Characteristic	Acoustic features
Sinusoid	Energy, frequency
Vibrato	Depth, rate
Tremolo	Depth, resonance frequency

Table 2: *Features considered.*

One issue we found was that for some recordings sometimes elements could overlap. One example can be seen in Figure 2, where two different frequency trajectories overlap in the marked area. Overlapping is not allowed in the HMM training since we use a single stream for the frequency feature. Thus, we opted for cutting some data, although it is not the ideal solution. This is something to be improved in the future.

2.3. HMM models and contextual factors

In general, for modeling a particular species, we should use a set of contextual descriptors that characterize the context-specificity of its vocal signals and that are aligned with the proposed requirements of the synthesizer. In speech, the acoustic features greatly depend on the context. Speech synthesizers take into account several linguistic contexts such as phoneme information, lexical stress, tone, pitch accent, and part-of-speech (POS) information. By contrast, in animal vocalizations in general and bird songs in particular, the context is simpler and the sound characteristics are less context-dependent [2].

Each recording in the database corresponds to one song.

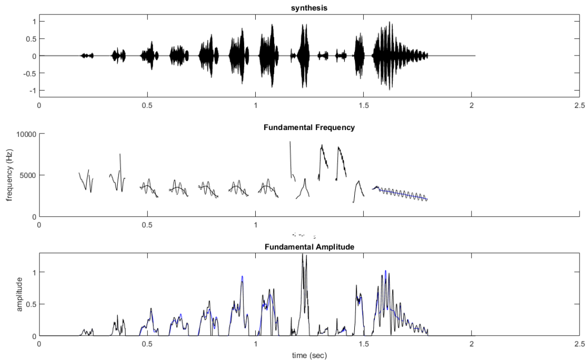


Figure 6: *f0 and amplitude estimated features*

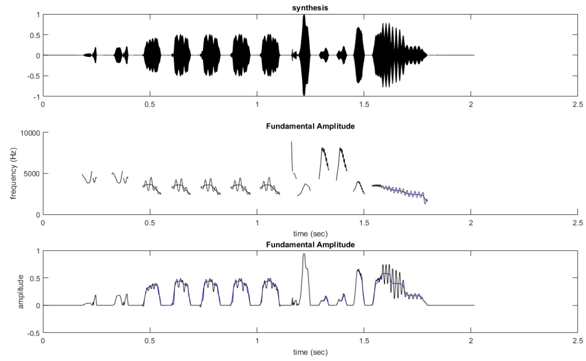


Figure 7: *f0 and amplitude features generated by the HMM models*

One limitation is that even though the hierarchical segmentation provided by experts specifies syllable and element segments, it does not identify repetitions of the same syllable. As we need to specify the context of each model using contextual labels (aka full context). In our case for HMM training, we manually inspected each song and labelled each syllable with a letter (in alphabetical order), using the same letter for identifying repetitions. This is a (rather easy) task not requiring an expert, so it can be automatized by estimating dynamic time warping based distances between syllable segments. For instance, the syllabic transcription of the song in Figure 6 is

a a b b b b b c d d e f

In addition we specify the song type by manually comparing the f0 curve between recordings. One good aspect of this approach is that the syllable name does not depend on the song type, but just on the position within the song.

As mentioned, syllables are segmented into elements. A finer segmentation is better for the HMM modeling for capturing more details of the f0 contours. In our case, HMM models are assigned to element segments, pauses within syllables and silences between syllables. We label HMM models with the syllable name plus a number indicating the position within the syllable (only if there are several elements). For Figure 6 the HMM model transcription is

sil a0 pau a1 sil a0 pau a1 sil b sil b sil
b sil b sil b sil c0 pau c1 pau c2 sil d sil
d sil e sil f sil

In the HMM training process, similar states and model parameters among several HMMs are automatically clustered in a

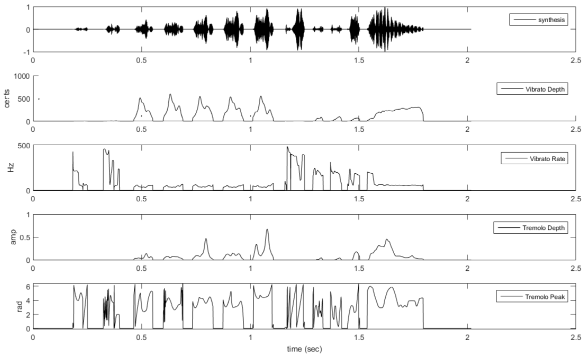


Figure 8: *Vibrato and tremolo estimated features*

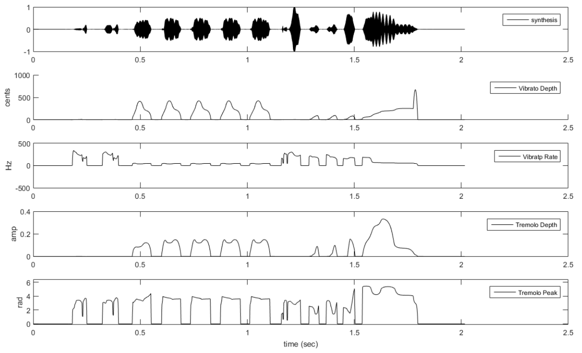


Figure 9: *Vibrato and tremolo features generated by the HMM models*

hierarchical tree structure [10] by means of the Minimum Description Length (MDL) criterion [11]. This hierarchical process is conducted answering a set of questions about the context labels. In our case, after several experiments, we ended using only three contextual labels for each HMM model: song type, syllable type and element type.

We use the HTS (HMM-based Speech Synthesis System [12]) framework to model and synthesize bird songs. The HMM configuration and training steps used in training of the model overall follow the HTS v2.3 demo scripts. The energy of the fundamental is modeled as a 1-dimensional continuous stream, f0 as a 1-dimensional multi-space distribution (MSD) stream to allow regions without pitch, and the vibrato and tremolo features are jointly modeled as a 4-dimensional continuous (non-MSD) stream as they are correlated and this way they are clustered based on the same contextual factors. In these experiments we disabled global variance in the parameter generation step. We use the standard 5 state model, but do allow a slightly lower minimum duration of 3 frames. MDL clustering (with default hyper-parameters), reduces the total of 1325 context-dependent states (over the model's 5 states) to 185 leaf nodes (around 14 %).

3. Results

We have re-synthesized the songs in our database using the trained HMM models. Figures and examples are available online from [13].

In Figures 6 and 8 we show the features estimated for a given bird song recording. For this particular recording, the syllabic transcription is

$$a\ a\ b\ b\ b\ b\ b\ c\ d\ d\ e\ f$$

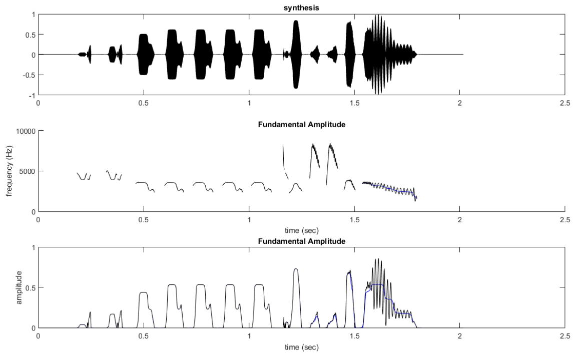


Figure 10: *f0 and amplitude features generated by the HMM models without extracting the modulations for the second syllable type.*

Figures 7 and 9 show the features generated by the HMM models out of the same transcription, in this case forcing the model duration to be aligned with the recording segmentation. As expected, HMM predicted features are smoother than input features. Time quantization in HMM states is also noticeable, especially in long elements such as those typically found in the ending part. On the other hand, modulation features seem to perform promisingly well. For instance, in Figure 10 we can see the synthesis result obtained for the same song when modulations are not extracted for the second syllable type. Clearly, the f0 contour is over smoothed, and many relevant details are missing.

It is challenging to assess how good the synthesis is. We plan to perform listening experiments with chaffinches. In general, when synthetic vocalizations are directed to an animal, one crucial aspect is that the acoustic representations should ideally consider the up-to-date scientific knowledge about the auditory perception of the target species (e.g. audible frequency range, acoustic masking mechanisms, sound level threshold, equal loudness curves, etc.). This helps to define which are the relevant characteristics to model from the acoustic signals.

4. Conclusions

We have presented a method for bird song synthesis based on Hidden Markov Models, discussing on the challenges and strategies followed at the different stages of the process.

Regarding the acoustic representation, features were computed in a semi-supervised manner, requiring the supervision of an expert to provide meaningful and robust features. Two relevant aspects for facilitating a general methodology in the future are to automate the feature extraction process and to provide a robust analysis. Regarding modelling strategies, an interesting refinement is to model call sequencing with Markov chains or first-order HMMs (as proposed in [4]).

One interesting direction to investigate, in the context of human-animal interaction and sound design, is how to generate call sequences from human vocal imitation or sketching (e.g. analogously to [14], build a parallel database of animal vocalization and human imitations, and train models for both), as well as methods to embed human emotions in the synthetic vocalizations.

5. References

[1] K. Tokuda, Y. Nankaku, T. Toda, H. Zen, J. Yamagishi, and K.Oura, "Speech synthesis based on hidden markov models," *Proceddings of the IEEE*, vol. 101, no. 5, pp. 1234–1252, May 2013.

[2] P. Marler, *The origins of music*. Cambridge, MA: MIT Pres, 2000, ch. Origins of Music and Speech: Insights from Animals (eds N. Wallin, B. Merker and S. Brown), pp. 31–48.

[3] C. ten Cate and K. Okanoya, "Revisiting the syntactic abilities of non-human animals: Natural vocalizations and artificial grammar learning," *Philosophical Transactions of The Royal Society B*, vol. 367, pp. 1984–1994, 2012.

[4] K. Katahira, K. Suzuki, K. Okanoya, and M. Okada, "Complex sequencing rules of birdsong can be explained by simple hidden markov processes," *PLoS ONE 6, e24516*, 2011.

[5] A. Amadorl, Y. S. Perl, G. B. Mindlin, and D. Margoliash, "Elemental gesture dynamics are encoded by song premotor cortical neurons," *Nature 495*, pp. 59–64, 2013.

[6] T. Smyth, J. S. Abel, and J. O. S. III, "The estimation of birdsong control parameters using maximum likelihood and minimum action," in *Proceedings of the Stockholm Music Acoustics Conference (SMAC)*, Stockholm, Sweden, August 2003.

[7] R. Lachlan, "Luscinia software." [Online]. Available: http://rflachlan.github.io/Luscinia

[8] J. G. Beckers and C. ten Cate, "Nonlinear phenomena and song evolution in streptopelia doves," *Acta Zoologica Sinica*, vol. 52(Supplement), pp. 482–485, 2006.

[9] S. Musevic and J. Bonada, "Distribution derivative method for generalised sinusoid with complex amplitude modulation," in *Proceedings of 18th Int. Conference on Digital Audio Effects (DAFx-15)*, Trondheim, Norway, 2015.

[10] J. J. Odell, "The use of context in large vocabulary speech recognition," in *PhD Thesis, Queens' College, University of Cambridge*, Cambridge, U.K., 1995.

[11] K. Shinoda and T. Watanabe, "Mdl-based context-dependent subword modeling for speech recognition," *Journal of Acoustic Society of Japan (E)*, vol. 21, no. 2, pp. 79–86, 2000.

[12] "Hts software." [Online]. Available: http://hts.sp.nitech.ac.jp

[13] "Examples of bird song synthesis based on hmm." [Online]. Available: http://www.dtic.upf.edu/~jbonada/HMMBirdsong16

[14] T. Nose and T. Kobayashi, "Speaker-independent hmm-based voice conversion using adaptive quantization of the fundamental frequency," *Speech Communication*, vol. 53, no. 7, pp. 973–985, 2001.