How Blue Can You Get? Learning Structural Relationships for Microtones via Continuous Stochastic Transduction Grammars

Dekai Wu
HKUST
Human Language Technology Center
Department of Computer Science and Engineering
Hong Kong University of Science and Technology
dekai@cs.ust.hk

Abstract

We describe a new approach to probabilistic modeling of structural inter-part relationships between continuous-valued musical events such as microtones, through a novel class of *continuous* stochastic transduction grammars. Linguistic and grammar oriented models for music commonly approximate features like pitch using discrete symbols to represent 'clean' notes on scales. In many musical genres, however, contextual relationships between continuous values are essential to improvisational and accompaniment decisions—as with the 'bent notes' that blues rely heavily upon. In this paper, we study how stochastic transduction grammars or STGs, which have until now only been able to handle discrete symbols, can be generalized to model continuous valued features for such applications. STGs are interesting for modeling the learning of musical improvisation and accompaniment where parallel musical sequences interact hierarchically (compositionally) at many overlapping levels of granularity. Each part influences decisions made by other parts while at the same time satisfying contextual preferences across multiple dimensions; applications to flamenco and hip hop have recently been shown using discrete STGs. We propose to use a formulation of continuous STGs in which musical signals are finely represented as continuous values without crude quantization into discrete symbols, yet still retaining the ability to model probabilistic structural relations between multiple musical languages. We instantiate this approach for the specific class of stochastic inversion transduction grammars (SITGs), which has proven useful in many applications, via a polynomial time algorithm for expectation-maximization training of continuous SITGs.

Introduction

Musical improvisation is the creative activity of spontaneous, on-the-fly musical composition without prior planning, in response to a novel context (typically provided by other musicians, who are often also improvising), in contextually relevant ways that adhere to stylistic conventions, yet are not constrained by *a priori* written scores. Throughout most of history, creative improvisation has been the norm in many, if not most, traditional and folk forms of music (unlike Western music in recent centuries, where written music has been a historically recent artifact). Musical improvisation is a uniquely human behavior, exhibiting creative expression that has not been found in other "singing" species.

It can be relatively easy to construct automatic music generation algorithms that can be parametrized by various conditions and constraints. On one hand, some approaches rely on manually constructed rules; these approaches can represent fairly complex kinds of structures and patterns, but the improvisation is limited to the rules that have been imagined by experts and hand coded in advance, which can only crudely be matched to true human improvisation. On the other hand, other approaches employ machine learning; these approaches attempt to match their performance more finely to human improvisation by training contextual predictors on actual music data, but improvisation tends to be restricted to what can be modeled via fairly simple representations such as HMMs to limit the complexity of the learning.

The problem is that real musical improvisation at human levels requires both complex structures and patterns, and also contextual prediction that is finely tuned to human performance data. Improvisational and accompaniment decisions in one part can be influenced strongly, or subtly, by decisions made in other parts, interacting hierarchically or *compositionally* at many overlapping levels of granularity. Improvisation and accompaniment decisions are not merely random; rather, participants understand how to communicate with each other within accepted conventions and frameworks—witness, for example, flamenco *palos*, Indian *ragas*, jazz and blues. Conventions in widespread use include tonal systems, metrical constraints, chord progressions, verse structures, rhythmic patterns, and melodic phrases that are re-used or swapped into different positions within the structures. Making improvisation decisions that integrate interacting contextual factors over many levels of granularity requires a representation that can encode such sophisticated phenomena, yet will not blow up machine learning complexity exponentially.

To attack this challenge, we are engaged in a long-term program to develop a general mathematical framework for creative improvisation, that is capable of representing a realistically broad range of the many different complex interactions among factors that should influence the improvisation, and yet which can still support efficient polynomial-time training and improvisation algorithms—so that ultimately, we should be able to build more realistic models of learning to improvise. A full solution to the representation, learning, and improvisation problems will obviously require many advances, but we have already begun to show how various aspects of these tasks can be accomplished, via bilingual **stochastic transduction grammar** or **STG** models that can

simultaneously capture contextual preferences across a wide variety of dimensions. In our work on hip hop learning models (Wu *et al.*, 2013), we showed how **stochastic inversion transduction grammars** or **SITGs** can be used to learn how to improvise responses in freestyle rap battling when confronted with arbitrary challenge raps, by learning complex relationships between challenges and responses. In our work on flamenco learning models (Wu, 2013), we showed how SITGs can be used to learn how to improvise complementary lines in, for example, *palmas* percussion in the context of perceiving *cajón* percussion, by learning complex hypermeter and rhythm biases in the relationship between the languages of different percussion instruments.

Applications like these have demonstrated how STGs (a) have the expressiveness to represent compositionally interacting factors between two different parts or instruments at many overlapping levels of granularity, (b) can be efficiently induced via the polynomial-time learning algorithms that exploit the combinatorial structure of SITGs, and (c) can then use the learned knowledge representation to creatively perform real-time improvisational expression. For capturing the complexity of hierarchical structural relationships between different musical languages, the linguistic bilingual approaches of STGs have many appealing properties. They allow idiomatic constructs of significant complexity to be encoded. They allow biasing of probabilities from many different contextual features. They allow idiomatic constructs to be combined in creative new ways inspired by the unplanned contextual factors. They accommodate correlations that are not necessarily aligned in time, which make them significantly more expressive than context-free grammars (CFGs); this is why the basic time complexities for stochastic CFG recognition and training are $O\left(n^3\right)$, in contrast to $O\left(n^6\right)$ for SITGs. Musical improvisation modeling approaches based on SITGs benefit from leveraging several decades of advances in the field of statistical machine translation, which exhibits very analogous challenges.

However, all SITG based models to date over the past two decades have exhibited a glaring weakness when it comes to learning creative improvisation knowledge in the domain of music: they are only capable of representing sequences of discrete symbolic events. This was not an obstacle in the rap battle improvisation domain, where words and phrases were modeled by discrete symbols. Likewise, it was not an obstacle in the flamenco improvisation domain, where each percussive event was modeled by a discrete symbol. However, this is a major limitation in the music domain in general, where the overwhelming majority of events are continuous values like pitch, timbre, or volume.

This paper proposes for the first time a formulation of SITGs that (a) have the expressiveness to represent compositionally interacting factors between different *continuous valued* parts or instruments at many overlapping levels of granularity, and yet (b) can still preserve all the aforementioned advantages of SITGs, including efficiently induction via the polynomial-time learning algorithms.

The motivation for this is that if continuous-valued probabilistic structured associations can be learned, then creative improvisation algorithms can be developed along analogous designs to those previously developed for discrete events. Our new approach bridges the gap between (a) computational models that leverage linguistic approaches to describing the complex structural relationships between different musical parts or languages, and (b) computational models that realistically describe truly continuous valued musical events such as pitch or volume. This gap is presently one of the impediments in modeling many creative decisions necessary in live musical improvisation and accompaniment.

Continuous stochastic transduction grammars represent perhaps the first completely integrated models that are capable of finely representing musical events as continuous values while modeling probabilistic structural compositional relations between multiple musical languages. Crude quantization into discrete symbols is no longer necessarily needed in STG modeling.

We illustrate (a) the new representational approach, and (b) the new EM training algorithm for continuous STGs. To illustrate how the new formulation works, we consider an example inspired by that fact that the same traditional and folk genres in which improvisation plays an important role also very often make heavy use of microtonal pitches, as opposed to 'clean' notes on a discrete scale. The degree to which notes are 'bent' may depend on a host of contextual factors both within and between parts, at various different granularities of musical structure. To approach human levels of improvisation quality, or to advance musicological studies, truly integrated computational modeling must natively handle not only discretized symbols but also continuous values.

We show how the relationship in blues music between microtonal melody pitches ('bent notes') and bass pitches can be modeled, by instantiating the idea of continuous STGs for a particularly useful kind of STGs known as a **stochastic inversion transduction grammar** or **SITG**. This is motivated by the fact that SITGs have been empirically shown over decades to exhibit an excellent balance of expressiveness and inductive biases, while maintaining practical polynomial computational complexity characteristics (including statistical machine translation, as well as the hip hop and flamenco models mentioned above). This enables an efficient polynomial time algorithm for expectation-maximization training of continuous SITGs.

Stochastic transduction grammars

Transduction grammars can be seen in the generative modeling paradigm of GTTM (Lerdahl and Jackendoff, 1983) and Steedman Steedman (1984) or Steedman (1996) in using formal grammars to model musical sequences—but instead of monolingual modeling of a single musical language, transduction grammars represent bilingual modeling of the *relationship* between two musical languages.

This makes sense because music is not primarily about a single sequence. Rather, what makes music *musical* more often than not concerns the loosely coupled relationships between parallel strands of different kinds of sequences. Transduction grammars are by nature *bilingual*, which renders them ideally suited for modeling the complex structural relationships between different musical sequences.

Just like natural language, music is highly nondeterministic. As with language, stochastic versions of transduction grammars must be used for any but the most trivial models of music. Much of the previous work on stochastic grammatical modeling of music has been based on flat Markov models and/or hidden Markov models (HMMs). The Continuator model of Pachet (2003) and the Factor Oracle models of Assayag *et al.* (2006) and Assayag and Dubnov (2004) both learned music improvisation conventions using Markov models, later further explored by François *et al.* (2007) and François *et al.* (2010). Jazz grammars were induced by Gillick *et al.* (2010) also under Markovian assumptions.

Much less has been done on modeling of musical structure via stochastic context-free grammars Lari and Young (1990). Unsupervised learning of CCMs (a variant of SCFGs) for musical grammars was described by Swanson *et al.* (2007) and in the DOP approach originally proposed by Bod (2001).

The work on machine learning of stochastic transduction grammars originated largely in the statistical natural language processing community. Stochastic transduction grammars generalize stochastic grammars to model two streams instead of one. As transduction grammars are strictly more powerful than their corresponding monolingual grammars, they are capable of modeling anything that stochastic grammars can model. **Inversion transduction grammars** or ITGs (Wu, 1997) are a subclass of syntax directed transduction grammars or SDTGs (Lewis and Stearns, 1968) that generalize context-free grammars to the bilingual case. Stochastic ITGs, or SITGs, are the bilingual generalization of stochastic CFGs and have proven extremely effective in machine translation as well as other NLP applications.

Whereas the production rules in CFGs probabilistically generate a monolingual subtree, the transduction rules in STGs probabilistically generate both input and output language subtrees. Just as in CFGs, subtrees are generated by recursively combining smaller subtrees (which describe the compositional structure of aligned input and output chunks) into larger subtrees. But unlike in monolingual CFGs, each leaf of a parse tree is a preterminal representing a bilingual *pair* of atoms, as opposed to simply a monolingual atom.

ITGs restrict the alignment between the children of any internal node to be only straight or inverted, rather than arbitrary permutations. This ITG restriction empirically (and somewhat surprisingly) provides sufficient alignment flexibility between the input and output language atoms across virtually every pair of natural languages (Zens and Ney, 2003; Saers *et al.*, 2009; Addanki *et al.*, 2012), but unlike general SDTGs, yields tractable polynomial time training and translation algorithms.

Stochastic transduction grammars appear quite promising for learning of probabilistic structural relations between musical languages. Wu (2013) learned a SITG that discovered structural relationships between flamenco *cajón* and *palmas* languages via transduction grammar induction driven by a Bayesian MAP (maximum *a posteriori*) criterion, in which metrical relations, hypermetrical relations, and probabilistic transduction relations were simultaneously integrated. Wu *et al.* (2013) used SITG induction to automatically learn hip hop freestyling by discovering structural relationships between challenge and response rap languages. However, these models suffer from the weakness mentioned above of only being able to model non-continuous musical information that can be represented in terms of discrete symbols.

Continuous STGs

We now describe how continuous stochastic transduction grammars represent continuous-valued musical information at various levels of structural granularity within an integrated model, by generalizing a step at a time from context-free grammars. For greater detail on the formal properties of STGs, the reader is referred to (Wu, 1997) and (Wu, 2010).

In the well-known **twelve-bar blues** form, verses consist of three lines: a first four bars, a second four, and a third four called a turnaround. The following syntactic rules, in a conventional context-free grammar, describe a twelve-bar blues in its typical 'quick to four' variant:

$$
\begin{aligned}
S &\rightarrow \text{VERSE} \\
S &\rightarrow [\text{VERSE } S\,] \\
\text{VERSE} &\rightarrow [\text{FIRST8 TURNAROUND}] \\
\text{FIRST8} &\rightarrow [\text{FIRST4 SECOND4}] \\
\text{FIRST4} &\rightarrow [\text{AD AA}] \\
\text{SECOND4} &\rightarrow [\text{DD AA}\,] \\
\text{TURNAROUND} &\rightarrow [\text{ED AA}] \\
\text{AA} &\rightarrow [A\ A] \\
\text{AD} &\rightarrow [A\ D] \\
\text{DD} &\rightarrow [D\ D] \\
\text{ED} &\rightarrow [E\ D]
\end{aligned}
$$

We can generalize this to a bilingual transduction grammar that expresses the relationship between, for example, a bassline language and a vocal melody language. Ordinary grammars have preterminal symbols corresponding to the monolingual lexical atoms of a single language. On the other hand, transduction grammars have bilingual preterminal symbols corresponding to a relation between two lexical atoms from two *different* languages, which is called a **biterminal**. Let us further decompose the nonterminal symbol A, which represents a single bar in the tonic, into a finer grained series of frames—we'll use eighth note durations for simplicity's sake here, though we could also use much finer granularities:

$$
\begin{aligned}
A &\rightarrow [\text{AT BU CV DW EX FY GZ H0}] \\
\text{AT} &\rightarrow a/t \\
\text{BU} &\rightarrow b/u \\
\text{CV} &\rightarrow c/v \\
\text{DW} &\rightarrow d/w \\
\text{EX} &\rightarrow e/x \\
\text{FY} &\rightarrow f/y \\
\text{GZ} &\rightarrow g/z \\
\text{H0} &\rightarrow h/\epsilon
\end{aligned}
$$

The preterminal AT, for instance, generates the biterminal a/t which stands for a bassline language atom a, representing some bass note, that is associated with a melody

language atom t, representing some melodic note. The special empty symbol ϵ, represents an absence or silence—for example, the preterminal H0 generates the **singleton** biterminal h/ϵ which represents a standalone bassline note h against which no melodic note occurs. Thus, the nonterminal A simultaneously generates *both* the bassline $abcdefgh$, and the melody $tuvwxyz\epsilon$. We use the convention of referring to the languages to the left and right of the slash as **language 0** and **language 1**, respectively.

Positional variation in musical phrases

Blues are a good example of an improvisational form in which often melodic phrases are re-used or swapped into different positions within the verses. Melodies from the first four are often re-used or swapped into the second four instead, and vice versa.

We can easily model such phenomena using inversion transduction grammars, since ITGs naturally model the possibility of such swapping of positions of various chunks (a constant phenomenon in natural language translation). Consider the ordinary **straight** rule for FIRST8 from above. If we also add a corresponding **inverted** rule, then we now have two alternatives, where the angle brackets signify that the order for language 1 is inverted:

$$\begin{aligned} \text{FIRST8} &\rightarrow [\text{FIRST4 SECOND4}] \\ \text{FIRST8} &\rightarrow \langle \text{FIRST4 SECOND4} \rangle \end{aligned}$$

This says that for the same language 0 bassline generated by the sequence of two constituents FIRST4 and SECOND4, the language 1 melodic phrase that was played against the bassline of the FIRST4 could also be played against the language 0 bassline of the SECOND4, and vice versa.

As a result, now the melody $tuvwxyz\epsilon$ (generated in language 1 by the nonterminal A, which leads off FIRST8) can not only be played against the bassline $abcdefgh$ (generated in language 0 again by the nonterminal A), but can also possibly be played against whatever bassline is generated in language 0 by the nonterminal D, which leads off SECOND8.

Probabilistic biases and preferences

Just as with monolingual stochastic CFGs, a stochastic transduction grammar is parameterized by associating a probability with each transduction rule. This imposes a probability distribution over the space of possible distributions.

Denoting the model being learned as Φ, the lexical rule $\text{AT} \rightarrow a/t$ for example has the probability $b_{\text{AT}}(a/t) \equiv P(\text{AT} \rightarrow a/t \mid \Phi)$. Likewise, the syntactic rule $\text{FIRST4} \rightarrow [\text{AD AA}]$ has the probability $a_{\text{FIRST4} \rightarrow [\text{AD AA}]} \equiv P(\text{FIRST4} \rightarrow [\text{AD AA}] \mid \Phi)$, and this could be used to bias the nondeterministic choice between the 'quick to four' and basic variants of twelve-bar blues:

$$\begin{aligned} \text{FIRST4} &\rightarrow [\text{AD AA}] \\ \text{FIRST4} &\rightarrow [\text{AA AA}] \end{aligned}$$

Continuous values

In conventional STG models, it is necessary to assign melodic symbols like a and x to 'clean' notes like F♯and C♯ in Western classical scales. This of course does not come close to adequately describing the microtonal pitch values of the characteristic 'bent notes' that are pervasive in blues. Pitches can be bent a little, or a lot, creating significantly different musical effects. Many other non-Western genres, such as flamenco or Indian genres, are even more sensitive to the microtones. A native approach to modeling such continuous values is needed if integrated STG modeling is to be realistically applied to music in general.

In continuous STGs, we replace biterminals that consisted of a pair of discrete symbols, like a/t, with biterminals that instead consist of a pair of continuous values. This means the probability of lexical rules in which preterminals generate biterminals, for example $b_{\text{AT}}(a/t) \equiv P(\text{AT} \rightarrow a/t \mid \Phi)$ which formerly had a scalar value, must be replaced by probability density functions. Using independent Gaussians, with x and y as real values:

$$b_{\text{AT}}(x/y) \equiv \frac{1}{\sqrt{2\pi\sigma^2_{\text{AT},0}}} e^{-\frac{(x-\mu_{\text{AT},0})^2}{2\sigma^2_{\text{AT},0}}} + \frac{1}{\sqrt{2\pi\sigma^2_{\text{AT},1}}} e^{-\frac{(y-\mu_{\text{AT},1})^2}{2\sigma^2_{\text{AT},1}}}$$

With this generalization, x can be used to represent a microtonal melodic pitch, while y can be used to represent an exact bass pitch.

EM training of continuous STGs

Applications

There are numerous applications for automatic simultaneous estimation of both the probabilities for syntactic transduction rules and the pdfs for lexical transduction rules.

In cases where full or partial knowledge of the high-level structure of musical forms is available, as with twelve-bar blues, we can estimate probabilities for the syntactic transduction rules from data. Note that it is not necessary for the the training set to be parsed or annotated.

In cases where no high-level structure is known in advance, as in Wu (2013), estimation of transduction rule probabilities is a basic building block in transduction grammar induction algorithms that automatically analyze and extract the high-level structure.

In either case, simultaneously estimating the pdfs for lexical transduction rules is both important for (a) anchoring estimation of the syntactic transduction rule probabilities from continuous data, and (b) automatically improving the modeling of phenomena like microtonal pitches and volumes.

Algorithm

Estimation of probabilities for both syntactic and lexical transduction rules in continuous SITGs like those in the previous section can be accomplished in $O\left(n^6\right)$ time via an expectation-maximization algorithm for iteratively improving the transduction rule parameters, driven by a maximum likelihood objective. As all ITGs can be normalized into an equivalent 2-normal form (Wu, 1997), we can simplify the description of the algorithm by assuming the SITG to be in 2-normal form, although EM can also readily be implemented for SITGs in arbitrary form. Unlike the inside-outside algorithm for estimating parameters of monolingual SCFGs

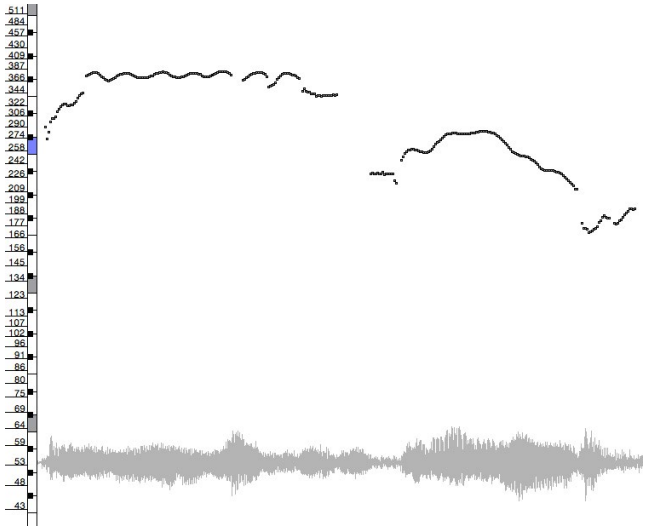

Figure 1: Example contour for a blues vocal melodic phrase that occurs repeatedly in verses at alternate positional variants, showing heavy use of microtonal 'bent' notes.

(Baker, 1979; Lari and Young, 1990), this algorithm handles bilingual SITGs allowing positional variance and pdfs over pairs of continuous-valued musical properties on two musical language streams.

Each iteration of EM first computes generalized inside and outside probabilities, as shown in Figure 2. These quantities are used in reestimating the model parameters Φ employing the procedure derived in Figure 3. We use the shorthand $e_{s..t}$ to denote the language 0 subsequence of continuous values in the span from s to t, or more precisely $e_s, e_{s+1}, \ldots, e_{t-1}$. Likewise, $f_{u..v}$ denotes a subsequence in language 1. We use the notation q_{stuv} to denote the nonterminal label on a bilingual span or bispan $s..t, u..v$.

How blue can you get?

Microtonal blues notes can be 'bent' to a larger or smaller degree; the musical effect is altered by the degree to which they are 'bent'. An accurate model of blues should be capable of learning what degree of microtonal 'bending' goes well with what other parts and in what contexts, so as to reflect biases and preferences in accompaniment and improvisation.

To test this, we trained a continuous SITG using data extracted from the twelve-bar blues 'Give Me One Reason', as recorded by Tracy Chapman. This 'quick to four' blues consisted of seven vocal verses (plus one instrumental verse), over the course of which all the phenomena described in the foregoing sections are exhibited.

The vocal melody and bassline were extracted using the Tony system (Mauch *et al.*, 2015), and then converted into a sequence of frames in language 0 and language 1 streams. Figure 1 shows the melody's heavy use of bent notes.

The transduction grammar encapsulated prior knowledge of the basic twelve-bar blues structures, including the syntactic rules discussed earlier. For the preterminal rules' Gaussian pdfs, on the other hand, the means were randomly initialized rather than trying to predefine microtonal values by hand, and the variances simply initialized to constants. For each nonterminal that directly dominated preterminals, two alternate versions were 'cloned', with separate randomly initialized preterminals allocated to each frame. This strategy provided exploration space to the continuous SITG to self-learn microtonal melodies, basslines, and their interrelationships.

EM training discovered the two main melodic phrases—assigning them to different nonterminals by allocating the 'clones'.

Because the SITG permits positional variation, EM training pays attention to similar melodic phrases, whether they occurred in the first four or the second four bars. For the numerous occurences of melodic phrases similar to Figure 1, we left it to EM training to determine whether a better fitting model could be learned by (a) grouping them all into the same melodic nonterminal category, thereby generalizing over the positional variation, versus (b) associating them with separate nonterminal categories for the first four versus the second four, due to systematic biases. Both possibilities are considered by EM, and they both influence generalization since the probabilities under both alternatives are aggregated when computing the expectations.

In this case EM decided in favor of the latter, despite the fact that the melodic phrases appear essentially the same when aggressively quantized into 'clean' notes on the scale. By instead modeling the continuous microtonal pitches, a correlation that previously would have been overlooked emerges, between the degree of melodic bending and the bassline pitch. For the 'same' melodic phrase, greater bending is associated with the tonic that introduces the first four, compared with the subdominant that introduces the second four. (The preference might be ascribed to greater dissonance in the latter case.)

It could well be that the preferences learned here were idiosyncratic to a particular performer. The EM technique can be used to adapt to mimic styles of particular individuals (as in this case), or alternatively it can be trained on data aggregated from many performers, in order to gain insight on general tendencies in a genre.

After the model parameters have been trained, it becomes possible to use the trained SITG for accompaniment or improvisation. This is accomplished via a transduction algorithm similar to that used in tree-based machine translation Wu and Wong (1998), but again generalized to handle continuous values instead of discrete symbols in analogous fashion to the EM algorithm. Either we can designate the melody (language 0) as the 'output' part to be improvised against a human 'input' bassline (language 1), or we can designate the bassline (language 1) as the 'output' part to accompany a human 'input' melody (language 0). In order to find the most likely improvisation or accompaniment (which we can think of as finding the best translation of the 'input'), we use dynamic programming based parsing to apply the 'input' half of the trained SITG rules to the 'input' language. Once the most likely parse is found, reading the 'output' half of the rules forming that parse yields the best translation.

Conclusion

We have discussed a new strategy for learning complex structural relationships between microtones, and other continuous valued musical features, that simultaneously models contextual influences both within and between different musical languages (players or parts) at many hierarchical or compositional levels of granularity, in improvisational and accompaniment settings. Using continuous stochastic transduction grammars, we bridge the computational modeling gap between (a) fully integrating structural, hierarchical inter-part factors, and (b) finely represented continuous valued signals, overcoming what has until now been one of the major weaknesses in realistically modeling of music based on STGs. Because continuous STGs natively handle continuous valued biterminals, phenomena like microtonal pitch can be modeled without crude quantization to 'clean' notes.

The degree to which melody notes in blues should be 'bent' in the context of decisions made by other players, such as that of the bassline, can be learned via a practical polynomial-time EM algorithm for the continuous instantiation of stochastic inversion transduction grammars—empirically one of the most useful subclasses of stochastic transduction grammars. Syntactic and preterminal probabilities are automatically learned, to model patterns at different contextual granularities between two different continuous valued parts while allowing positional variance.

We are currently exploring whether neural networks, which employ inherently continuous valued representations, could be used to augment continuous STGs. The recursive neural network implementation of STGs described by Wu and Addanki (2015) still only use continuous valued vectors to represent discrete symbols. We believe such neural networks may be directly useful for true continuous valued musical signals, but perhaps in combination with the approach discussed in this paper because the neural models are significantly more lossy and noisy, and difficult to analyze in terms of what musical knowledge they encode.

Acknowledgements

This work is supported in part by the Hong Kong Research Grants Council (RGC) research grants GRF16210714, GRF16214315, GRF620811 and GRF621008; by the Defense Advanced Research Projects Agency (DARPA) under LORELEI contract HR0011-15-C-0114, BOLT contracts HR0011-12-C-0014 and HR0011-12-C-0016, and GALE contracts HR0011-06-C-0022 and HR0011-06-C-0023; and by the European Union under the Horizon 2020 grant agreement 645452 (QT21) and FP7 grant agreement 287658. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of DARPA, the EU, or RGC.

References

Karteek Addanki, Chi-Kiu Lo, Markus Saers, and Dekai Wu. LTG vs. ITG coverage of cross lingual verb frame alternations. In *16th Annual Conference of the European Association for Machine Translation (EAMT-2012)*, pages 295–302, Trento, Italy, May 2012.

Gérard Assayag and Shlomo Dubnov. Using factor oracles for machine improvisation. *Soft Computing*, 8:1432–7643, 2004.

Gérard Assayag, Georges Bloch, Marc Chemillier, Arshia Cont, and Shlomo Dubnov. OMax Brothers: A dynamic topology of agents for improvization learning. In *First ACM Workshop on Audio and Music Computing Multimedia*, pages 125–132, 2006.

James K. Baker. Trainable grammars for speech recognition. In D. H. Klatt andJ. J. Wolf, editor, *Speech Communication Papers for the 97th Meeting of the Acoustic Society of America*, pages 547–550, 1979.

Rens Bod. Stochastic models of melodic analysis: Challenging the gestalt principles. *Journal of New Music Research*, 30(3), 2001.

Alexandre R.J. François, Elaine Chew, and Dennis Thurmond. Mimi - a musical improvisation system that provides visual feedback to the performer. Technical Report 07-889, USC Computer Science Department, Apr 2007.

Alexandre R.J. François, Isaac Schankler, and Elaine Chew. Mimi4x: An interactive audio-visual installation for high-level structural improvisation. In *IEEE International Conference on Multimedia and Expo (ICME 2010)*, pages 1618–1623, 2010.

Jon Gillick, Kevin Tang, and Robert M. Keller. Machine learning of jazz grammars. *Computer Music Journal*, 34(3):56–66, Fall 2010.

Karim Lari and Steve J. Young. The estimation of stochastic context-free grammars using the inside-outside algorithm. *Computer Speech and Language*, 4:35–56, 1990.

Fred Lerdahl and Ray Jackendoff. *A Generative Theory of Tonal Music*. MIT Press, 1983.

Philip M. Lewis and Richard E. Stearns. Syntax-directed transduction. *Journal of the Association for Computing Machinery*, 15(3):465–488, 1968.

Matthias Mauch, Chris Cannam, Rachel Bittner, George Fazekas, Justin Salamon, Jiajie Dai, Juan Bello, and Simon Dixon. Computer-aided melody note transcription using the Tony software: Accuracy and efficiency. In *First International Conference on Technologies for Music Notation and Representation (TENOR 2015)*, 2015.

François Pachet. The Continuator: Musical interaction with style. *Journal of New Music Research*, 32(3):33–341, 2003.

Markus Saers, Joakim Nivre, and Dekai Wu. Learning stochastic bracketing inversion transduction grammars with a cubic time biparsing algorithms. In *11th International Conference on Parsing Technologies (IWPT'09)*, pages 29–32, Paris, Oct 2009.

Mark J. Steedman. The formal description of musical perception. *Music Perception*, 2:52–77, 1984.

Mark J. Steedman. The blues and the abstract truth: Music and mental models. In A. Garnham and J. Oakhill, editors, *Mental Models in Cognitive Science*, pages 305–318. Erlbaum, 1996.

1. Recursive computation of generalized inside probabilities $\beta_{stuv}(i) \equiv P[i \stackrel{*}{\Rightarrow} e_{s..t}/f_{u..v} | q_{stuv} = i, \Phi]$

   (a) Basis

$$\begin{aligned}
\beta_{ttvv}(i) &= 0 && 0 \le t \le T, 0 \le v \le V \\
\beta^0_{stuv}(i) &= \begin{cases} b_i(e_s/f_u) & \text{if } s+1=t, u+1=t, 0 \le s < t \le T, 0 \le u < v \le V \\ 0 & \text{otherwise} \end{cases}
\end{aligned}$$

   (b) Recursion

$$\begin{aligned}
\beta_{stuv}(i) &= \beta^{[]}_{stuv}(i) + \beta^{\langle\rangle}_{stuv}(i) + \beta^0_{stuv}(i) \\
\beta^{[]}_{stuv}(i) &= \sum_{\substack{1 \le j \le N \\ 1 \le k \le N \\ s \le S \le t \\ u \le U \le v \\ (S-s)(t-S)+(U-u)(v-U) \neq 0}} a_{i \to [jk]}\, \beta_{sSuU}(j)\, \beta_{StUv}(k) \\
\beta^{\langle\rangle}_{stuv}(i) &= \sum_{\substack{1 \le j \le N \\ 1 \le k \le N \\ s \le S \le t \\ u \le U \le v \\ (S-s)(t-S)+(U-u)(v-U) \neq 0}} a_{i \to \langle jk \rangle}\, \beta_{sSUv}(j)\, \beta_{StuU}(k)
\end{aligned}$$

2. Recursive computation of generalized outside probabilities $\alpha_{stuv}(i) \equiv P[\mathrm{S} \stackrel{*}{\Rightarrow} e_{0..s} i e_{t..T}/f_{0..u} i f_{v..V}, q_{stuv} = i | \Phi]$

   (a) Basis

$$\begin{aligned}
\alpha_{0,T,0,V}(i) &= \begin{cases} 1 & \text{if } i = \mathrm{S} \\ 0 & \text{otherwise} \end{cases} \\
\alpha_{ttvv}(i) &= 0 && 0 \le t \le T, 0 \le v \le V
\end{aligned}$$

   (b) Recursion

$$\begin{aligned}
\alpha_{stuv}(i) &= \alpha^{[]}_{stuv}(i) + \alpha^{\langle\rangle}_{stuv}(i) \\
\alpha^{[]}_{stuv}(i) &= \sum_{\substack{1 \le j \le N \\ 1 \le k \le N \\ 0 \le S \le s \\ 0 \le U \le u \\ (s-S)(u-U) \neq 0}} \alpha_{StUv}(j)\, a_{j \to [ki]}\, \beta_{SsUu}(k) \;+\; \sum_{\substack{1 \le j \le N \\ 1 \le k \le N \\ t \le S \le T \\ v \le U \le V \\ (S-t)(U-v) \neq 0}} \alpha_{sSuU}(j)\, a_{j \to [ik]}\, \beta_{tSvU}(k) \\
\alpha^{\langle\rangle}_{stuv}(i) &= \sum_{\substack{1 \le j \le N \\ 1 \le k \le N \\ 0 \le S \le s \\ v \le U \le V \\ (s-S)(U-v) \neq 0}} \alpha_{StuU}(j)\, a_{j \to \langle ki \rangle}\, \beta_{SsvU}(k) \;+\; \sum_{\substack{1 \le j \le N \\ 1 \le k \le N \\ t \le S \le T \\ 0 \le U \le u \\ (S-t)(u-U) \neq 0}} \alpha_{sSUv}(j)\, a_{j \to \langle ik \rangle}\, \beta_{tSUu}(k)
\end{aligned}$$

Figure 2: Dynamic programming for computing generalized inside and outside probabilities for continuous SITGs.

Reid Swanson, Elaine Chew, and Andrew S. Gordon. Supporting musical creativity with unsupervised syntactic parsing. In *AAAI Spring Symposium on Creative Intelligent Systems*, 2007.

Dekai Wu and Karteek Addanki. Learning to rap battle with bilingual recursive neural networks. In *24th International Joint Conference on Artificial Intelligence (IJCAI 2015)*, pages 2524–2530, Buenos Aires, Jul 2015.

Dekai Wu and Hongsing Wong. Machine translation with a stochastic grammatical channel. In *36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics (COLING-ACL'98)*, Montreal, Aug 1998.

Dekai Wu, Karteek Addanki, Markus Saers, and Meriem Beloucif. Learning to freestyle: Hip hop challenge-response induction via transduction rule segmentation. In *2013 Conference on Empirical Methods in Natural Language Processing (EMNLP 2013)*, pages 102–112, Seattle, Oct 2013.

Dekai Wu. Stochastic Inversion Transduction Grammars and bilingual parsing of parallel corpora. *Computational Linguistics*, 23(3):377–404, Sep 1997.

Dekai Wu. Alignment. In Nitin Indurkhya and Fred J. Damerau, editors, *Handbook of Natural Language Processing*, pages 367–408. Chapman and Hall / CRC, second edition, 2010.

Dekai Wu. Simultaneous unsupervised learning of flamenco metrical structure, hypermetrical structure, and multipart

1. Probability of using each nonterminal in a derivation of the observed training pair:

$$P[i \text{ used} \mid \mathrm{S} \stackrel{*}{\Rightarrow} \mathbf{e}/\mathbf{f}, \Phi] = \frac{\sum_{s=0}^{T}\sum_{t=s}^{T}\sum_{u=0}^{V}\sum_{v=u}^{V} P[\mathrm{S} \stackrel{*}{\Rightarrow} \mathbf{e}/\mathbf{f} | q_{stuv} = i, \Phi]}{P[\mathrm{S} \stackrel{*}{\Rightarrow} \mathbf{e}/\mathbf{f}|\Phi]} = \frac{\sum_{s=0}^{T}\sum_{t=s}^{T}\sum_{u=0}^{V}\sum_{v=u}^{V} \alpha_{stuv}(i)\beta_{stuv}(i)}{P[\mathrm{S} \stackrel{*}{\Rightarrow} \mathbf{e}/\mathbf{f}|\Phi]}$$

2. Probability of using each straight or inverted transduction rule in a derivation of the observed training pair:

$$P[i \to [jk] \text{ used} \mid \mathrm{S} \stackrel{*}{\Rightarrow} \mathbf{e}/\mathbf{f}, \Phi] = \sum_{s=0}^{T}\sum_{t=s}^{T}\sum_{u=0}^{V}\sum_{v=u}^{V} P[i \Rightarrow [jk] \stackrel{*}{\Rightarrow} e_{s..t}/f_{u..v} \mid \mathrm{S} \stackrel{*}{\Rightarrow} \mathbf{e}/\mathbf{f}, \Phi] = \frac{\sum_{s=0}^{T}\sum_{t=s}^{T}\sum_{u=0}^{V}\sum_{v=u}^{V}\sum_{S=s}^{t}\sum_{U=u}^{v} a_{i\to[jk]}\alpha_{stuv}(i)\beta_{sSuU}(j)\beta_{StUv}(k)}{P[\mathrm{S} \stackrel{*}{\Rightarrow} \mathbf{e}/\mathbf{f}|\Phi]}$$

$$P[i \to \langle jk \rangle \text{ used} \mid \mathrm{S} \stackrel{*}{\Rightarrow} \mathbf{e}/\mathbf{f}, \Phi] = \sum_{s=0}^{T}\sum_{t=s}^{T}\sum_{u=0}^{V}\sum_{v=u}^{V} P[i \Rightarrow \langle jk \rangle \stackrel{*}{\Rightarrow} e_{s..t}/f_{u..v} \mid \mathrm{S} \stackrel{*}{\Rightarrow} \mathbf{e}/\mathbf{f}, \Phi] = \frac{\sum_{s=0}^{T}\sum_{t=s}^{T}\sum_{u=0}^{V}\sum_{v=u}^{V}\sum_{S=s}^{t}\sum_{U=u}^{v} a_{i\to\langle jk\rangle}\alpha_{stuv}(i)\beta_{sSUv}(j)\beta_{StuU}(k)}{P[\mathrm{S} \stackrel{*}{\Rightarrow} \mathbf{e}/\mathbf{f}|\Phi]}$$

3. Transduction rule probabilities (by definition):

$$a_{i\to[jk]} \equiv P[i \to [jk] \text{ used} \mid i \text{ used}, \mathrm{S} \stackrel{*}{\Rightarrow} \mathbf{e}/\mathbf{f}, \Phi]$$
$$a_{i\to\langle jk\rangle} \equiv P[i \to \langle jk \rangle \text{ used} \mid i \text{ used}, \mathrm{S} \stackrel{*}{\Rightarrow} \mathbf{e}/\mathbf{f}, \Phi]$$

4. Re-estimation procedure for transduction rule probabilities $\hat{a}$ (by substitution):

$$\hat{a}_{i\to[jk]} = \frac{\sum_{s=0}^{T}\sum_{t=s}^{T}\sum_{u=0}^{V}\sum_{v=u}^{V}\sum_{S=s}^{t}\sum_{U=u}^{v} a_{i\to[jk]}\alpha_{stuv}(i)\beta_{sSuU}(j)\beta_{StUv}(k)}{\sum_{s=0}^{T}\sum_{t=s}^{T}\sum_{u=0}^{V}\sum_{v=u}^{V} \alpha_{stuv}(i)\beta_{stuv}(i)}$$

$$\hat{a}_{i\to\langle jk\rangle} = \frac{\sum_{s=0}^{T}\sum_{t=s}^{T}\sum_{u=0}^{V}\sum_{v=u}^{V}\sum_{S=s}^{t}\sum_{U=u}^{v} a_{i\to\langle jk\rangle}\alpha_{stuv}(i)\beta_{sSUv}(j)\beta_{StuU}(k)}{\sum_{s=0}^{T}\sum_{t=s}^{T}\sum_{u=0}^{V}\sum_{v=u}^{V} \alpha_{stuv}(i)\beta_{stuv}(i)}$$

5. Re-estimation procedure for preterminal rules' Gaussian means $\hat{\mu}$ and variances $\hat{\sigma}$:

$$\hat{\mu}_{i,0} = \frac{\sum_{s=0}^{T-1}\sum_{u=0}^{V-1} e_s \alpha_{s,s+1,u,u+1}(i)\beta_{s,s+1,u,u+1}(i)}{\sum_{s=0}^{T}\sum_{t=s}^{T}\sum_{u=0}^{V}\sum_{v=u}^{V} \alpha_{stuv}(i)\beta_{stuv}(i)} \qquad \hat{\sigma}^2_{i,0} = \frac{\sum_{s=0}^{T-1}\sum_{u=0}^{V-1} (e_s - \mu_{i,0})^2 \alpha_{s,s+1,u,u+1}(i)\beta_{s,s+1,u,u+1}(i)}{\sum_{s=0}^{T}\sum_{t=s}^{T}\sum_{u=0}^{V}\sum_{v=u}^{V} \alpha_{stuv}(i)\beta_{stuv}(i)}$$

$$\hat{\mu}_{i,1} = \frac{\sum_{s=0}^{T-1}\sum_{u=0}^{V-1} f_u \alpha_{s,s+1,u,u+1}(i)\beta_{s,s+1,u,u+1}(i)}{\sum_{s=0}^{T}\sum_{t=s}^{T}\sum_{u=0}^{V}\sum_{v=u}^{V} \alpha_{stuv}(i)\beta_{stuv}(i)} \qquad \hat{\sigma}^2_{i,1} = \frac{\sum_{s=0}^{T-1}\sum_{u=0}^{V-1} (f_u - \mu_{1,1})^2 \alpha_{s,s+1,u,u+1}(i)\beta_{s,s+1,u,u+1}(i)}{\sum_{s=0}^{T}\sum_{t=s}^{T}\sum_{u=0}^{V}\sum_{v=u}^{V} \alpha_{stuv}(i)\beta_{stuv}(i)}$$

Figure 3: Derivation of EM reestimation of model parameters Φ for continuous SITGs, using inside and outside probabilities.

structural relations. In *14th International Society for Music Information Retrieval Conference (ISMIR 2013)*, Curibata, Brazil, Nov 2013.

Richard Zens and Hermann Ney. A comparative study on reordering constraints in statistical machine translation. In *41st Annual Meeting of the Association for Computational Linguistics (ACL-2003)*, pages 192–202, Sapporo, Aug 2003.