CS447 Literature Review: How can Natural Language Processing (NLP) be Used to Understand the Linguistics of Birdsongs

Chris Heejun Park,
hpark102@illinois.edu

November 29, 2021

Abstract

There are significant challenges to model songbird song NLP, let alone a music. The challenge presents its opportunity for researchers to investigate the song's complex sequential rules using different variations of Hidden Markov (HM) processes, simple recurrent network (SRN), a modified Chomsky Hierarchy (CH), and a unique representation of songs, a partially observable Markov model (POMM). This review explores each studies' natural language processing models that could explain the complex songs by various songbird species.

1 Introduction

Satisfactory, explanatory, and descriptive models have yet to be found for the complex vocal sequences of birdsongs, which constitute ways of sonic communication that evolved a remarkable degree of structural complexity. Jin and Kozhevnikov study shows complex action sequences in animals and humans are often organized according to syntactical rules that specify how actions are strung together into sequences. Many examples are found in birdsong. Songs of songbird species such as Bengalese finch, sedge warbler, nightingale, and willow warbler, consist of a finite number of stereotypical syllables (or notes) arranged in variable sequences. Quantitative analysis of the action syntax is critical for understanding how complex sequences are generated. Katahira et al. (2011) believes a NLP model on Bengalese finches song that has the more complex sequencing rules with branching points can explain the complex sequencing rules in speech and musical performance. Berwick et al. (2011) sees the comparison between birds songs and humans linguistics. They both consist of complex, patterned vocalizations. such sequential structures can be analysed and compared via formal syntactic methods. NLP models capable of explaining the complex linguistics of birdsongs can not only generate songs but also provide useful perspective that the choice of model determines which aspects of structure building we are comparing across domains. It's more appropriate to decide on a NLP model for a certain purpose than to choose the best overall model. Honing certain NLP models for specific tasks serve to solve underlying problems in the NLP models in linguistics.

2 Background

These studies share the same achievement of conducting independent studies on songbird songs to create NLP models explaining the songs while aware of other similar studies method of approach and results. Vocalizations of the songbirds were recorded using microphones to obtain spectrogram or sonogram. Experts annotate syllables from the visuals and pauses between notes of the birdsongs. Depending on the hierarchical structure of choice by the authors, the syllables are labeled accordingly. The authors treat the retrieved motifs, which are repeated sequences of syllables, on the spectrum of the tripartite diagram of abstract components encompassing birdsong and human language or the CH rings. Lastly, they base their NLP models on the statistical properties of the syllable sequences. I share these studies under the literature reviews section with caution as I have not conducted my own samples of birdsongs. With this limitation, I focus on the significance of the authors focus on their approach method and NLP models with critical view.

3 Literature Reviews

3.1 A Compact Statistical Model of the song syntax in Bengalese finch

The authors, Jin and Kozhevnikov (2011), analyze the song syntax in two Bengalese finches. They show that the Markov model, even with attention, fails to capture the statistical properties of the syllable sequences. In contrast, they show that the statistical properties of the sequences are well captured by a state transition model, the partially observable Markov model with adaptation (POMMA), an extension of the Markov model, in which the repeat probabilities of the syllables adapt and many-to-one mappings from the states to the syllables are allowed.

To obtain data, spontaneous vocalizations of two Bengalese finches were recorded in an acoustic chamber using a microphone. They obtained the Bengalese finch song's acoustically continuous segments, called "song elements" or syllables" which are separated by silent intervals. They adopted two-stage methods for extracting syllable sequences from audio signals: they first assigned syllable labels for the audio signals of segmented syllables, such as A, B, C, and etc... by using an automatic clustering method and then they constructed syntax models for describing the sequences of syllable labels.

The syntactical rules of the songs of songbird species such as nightingale, willow warbler, Bengalese finch, and sedge warbler are often characterized by the Markov model, which each syllable is associated with one state, the transitions between the states are stochastic, and the transition probabilities between the states do not depend on the state transition history. They show that this Markov model fails to capture the statistical properties of the syllable sequences, including the probability of observing a given syllable at a given step from the start of the sequences, the repeat number distributions of individual syllables, and the distributions of the n-grams (sequences of length N).

Instead, introducing two modifications to the Markov model show that the extended model is successful in describing the syntax of the Bengalese finch songs (Figure 1). The first modification is adaptation, similar to domain adaptation in natural language processing (NLP). Since syllable repetitions are common in the Bengalese finch songs, allowing the repeat probabilities of syllables to decrease with the number of repetitions leads to a better fit of the repeat number distributions. The second modification is many-to-one mapping from the states to the syllables. A given syllable can be generated by more than one state. Even if the transitions between the states are Markovian, the syllable statistics are not Markovian due to the multiple representations of the same syllables.

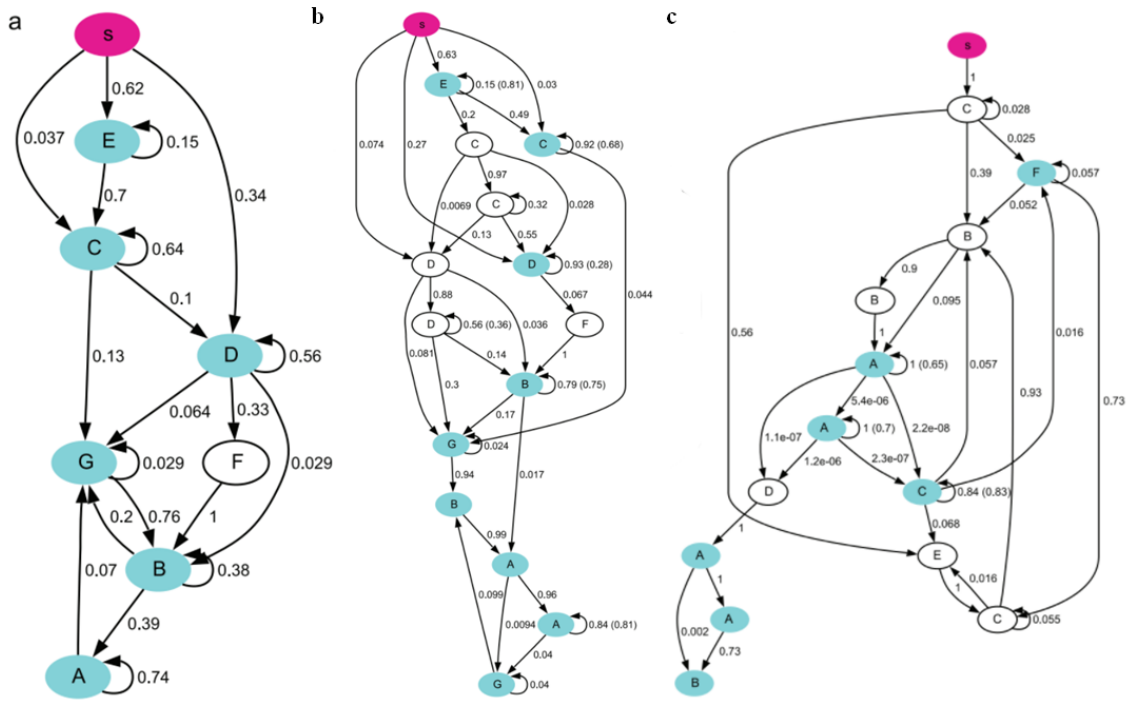

Figure 1: **a.** The Markov model. The pink oval represents the start state. The cyan ovals are the states with finite probabilities of transitioning to the end state. The numbers near the transition lines indicate the transition probabilities. **b.** The POMMA derived from the observed syllable sequences with the repetitions fitted with adaptation models by Bird 1. **c.** The POMMA derived from the observed syllable sequences with the repetitions fitted with adaptation models by Bird 2.

The resulting model, a partially observable Markov model with adaptation (POMMA), has history-dependent transition probabilities between the states and many-to-one mappings from the states to the syllables. This state transition model that accurately describes the statistics of the syllable sequences includes adaptation of the self-transition probabilities when states are revisited consecutively, and allows associations of more than one state to a given syllable.

The three different models, Markov model, Markov model with adaptation, and POMMA, were compared by three different statistical properties (Figure 2). The authors compared the distribution of the syllable repeats, the n-gram distribution, and the step probability of the syllables. To evaluate how well the models describe the statistics of the observed song sequences, they generated 1000 sequences from the models and compared three statistical properties of the generated sequences and the observed sequences. To quantify the difference between distributions, they defined the maximum normalized difference d, which is the maximum of the absolute differences divided by the maximum values in the two distributions, i.e. $d = max_x|f_1(x) - f_2(x)|/max(f_1(x), f_2(x))$. A good statistical model of the syllable sequences should produce d-values smaller than the benchmarks or close to them. The d-values obtained from the Markov model and the Markov model with adaptation are mostly far beyond the benchmarks. It is clear that these models fail to capture the statistical properties of the songs. In contrast, the POMMA accurately describes the statistical properties of the syllable sequences. The d-values between the model and the observed distributions are below or very close to the benchmarks obtained from the observed sequences. In conclusion, they have

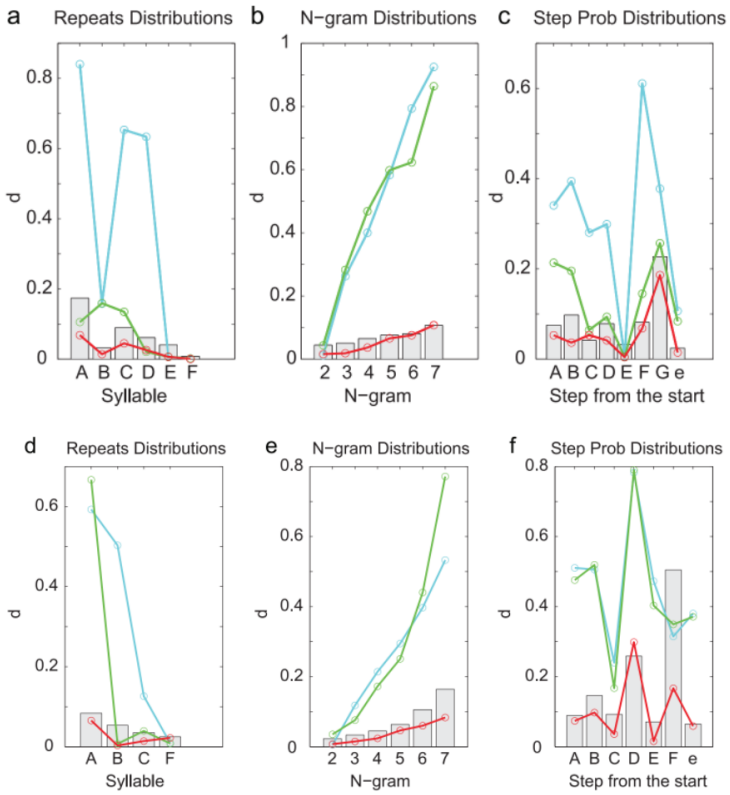

Figure 2: **a.** The Markov model. The pink oval represents the start state. The cyan ovals are the states with finite probabilities of transitioning to the end state. The numbers near the transition lines indicate the transition probabilities. **b.** The POMMA derived from the observed syllable sequences with the repetitions fitted with adaptation models by Bird 1. **c.** The POMMA derived from the observed syllable sequences with the repetitions fitted with adaptation models by Bird 2.

derived a compact POMMA that successfully describes the statistical properties of Bengalese finch songs.

3.2 Complex Sequencing Rules of Birdsong Can be Explained by Simple Hidden Markov Processes.

The authors, Katahira et al. (2011), analyze the complex sequencing rules in 16 normal adult male Bengalese finches. They applied hidden Markov models (HMMs) with various context dependencies to the acoustic features of a Bengalese finch song and selected a suitable model based on the degree of agreement with manual annotation, the Bayesian model comparison, and its predictive performance. As a result, they found that the first-order HMM, in which the current state appears depending only on the last state, is sufficient and suitable for describing the Bengalese finch song. Although the first-order HMM may seem counter-intuitive since the song sequences have higher-order dependency, the many-to-one state mapping to syllables enables the HMM to generate apparently complex sequences.

To gather data, They obtained the Bengalese finch song's acoustically continuous segments,

called “song elements” or syllables” which are separated by silent intervals. Based on visual inspection on the sonogram, the acoustically similar syllables were labeled as such A, B, C, and etc.... Following this approach, they first analyzed the statistical properties of the syllable label strings. They then directly analyzed the acoustic features using statistical models and compared the results to those of an analysis on manual annotated labels. The manual annotated labels based on visual inspection on sonogram by experts were cross checked by computing Fleiss’s k coefficient, which measures the degree of agreement among more than two annotators. As a result, the k coefficients were 0.972 ± 0.028 for all the 16 birds, and all within the range of “Almost perfect agreement,” indicating annotation was reliable. The authors use the labeling results by only one of the labeling experts.

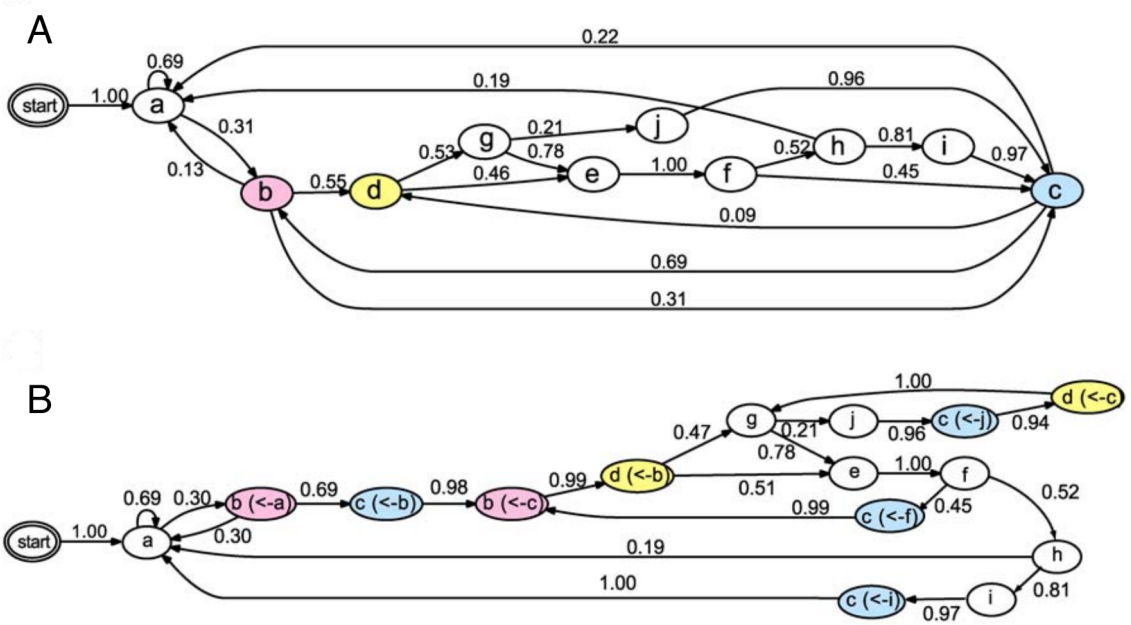

Figure 3: **a.** Bigram automaton representation of syllable sequences obtained from song set. **b.** POMM representation of same sequences.

The most effective models they considered were the first-order HMM and second-order HMM. They conclude that the sequencing rules of the Bengalese finch song have higher order Markov dependency, and cannot be described using a simple Markov process, where states and syllables have one-to-one mapping. With the help of a model called the partially observable Markov model (POMM) representation, the second-order context-dependency can be visually captured by splitting the syllables into distinct states depending on the preceding states. For example, it may seem in Figure 3A that transition from syllable “b” to “a”, “d”, and “c” are random, but with POMM representation (Figure 3B), we can capture the tendency that states b(←a) and b(←c) depending on the preceding syllables (a or c).

However, combining two succeeding states into one context states can transform this graphical model into one with the same form as the first-order HMM. Even if the hidden state sequences of the first-order HMMs have only first-order dependency, the emitting syllable acoustic features can have higher-order dependency. This can occur when the different hidden states (States 2 and 4 in Figure 4) have similar emission distributions (corresponding syllable “b” in Figure 4). Although the

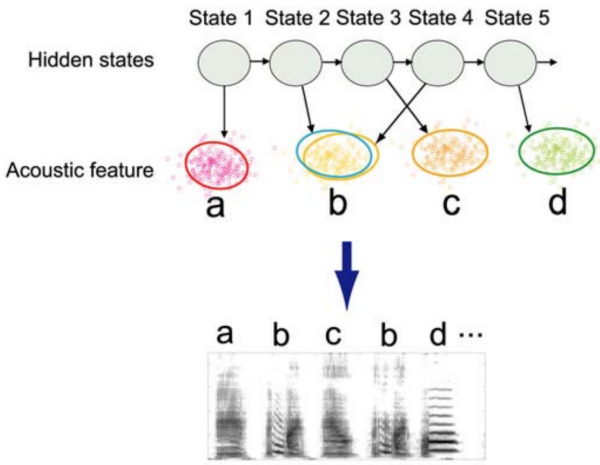

Figure 4: States 2 and 4 can generate similar acoustic feature space (”b”). This mechanism allows observed song sequences to have higher-order context dependency, even if hidden state sequences are generated from simple Markov process.

hidden state sequence is a first-order Markov sequence (“12345...”), the emitted syllables can have second-order dependency (“abcbd...”). This representation was that the first-order HMM was indeed attained through an automatic parameter fitting process using the Variational Bayes method. Their results suggest that the statistical structure of the Bengalese finch song is close to the first-order HMM.

One could notice the different models by Jin and Kozhevnikov (2011) and Katahira et al. (2011). The studies independent from each other propose different models on Bengalese finch songs. Their models share similar mechanism of many-to-one mapping to generate sequences with higher-order history dependency. However, the differences in the method of extracting sequence structures from audio features, consideration of state transition that obey higher-order Markov processes, and recognition of repetitive syllables resulted in two different models.

3.3 Songs to syntax: the linguistics of birdsong.

The authors show that birdsong structure is best characterized as ‘phonological syntax,’ which shares the hierarchically organized human language with particular syntactic constraints. Their study suggests that birdsong syntax lies well beyond the power of bigram descriptions, but is at most only as powerful as k-reversible regular languages, lacking the nested dependencies that are characteristic of human syntax. This is probably because of the lack of semantics and words in birdsong, because song sequence changes typically alter message strength but not message type.

In birdsongs’ songs, individual notes can be combined as particular sequences into syllables, syllables into ‘motifs’, and motifs into complete song ‘bouts.’ Songs might consist of fixed sequences with only sporadic variation (e.g. zebra finches), or more variable sequences (e.g. starlings, or Bengalese finches, nightingales), where a song element might be followed by several alternatives, with overall song structure describable by probabilistic rules between a finite number of states.

Through their studies of songs of nightingales, zebra finches, starlings, and other songbirds, under the resulting ‘rings’ under the Chomsky hierarchy, the sets of strings of songs overlaps the finite languages (most inner ring) and the finite-state automaton (FSA) generating the regular languages

(enveloping the inner ring). For the majority of the songbirds, birds form simple songs by using two bigram templates: [a-b] and [b-a], the simplest kind of pattern recognizable by a FSA. For instance, skylarks mark individual identity by particular song notes, as starlings do with song sequences; and canaries use special 'sexy syllables' to strengthen the effect of mate attraction. A more complex Bengalese finch song explicitly shows the probability that one state follows another via the numbers on the links between states. Another more complex song by nightingales can sing motifs with notes that are similarly embedded within looped note chunks. The finch and nightingales construct a particular kind of restricted FSA generating the observed sequences (a *k*-reversible FSA). Thus, birdsong might best be regarded as 'phonological syntax', a formal language; that is, a set of units (here acoustic elements) that are arranged in particular ways but not others according to a definable rule set.

Regarding the birdsong as a 'phonological syntax', the author's *k*-reversible FSA notices the songbirds acoustic domain, rearranging existing 'chunks' to produce hundreds of distinct song types that might serve to identify individual birds to measuring the degree of sexual arousal. Song variants do not result in distinct meanings with completely new semantics, but serve to modify the entirety of the song within the context of its behavior and communication system.

3.4 Formal models of Structure Building in Music, Language and Animal Song.

The authors, Zuidema et al. (2019), bring together accounts of the principles of structure building in songbird songs, relating them to the corresponding models in formal language theory, with a special focus on evaluating the benefits of using the Chomsky hierarchy (CH). They further discuss shortcomings concerning the CH, as well as extensions or augmentations of it that address some of these issues, including extending the CH with more fine-grained classes, the addition of probabilities and meaning representations to symbolic grammars, and replacing abstract symbols with numerical vectors.

The building blocks of songs combine units, which can be identified by numerous methods, of sound into larger units in a hierarchical way. A common way of identifying units in animal songs is to study their spectrogram and delineate units based on acoustic properties such as silent gaps.

For many bird species, bigrams in fact seem to give a very adequate description of the sequential structure. *n*-gram models seem to suffice for modelling songs of for instance mistle thrushes and zebra finches. It is however often argued that they are unable to model certain key structural aspects of songs by other bird species. In addition, *n*-grams are inadequate models of the structure of the vocalisation on both the syllable and phrase.

The CH dismisses *n*-grams as useful models of syntactic structure in natural language. It argues against incorporating probabilities into language models and concerns the symbolic, syntactic structure of language. It proposes an idealisation of natural language where a language is conceived of as a potentially infinite set of sentences, and a sentence is simply a sequence of words (or morphemes).

Richer models are needed to characterize the vocalisations of certain bird species. However, this difference in complexity is not captured by the CH, as the complex models proposed are still finite-state models that fall into the lowest complexity class of the CH: regular languages. This issue can be addressed by describing a hierarchy of sub-regular languages that contains the set of strictly local (SL) languages, which constitute the non-probabilistic counterpart of *n*-gram models.

An entirely different approach to modelling natural language - parallel to the symbolic one employed by the Chomsky hierarchy - is one where the symbols and categories of the CH are replaced by vectors and the rules are projections in a vector space (implicitly) defined in matrix vector algebra. Thus, instead of having a rule 'Z $\rightarrow$ X Y', where X, Y and Z are symbolic objects

(such as a 'prepositional phrase' (PP) in linguistics, or a motif in a zebra finch song), we treat them as n-dimensional vectors of numbers.

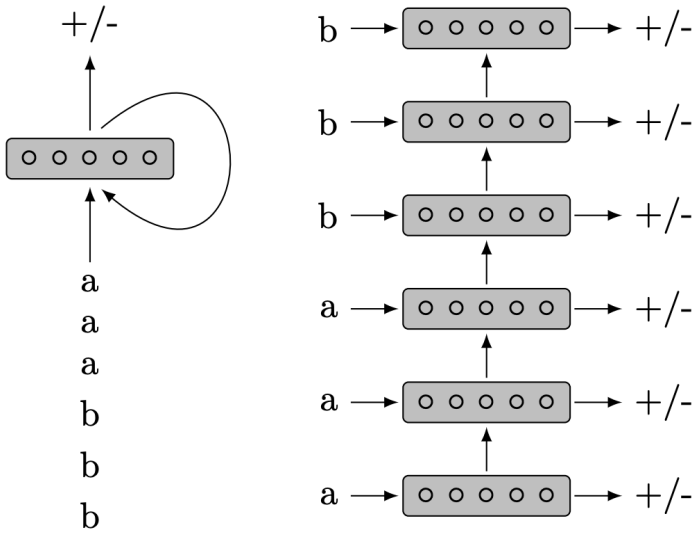

Figure 5: **Left)** A simple recurrent network (SRN). **Right)** The same network, but unfolded over time.

The expressivity of the model reduces the need for more complex architectures, as vector grammars are computationally much more expressive than symbolic systems with similar architectures. For instance, a simple recurrent network (or SRN, on an architectural level similar to an FSA) in Figure 5 can implement the counter language $a^n b^n$, a prime example of a context-free language. The network receives a sequence of inputs (in this case "a a a b b b") and outputs whether this is a grammatical sequence (+/-). The arrows represent so-called "weight matrices" that define the projections in the vector space used to compute the new vector activation from the previous ones. Vector grammars provide a motivation to move on to probabilistic, non-symbolic models that go beyond the constraints of the CH.

The authors lay out the commonalities as well as differences between linguistics and between species. Uncovering many useful models for how sequences of sound might be generated and processed, they've come to a graded category model that add probabilities, semantics and graded categories to classical formal grammars. It links neutral network approaches with formal grammar to deal with inherently continuous structures.

4 Discussion

Extensive studies by the authors in the Literature Reviews (Section 3) explains the songs of songbird species using existing NLP models. Testing these existing models to the retrieved data, songbirds syllable, they discover the limitation and shortcoming in explaining the songs sequence, syntax, or grammar. This leads to necessary modification to better the models of songbirds. The authors performed statistical modelling tests by either comparing the modified model's generated songs to the existing models or generating a more efficient model architecturally from the existing ones. Although confident in their own models, they acknowledge that one model does not fit all the songs

of many songbird species. They believe that the models are significantly different from the NLP models explaining human languages, but understanding these models can be the building block to answering some of the questions occurring in current NLP models on human languages.

5 Conclusion

To non-experts, the linguistics of birdsongs may seem simpler than human linguistics that existing NLP models are sufficient to generate songs similar to ones by songbirds. To test this theory, the authors prepare/retrieve syllables of songbirds and generate songs using existing models only to find underlying problems with using simple NLP models to explain the complex sequences of songbird syllables. The literature review's extensive analysis on the birdsong syllables prove that modifications to the current NLP models' architecture and statistical properties are necessary to understand birdsong linguistics and generate songs.

References

Robert C. Berwick, Kazuo Okanoya, Gabriel J. L. Beckers, and Johan J Bolhuis. 2011. Songs to syntax: the linguistics of birdsong. volume 15, pages 113–121.

Dezhe Jin and Alexay Kozhevnikov. 2011. A compact statistical model of the song syntax in bengalese finch. volume 7, page e1001108.

Kentaro Katahira, Kenta Suzuki, Kazuo Okanoya, and Masato Okada. 2011. Complex sequencing rules of birdsong can be explained by simple hidden markov processes. volume 6.

Willem Zuidema, Dieuwke Hupkes, Geraint Wiggins, Constance Scharff, and Martin Rohrmeier. 2019. Formal models of structure building in music, language and animal songs.