GENERALIZING TRANSDUCTION GRAMMARS TO MODEL CONTINUOUS VALUED MUSICAL EVENTS

Dekai Wu
Hong Kong University of Science & Technology
dekai@cs.ust.hk

ABSTRACT

We describe a generalization of stochastic transduction grammars to be able to model *continuous* values, the first models to natively handle continuous-valued musical events such as microtones while still gaining the advantages of STGs for describing complex structural, hierarchically compositional inter-part relationships. Music transduction modeling based on linguistic or grammatical models have commonly approximated continuous valued features like pitch by quantizing them into discrete symbols, which represent 'clean' notes on a scale. The sacrifice is worthwhile for modeling the learning of musical improvisation and accompaniment where musical sequences interact hierarchically at many overlapping levels of granularity; previous work has shown in flamenco and hip hop how discrete STGs allow each part to influence decisions made by other parts while also satisfying contextual preferences across multiple dimensions. We extend the modeling machinery toward the many musical genres where contextual relationships between *continuous* values influence improvisational and accompaniment decisions. Illustrating using the 'bent notes' prevalent in blues, we show how continuous STGs can be generalized from conventional discrete STGs which have until now only been able to handle symbolic events, thereby allowing musical signals to remain finely represented as continuous values without crude quantization into discrete symbols, while still retaining the ability to model probabilistic structural relations between multiple musical languages. We exemplify this new approach in learning blues notes biases via a new polynomial time algorithm for expectation-maximization training of continuous SITGs (stochastic inversion transduction grammars), a specific subclass of STGs that has already proven useful in numerous applications in both music and language.

1. STOCHASTIC TRANSDUCTION GRAMMARS

To introduce continuous STGs (see Wu [4, 5] for details of conventional STGs) we will generalize a step at a time from context-free grammars, illustrating with the well-known *twelve-bar blues* form in which verses consist of three lines: a first four bars, a second four, and a third four called a turnaround. A stochastic context-free grammar could use these syntactic rules to describe twelve-bar blues with the first four in both basic [AA AA] and 'quick to four' [AD AA] variants, biased by rule probabilities:

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

Transduction grammars can generalize this bilingually to express the relationship between, for example, a bassline language and a vocal melody language. Rather than monolingual grammars' preterminal symbols, transduction grammars instead have *biterminals*, bilingual preterminals representing a relation between two lexical atoms from two *different* languages. Let us further decompose the nonterminal A, which represents a single bar in the tonic, into a finer grained series of frames—for simplicity eighth note durations (but could use finer granularities):

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

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 $abcdefgh$ in language 0, and the melody $tuvwxyz\epsilon$ in language 1.

Positional variation of musical phrases is common in improvisational forms like blues, where melodic phrases are often 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. Such swapping of positions of various chunks (a constant phenomenon in natural language translation) are naturally modeled using **inversion transduction grammars** or **ITGs**. Earlier we saw the ordinary *straight* rule for FIRST8; we can now add an alternative corresponding *inverted* rule where the angle brackets invert the order for language 1:

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

This says that for the same language 0 bassline generated by the sequence of 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 be played not only against the bassline $abcdefgh$ (generated in language 0 again by the nonterminal A), but possibly also against whatever bassline is generated in language 0 by the nonterminal D, which leads off SECOND8.

2. CONTINUOUS STG MODELS

Conventional STG models are limited to defining melodic symbols like a and x as discrete notes in Western classical scales, which is far from adequate for describing the microtonal pitch values of the 'bent notes' characteristic of blues (and many other non-Western genres). Bending pitches a little, or a lot, creates very different musical effects.

Continuous STGs introduce a more realistic, native approach by replacing symbolic biterminals (pairs of discrete symbols, like a/t) with continuous biterminals that instead consist of a pair of continuous values x and y. 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, is replaced by probability density functions. Using independent Gaussians, we instead define $b_{\text{AT}}(x/y)$ as:

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

For example, x can be used to represent a microtonal melodic pitch, while y can be used to represent an exact bass pitch, in Tracy Chapman's twelve-bar blues 'Give Me One Reason', a 'quick to four' blues consisting of seven vocal verses (plus one instrumental verse). Figure 1 shows the melody's heavy use of bent notes as extracted with the Tony system [3]; we converted the vocal melody and bassline into a sequence of frames in language 0 and language 1 streams.

EM for continuous STGs estimates probabilities for both syntactic and lexical transduction rules in our new algorithm in $O\left(n^6\right)$ time. As all ITGs can be normalized into an equivalent 2-normal form [4], we can simplify the dynamic programming by assuming the SITG to be in 2-normal form, though EM can also readily be implemented for SITGs in arbitrary form. Unlike the inside-outside algorithm for estimating parameters of monolingual SCFGs [1, 2], our algorithm handles bilingual SITGs allowing positional variance and pdfs over pairs of continuous-valued musical properties on two musical language streams.

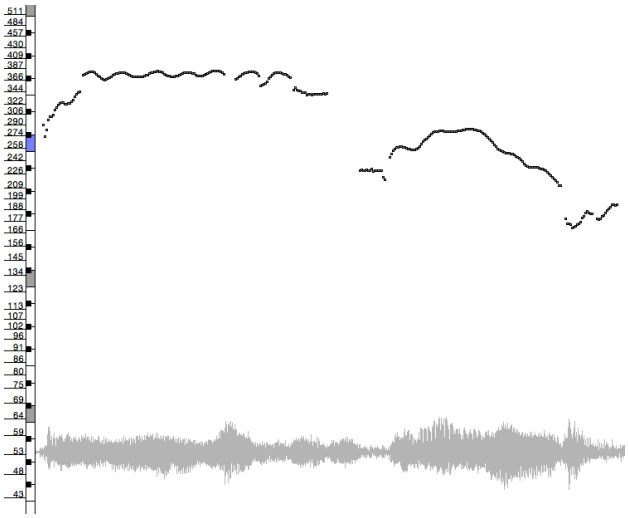

Figure 1. Example contour for a blues vocal melodic phrase that is repeated in verses at alternate positional variants, with heavy use of microtonal 'bent' notes.

This enables a broad range of new applications. In cases where full or partial knowledge of the high-level structure of musical forms is available, as with twelve-bar blues, we estimate probabilities for the syntactic transduction rules from unannotated data—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 (possibly arising from greater dissonance in the latter case). In cases where no high-level structure is unknown, as in [6], this provides the continuous valued handling for the basic EM 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.

The trained continuous SITG can be used for accompaniment or improvisation, via a dynamic programming based transduction algorithm similar to that in tree-based machine translation [7] but again generalized to handle continuous values instead of discrete symbols. We can designate the melody (language 0) as the 'output' part to be improvised against a human 'input' bassline (language 1), or vice versa. We are presently investigating a number of extensions to this approach.

3. ACKNOWLEDGMENTS

This work is supported in part by Hong Kong RGC research grants GRF16210714, GRF16214315, GRF620811 and GRF621008; by the EU under the Horizon 2020 grant agreement 645452 (QT21) and FP7 agreement 287658; and by 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. 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.

4. REFERENCES

[1] 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.

[2] 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.

[3] 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.

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

[5] 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.

[6] Dekai Wu. Simultaneous unsupervised learning of flamenco metrical structure, hypermetrical structure, and multipart structural relations. In *14th International Society for Music Information Retrieval Conference (ISMIR 2013)*, Curibata, Brazil, Nov 2013.

[7] 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.