Harmonics co-occurrences bootstrap pitch and tonality perception in music: Evidence from a statistical unsupervised learning model

Kat Agres (kathleen.agres@qmul.ac.uk)
Queen Mary, University of London
Department of Electronic Engineering and Computer Science
London E1 4NS, UK

Carlos Cancino, Maarten Grachten, Stefan Lattner (firstname.lastname@ofai.at)
Austrian Research Institute for Artificial Intelligence (OFAI)
Freyung 6/6, A-1010 Vienna, Austria

Abstract

The ability to extract meaningful relationships from sequences is crucial to many aspects of perception and cognition, such as speech and music. This paper explores how leading computational techniques may be used to model how humans learn abstract musical relationships, namely, tonality and octave equivalence. Rather than hard-coding musical rules, this model uses an unsupervised learning approach to glean tonal relationships from a musical corpus. We develop and test a novel input representation technique, using a perceptually-inspired harmonics-based representation, to bootstrap the model's learning of tonal structure. The results are compared with behavioral data from listeners' performance on a standard music perception task: the model effectively encodes tonal relationships from musical data, simulating expert performance on the listening task. Lastly, the results are contrasted with previous findings from a computational model that uses a more simple symbolic input representation of pitch.

Keywords: Music perception; tonality; unsupervised learning; Restricted Boltzman Machines

Introduction

Learning the rules and structure of sequential information is of fundamental importance to human perception and cognition, yet the process by which this occurs is still debated and widely investigated across domains. In language, for example, linguistic nativists posit that innate, domain-specific mechanisms are responsible for grammar learning (e.g., Berwick, Pietroski, Yankama, & Chomsky, 2011; Pinker, 1994), while others argue that more general, statistical learning mechanisms underlie the induction of grammatical rules (Chater & Manning, 2006; Saffran & Wilson, 2003; Gomez & Gerken, 1999). This debate has spread to other domains, such as the perception of tonal music, which, like language, is highly structured, and is governed by a set of grammatical rules that can be described in music-theoretic terms (Lerdahl & Jackendoff, 1983). Indeed, listeners' ability to implicitly extract statistical regularities and knowledge of tonal relationships has received much attention in recent years (Pearce, 2005; Saffran, Johnson, Aslin, & Newport, 1999).

In an effort to model mechanisms for learning statistical structure, unsupervised learning methods and Restricted Boltzman Machines (RBMs) have garnered enthusiastic support for examining questions of learning, feature representations, and the probabilistic structure of (big) data. Once an RBM has learned the properties of the given data, its latent (learned) feature spaces may be explored, and used for clustering or categorization. These abstracted representations may also model human perception (Hinton, 2007; Bartlett, 2001; Grachten & Krebs, 2014).

In music, unsupervised learning techniques have been used effectively to learn feature representations for the harmonic relationships between keys (Leman, 1995) and tonal pitch relationships within a key (Cancino, Lattner, & Grachten, 2014). The model proposed by Cancino et al. (2014) successfully replicates certain aspects of pitch perception, but fails to replicate others, such as the perception of octave similarity (the perceptual similarity of tones one octave apart) displayed by musicians. This is likely due to the symbolic pitch input representation used, which fails to capture harmonic relationship between tones. The current work uses a novel harmonics-based input representation inspired by human pitch perception, with the hypothesis that the additional information provided from lower resolved harmonics will bootstrap both the perception of tonal relationships and octave similarity. The present research investigates this topic through the use of unsupervised statistical learning, and tests the extent to which these methods are capable of modelling the perception of tonality, through the use of this more rich input representation.

Pitch Perception in Listeners

Arguably, the statistical properties of music (such as pitch occurrences and transitional probabilities between tones or chords) enable its structure to be *learned* from exposure. For example, the transitional probabilities between musical events, and the frequency of occurrence of pitches in tonal music, contribute to listeners perception of the hierarchical relationship of pitches within a key (Smith & Schmuckler, 2004). This is known as the "tonal hierarchy", a phrase that highlights the relative stability or importance of certain pitches in a musical key. In other words, due to the predominance of some notes over others within a tonality (such as the tonic and fifth scale degree), certain notes are perceived as belonging more or less to the key than others, and are consequently perceived as having different functional roles in the tonality. In the case of C Major, for example, the notes C and G (the tonic and fifth scale degree) have greater stability than the leading tone (B, the seventh in the scale), or chromatic pitches not in the key (e.g., F sharp).

Discovery of the tonal hierarchy was the result of seminal

studies by Krumhansl and colleagues (Krumhansl, 1990) using a "probe tone paradigm". In this task, listeners hear a musical context that clearly establishes a key (such as an ascending or descending scale), but is left incomplete (e.g., without the final note of the scale). After this context, a subsequent "probe tone" is played, and listeners rate how well the tone completes the prior context, usually on a scale from 1 ("very bad") to 7 ("very good") (Krumhansl & Shepard, 1979). The results of probe tone tasks have repeatedly shown that different pitches have different functions in the key. There is historical precedence for using human probe tone results as a measure of model performance in music, and our computational model follows this tradition.

In addition to the statistical properties of music, the characteristics of the acoustic signal also impact pitch and tonality perception (McDermott & Oxenham, 2008; De Cheveigne, 2005). Pitch, the psychological perception of frequency, is perceived in logarithmic relation to frequency. Whereas octaves on the linear frequency spectrum become farther apart the higher the absolute pitch, octaves are equally-spaced on the mel scale (such that doubling a frequency creates the perception of a pitch one octave higher). There is some evidence that the perceptual similarity of pitches an octave apart is universal and innate (Demany & Armand, 1984), and nearly all cultures base their musical scale on a one-octave range.

From a developmental perspective, given that most voices and instruments produce tones in which the fundamental pitch (F0) is much stronger than the partials, listeners may gradually build up pitch and tonal perception from weak individual harmonics. Empirical studies show that adults tend to be more sensitive to tonal relationships and less influenced by pitch proximity than children (Cuddy & Badertscher, 1987). If greater perception of individual harmonics is gained over the developmental trajectory, models using F0 as input may better simulate children and novice listeners, while models using harmonics information may reflect more experienced listeners.

Because both low-level acoustic information and implicit statistical learning mechanisms contribute to tonal perception in listeners, the present research sought to model how the hierarchical perception of tonality may be learned through exposure to music, utilizing an input representation inspired by the perception of pitch.

Computational approaches

Hard-coded, rule-based models can describe various cognitive phenomena with notable accuracy, possibly capturing some of the innate structure that constrains bottom-up, domain-general cognitive processing. Nevertheless, perception reflects, to a substantial degree, what is learned based on experience. Accordingly, an emphasis has recently been placed on investigating *how* features of data are learned from exposure. The development of such systems allows researchers to model perception without requiring user input or the pre-specification of rules. To this end, statistical and probabilistic approaches have elucidated aspects of music perception, such as tonal relationships and the perception of musical-phrase boundaries. Although this data-driven approach has been fairly successful, many statistical approaches lack robustness (e.g., they do not capture an entire conditional probability distribution), resistance to noise, and flexibility regarding different prior contexts. To circumvent these issues, an unsupervised RBM model is presently used to learn the probabilistic structure of tonal music through repeated exposure to a musical corpus.

An advantage of RBMs over the Self-Organizing Maps (SOMs) used in prior computational modeling approaches to the perception of tonality (Leman, 1995; Tillmann, Bharucha, & Bigand, 2000) is that the learned representation space in SOMs is typically 2- or 3-dimensional, whereas RBMs can learn spaces of arbitrary dimensionality. Low-dimensional space is convenient for visualization, but there are few biologically-motivated reasons for enforcing learned representations to be low-dimensional. Although RBMs are not claimed to be plausible models of neural structures, stacked RBMs have been shown capable of learning biologically realistic receptive fields in vision (Lee, Ekanadham, & Ng, 2008).

Perception-Based Input Representation Applications of neural networks to music often start from symbolic representations of music, midi notes, or piano roll notation (Cancino et al., 2014; Grachten & Krebs, 2014; Boulanger-Lewandowski, Bengio, & Vincent, 2012). This usually implies that pitch (octave-specific note name, e.g., 'G4') or pitch chroma (octave-invariant note name, e.g., 'G') are used, but this approach means losing potentially useful information from harmonics that can ain in the extraction of tonal relationships. For example, human listeners perceive co-occurring harmonics for pitches that are an octave or a fifth apart; this consonance may help listeners develop abilities such as octave similarity perception and relative pitch. Therefore, we developed an input representation that could enable the RBM to use harmonics to bootstrap tonal learning.

A harmonics representation provides the model with richer input than using only note-names or fundamental pitches. Other computational approaches have represented even lower-level information; for example, autocorrelation temporal models (Licklider, 1951; van Noorden, 1982; Meddis & Hewitt, 1991; Cariani, 2001) have shown that neural interspike interval representations and their subharmonic representations may potentially underlie the perception of pitch as well as some basic aspects of tonality. Complementary to this tradition, we endeavored to test whether resolved harmonics within the range of the piano (which covers the range of musical tonality) were sufficient to simulate listeners' performance on a music perception task addressing the tonal function of pitches within a key. While innate properties of the auditory system (e.g., neural spiking activity) may subserve representations of tonality, tonal perception is likely mediated by experience. We were therefore interested in whether different input representations (harmonics vs F0s) would better simulate listeners with varying degrees of musical expertise.

When examining the perception of tonal structure, our harmonics representation has the advantage over audio-based representations (such as acoustic spectra computed from

tones) that it allows us to focus solely on the effect of coinciding resolved harmonics between tones. When working with acoustic spectra, this effect is blurred by phenomena like inharmonicity, and tone quality (timbre). It is beyond the scope of this article to account for the effect of these phenomena on the perception of tonal structure. Thus, the following approach employs an abstract representation based on human pitch perception, with the hypothesis that co-occurring harmonics may scaffold the development of relative pitch and octave affinity found in musically-trained listeners.

Method

Restricted Boltzman Machine model

The present research implemented a Restricted Boltzmann Machine, a generative stochastic neural network (Hinton, 2002). This model consists of a layer of visible units $\mathbf{v} \in \mathbb{R}^n$, which represent the observed data, and a layer of binary hidden units $\mathbf{h} \in \{0,1\}^l$. Both layers form a bipartite graph, i.e. there are no connections between units from each layer. The joint probability distribution of $\mathbf{v}$ and $\mathbf{h}$ described by the RBM is given by

$$p(\mathbf{v},\mathbf{h}) = \frac{1}{Z}\exp\left(-E(\mathbf{v},\mathbf{h} \mid \theta)\right),$$

where Z is a normalization term, and $E(\cdot)$ is an energy function, usually a quadratic function of the visible and hidden units. This energy function is proportional to the log-likelihood function of the model parameters θ given the visible and hidden units, and its name was inspired by the Ising model from statistical thermodynamics. For the standard Bernoulli-Bernoulli RBM[1], the energy function is

$$E(\mathbf{v},\mathbf{h} \mid \mathbf{W},\mathbf{a},\mathbf{b}) = -\mathbf{v}^T\mathbf{a} - \mathbf{h}^T\mathbf{b} - \mathbf{v}^T\mathbf{W}\mathbf{b},$$

where $\theta = \{\mathbf{W},\mathbf{a},\mathbf{b}\}$, with $\mathbf{W} \in \mathbb{R}^{n\times l}$ a weight matrix, $\mathbf{a} \in \mathbb{R}^n$ a bias vector for the visible units, and $\mathbf{b} \in \mathbb{R}^l$ a bias vector for the hidden units.

The free energy (FE), denoted by $\mathcal{F}(\mathbf{v})$, is a measure of the expectancy of an input (visible) configuration, since it is proportional to the expected value of the conditional probability of the visible units given all possible configurations of the hidden units, i.e.

$$\mathcal{F}(\mathbf{v}) \propto -\log\left(\mathbb{E}\{p(\mathbf{v} \mid \mathbf{h})\}\right).$$

Model training

The model parameters θ are optimized to maximize the expected log-likelihood of the observed data. In the machine learning literature, this optimization process for neural networks is known as *training* (Bishop, 1995). The standard method for training RBMs is known as Contrastive Divergence, proposed in (Hinton, 2002). In this gradient-descent-like algorithm, the gradient of the log-likelihood of the observed data is approximated using Gibbs sampling, a Markov Chain Monte Carlo technique that is well suited for energy based models such as RBMs (Hinton, 2002).

For this paper, we train a model with 100 hidden units for 200 epochs, using a single Gibbs sampling step and a mini-batch size of 100. Different model parameters were explored, such as the size of hidden layer and the amount of training epochs. All hyperparameters (learning rate, momentum, number of steps of Gibbs sampling) were selected according to the guidelines proposed by Hinton in (Hinton, 2012).

Harmonics input representation

A distributed binary input vector was computed for every pitch of the piano keyboard, from A0 to C8, tuned in equal temperament. For each pitch the first four harmonics were represented, comprising the fundamental frequency and three successive harmonics for each pitch. The harmonic series was computed by multiplying the pitch's F0 by integer values (2 for the second harmonic, 3 for the third harmonic, etc). The four harmonics encoded represent the F0, an octave interval above the F0 (second harmonic), a fifth above the second harmonic (third harmonic), and two octaves above the F0 (fourth harmonic). The harmonics for all 88 piano tones formed a total of 112 frequency bins, which served as the 112 visible input nodes for the model. The binary input vector (visible units) for each pitch encoded that pitch's harmonics, i.e., there were four "on" nodes in each input vector.

Training corpus

Our training corpus consisted of the entire set of 48 fugues from J.S. Bachs Well-Tempered Clavier, regarded as one of the most seminal works of classical music. Previous computational modeling shows that the representations derived from this corpus reflect the "Circle of Fifths"; in other words, the statistics of this corpus yield meaningful relationships between the musical keys (Cancino et al., 2014). Because the fugues span every key and therefore have different distributions of pitch occurrences, they were all transposed to the key of C. Transposing or otherwise accounting for key (e.g., by representing scale degree and pitch interval) is common practice for training computational models on tonal corpora that span different keys. Without transposition, the statistics defining tonal relationships from different keys will provide conflicting information to a model that uses absolute pitch representation. Each fugue was decomposed into its constituent voices (two to five per fugue), where "voices" refers to the number of parts in the musical score. Voices in the bass register were moved to the C3 to C6 range to enable their tonal information to be used and integrated by the model. This yielded a total of 166 voices used for training, and each voice was considered as a single monophonic melody in the corpus.

The set of voices were converted from their original MIDI format into the harmonics representation described above (every pitch was replaced by its binary harmonics vector). The RBM was then trained on n-grams of these harmonics vectors, where an n-gram is defined as a successive set of n tones in the corpus. N-grams were each eight notes long, and were

[1] For more details on the derivation of energy functions for several RBM architectures see (Cancino, Lattner, & Grachten, 2015)

created by means of a sliding window (e.g., for a particular melody, notes 1-8 formed the first n-gram, followed by notes 2-9 for the second n-gram, etc). This n-gram length was chosen to allow for the presentation of a seven tone stimulus plus a single probe tone to the model, as is necessary for comparison with human ratings on a probe tone test (see the next section on Model Evaluation). Moreover, Cancino et al. (2014) found that a minimum of eight notes in an n-gram was necessary for optimal categorization of the n-gram in terms of tonal key. The 8-grams were presented in randomized order to provide more robust training for the model. Note that the RBM computes the probabilities of the elements in each input vector (the set of visible nodes that encode the input, in our case, the set of eight pitches), *not* the probability of a sequence of n-grams. As such, there is no temporal aspect with regard to the order of training instances themselves; rather, each n-gram is treated as another time-invariant training instance. The RBM extracts meaningful relationships between the pitches *within*, and not between, each training n-gram.

Model evaluation

After training the RBM, the model's internalization of the tonal pitch hierarchy was tested. To this end, we implemented a Krumhansl-style probe tone test: The model was given either an ascending scale (the octave from C3 to C4) or a descending (from C6 to C5), without the final C to complete the octave. This musical context was immediately followed by a probe tone which was selected from the chromatic pitches between C4 and C5 (see Figure 1).

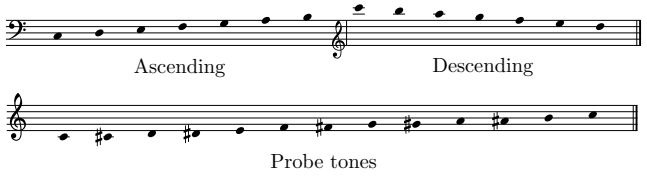

Figure 1: Ascending and descending C major scale context, and the set of possible chromatic probe tones.

To provide these stimuli to the model for testing, we constructed n-grams of length 8, each of which contained the seven pitches from the ascending or descending scale, followed by a probe tone. This yielded a test set of 26 stimuli. The Free Energy (FE) was calculated for each of these probe test stimuli, and then normalized and scaled for comparison with human ratings.

The model's performance was compared with that of listeners for both ascending and descending scale stimuli, as reported in Krumhansl and Shepard (1979). This classic study was chosen because 1) the probe tone context featured scales rather than chords, 2) tones containing harmonics were used (as opposed to pure tones, or Shepard tones as in Krumhansl and Kessler (1982)), and 3) listeners with different levels of training were tested. This last point enabled us to test the hypothesis that this richer input representation will allow the model to better simulate listeners with greater musical experience. The model results were therefore compared to highly trained musicians (experts) and musically-untrained listeners (novices). We refer the reader to this paper for further details regarding the study.

Results and Discussion

The performance of the model, as assessed by the FEs of the probe test stimuli, was compared with average probe tone ratings by expert and novice listeners (Krumhansl & Shepard, 1979). We were interested in comparing the model with the *pattern* (or profile) of human responses across probe tones, but the original variance data of listeners' responses is no longer available, which precludes statistical significance testing. Therefore, to compare the patterns of results, we calculated the Kullback-Leibler (KL) divergence (Kullback & Leibler, 1951) between the two sets of data, which measures the distance between two discrete distributions, $\mathbf{p}^{(1)}$ and $\mathbf{p}^{(2)}$. KL divergence was then used as the kernel for a distance-based Similarity measure (Shepard, 1987) that is used to quantify the similarity between the two distributions:

$$\text{Similarity}\left(\mathbf{p}^{(1)} \mid \mathbf{p}^{(2)}\right) = \exp\left(-D_{\text{KL}}\left(\mathbf{p}^{(1)} \mid \mathbf{p}^{(2)}\right)\right).$$

This similarity measure has the property of being exactly one if both distributions are identical, and tends asymptotically to zero if the KL divergence between both distributions goes to infinity. Similarity (e^{-KL}) values and Pearson correlations between model and human ratings are provided in Table 1 for an RBM tested on probe stimuli with ascending scale and descending scale contexts.

Table 1: Comparison of expert and novice listeners' probe tone ratings (for ascending or descending stimuli) with an RBM model tested on ascending or descending scale contexts. The highest Similarity values are in bold for both of the model test conditions.

	Asc model context		Desc model context	
Expertise	*R*	*Similarity*	*R*	*Similarity*
Expert (Asc)	0.82	**0.88**	0.72	0.57
Expert (Desc)	0.83	0.84	0.83	**0.88**
Novice (Asc)	0.59	0.75	0.00	0.42
Novice (Desc)	0.54	0.52	0.75	0.54

Given the model's results for ascending test stimuli, the KL divergence is lowest (i.e., the distributions were least different), and the Similarity is greatest, for Expert listeners' ratings of ascending probe stimuli. In other words, the model reflects expert listeners' behavioral results for this set of stimuli. The RBM results are most highly correlated with expert listeners for both ascending and descending stimuli.

These findings are mirrored by the descending stimuli results. For these test stimuli, the KL divergence is lowest and the Similarity is highest for Expert listeners who rated descending probe stimuli. The RBM results are most highly correlated with descending ratings from Expert listeners. Once again, the model best reflects expert listeners' results when

the two are compared on the same set of stimuli, and further demonstrates the stimulus-specific response of the RBM.

The comparisons between RBM Free Energy results and expert listeners' ratings are plotted in Figure 2 for visualization. These graphs illustrate that the RBM was able to model the hierarchical tonal relationships exhibited by listeners: The model learned the privileged role of diatonic pitches in the C major scale, and exhibits a degree of octave similarity.

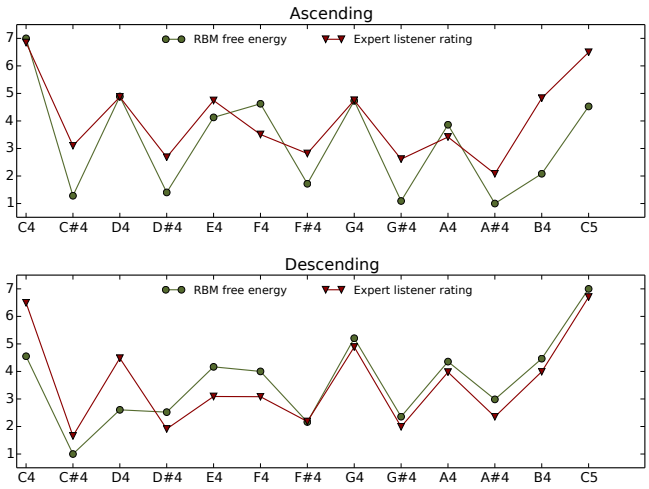

Figure 2: RBM Free Energy results compared with average probe tone ratings by expert listeners for ascending and descending scale contexts.

These results were also compared to a version of the RBM model that was instead trained on MIDI pitch with no harmonics, as reported in Cancino et al. (2014), which was configured to have the same parameters and hyperparameters as the model discussed above. This pitch-only model, trained to the same number of epochs, yielded worse performance when compared with listeners. The highest Similarity value for ascending contexts was 0.58 (for non-expert listeners rating descending stimuli). The highest Similarity value for descending contexts was 0.76 (with novice listeners rating descending stimuli). The greater similarity to untrained listeners rating descending contexts may reflect the prevalence of C4 over C3 in the corpus. Compared to an RBM model using only a local binary pitch representation, the harmonics representation yields better overall results. Also, whereas the harmonics representation best models expert listeners, the pitch-only representation reflects less experienced listeners.

Conclusion

In this paper, we use unsupervised learning techniques to train a computational model on pitch relationships from a corpus of fugues from Bach's Well Tempered Clavier. Our approach allows the RBM to learn musical structure (i.e., tonal relationships) from a training corpus without having to hard-code tonal rules into the model. In fact, the high correlations between the model and listeners' performance lend support to the claim that domain general processing mechanisms, based on learning the probabilistic structure of sequential information, contribute to the acquisition of abstract, high-level relationships in music.

Our novel representation of musical input was inspired by how human listeners process pitch, and this method takes our model one step closer to an embodied approach to modeling music cognition. Future work will investigate using the entire frequency spectrum of every tone (e.g., as sampled from audio recordings). The full spectrum of pitch information may result in even better model performance on pitch-related tasks, especially with regard to octave equivalence.

As hypothesized, a harmonics-based representation assists the model in learning the tonal hierarchy and octave similarity from pitches that share harmonics. The model best simulated expert listeners, which can be taken as evidence that trained musicians likely take advantage of the harmonic spectrum of musical pitches in order to (implicitly) perceptually organize the pitches within a key. Our findings may also support the claim that novice listeners focus more on fundamental frequency and pitch proximity than harmonics. More generally, these findings highlight how the choice of representation can have a notable impact on learned features, and that alternative representations may be used to simulate different populations.

An extension of this work will consider using representations based on subharmonic patterns, as these are consistent with temporal models of pitch perception (Cariani, 2001)). In addition, superior model performance may result from using stacked RBMs, a method currently popular in the area of deep learning, as additional layers (model depth) may allow the model to learn increasingly abstract features of tonal relationships.

Acknowledgments

This research was made possible through support from the European Commission. The Lrn2Cre8 and PROSECCO projects acknowledge the financial support of the Future and Emerging Technologies (FET) programme within the Seventh Framework Programme for Research of the European Commission, under FET grant number 610859 (Lrn2Cre8) and FET grant number 600653 (PROSECCO).

References

Bartlett, M. S. (2001). *Face image analysis by unsupervised learning*. Springer Science & Business Media.

Berwick, R. C., Pietroski, P., Yankama, B., & Chomsky, N. (2011). Poverty of the stimulus revisited. *Cognitive Science*, *35*(7), 1207–1242.

Bishop, C. M. (1995). *Neural networks for patter recognition*. Claredon Press, Oxford.

Boulanger-Lewandowski, N., Bengio, Y., & Vincent, P. (2012). Modeling temporal dependencies in high-dimensional sequences: Application to polyphonic music generation and transcription. In *Proceedings of the 29th international conference on machine learning*.

Cancino, C., Lattner, S., & Grachten, M. (2014). Developing tonal perception through unsupervised learning. In

Proceedings of the 15th international society for music information retrieval conference.

Cancino, C., Lattner, S., & Grachten, M. (2015). *Derivations of the free energy of restricted boltzmann machines* (Technical Report). Austrian Research Institute for Artificial Intelligence.

Cariani, P. (2001). Temporal codes, timing nets, and music perception. *Journal of New Music Research*, *30*(2), 107–135.

Chalnick, A., & Billman, D. (1988). Unsupervised learning of correlational structure. In *Proceedings of the tenth annual conference of the cognitive science society* (pp. 510–516). Hillsdale, NJ: Lawrence Erlbaum Associates.

Chater, N., & Manning, C. D. (2006). Probabilistic models of language processing and acquisition. *Trends in cognitive sciences*, *10*(7), 335–344.

Cuddy, L. L., & Badertscher, B. (1987). Recovery of the tonal hierarchy: Some comparisons across age and levels of musical experience. *Perception & Psychophysics*, *41*(6), 609–620.

De Cheveigne, A. (2005). Pitch perception models. In *Pitch* (pp. 169–233). Springer.

Demany, L., & Armand, F. (1984). The perceptual reality of tone chroma in early infancy. *The journal of the Acoustical Society of America*, *76*(1), 57–66.

Feigenbaum, E. A. (1963). The simulation of verbal learning behavior. In E. A. Feigenbaum & J. Feldman (Eds.), *Computers and thought.* New York: McGraw-Hill.

Gomez, R. L., & Gerken, L. (1999). Artificial grammar learning by 1-year-olds leads to specific and abstract knowledge. *Cognition*, *70*(2), 109–135.

Grachten, M., & Krebs, F. (2014). An assessment of learned score features for modeling expressive dynamics in music. IEEE.

Hill, J. A. C. (1983). A computational model of language acquisition in the two-year old. *Cognition and Brain Theory*, *6*, 287–317.

Hinton, G. E. (2002, July). Training products of experts by minimizing contrastive divergence. *Neural Computation*, *14*(8), 1771–1800.

Hinton, G. E. (2007). Learning multiple layers of representation. *Trends in cognitive sciences*, *11*(10), 428–434.

Hinton, G. E. (2012). A practical guide to training restricted boltzmann machines. *Neural Networks: Tricks of the Trade*.

Krumhansl, C. L. (1990). *Cognitive foundations of musical pitch* (Vol. 17). Oxford University Press New York.

Krumhansl, C. L., & Kessler, E. J. (1982). Tracing the dynamic changes in perceived tonal organization in a spatial representation of musical keys. *Psychological review*, *89*(4), 334.

Krumhansl, C. L., & Shepard, R. N. (1979). Quantification of the hierarchy of tonal functions within a diatonic context. *Journal of experimental psychology: Human Perception and Performance*, *5*(4), 579.

Kullback, S., & Leibler, R. A. (1951). On information and sufficiency. *The Annals of Mathematical Statistics*, *22*(1), 79-86.

Lee, H., Ekanadham, C., & Ng, A. Y. (2008). Sparse deep belief net model for visual area V2. In *Advances in neural information processing systems 20* (pp. 873–880).

Leman, M. (1995). A model of retroactive tone-center perception. *Music Perception*.

Lerdahl, F., & Jackendoff, R. (1983). *A generative theory of tonal music*. MIT Press.

Licklider, J. (1951). A duplex theory of pitch perception. *The Journal of the Acoustical Society of America*, *23*(1), 147–147.

Matlock, T. (2001). *How real is fictive motion?* Doctoral dissertation, Psychology Department, University of California, Santa Cruz.

McDermott, J. H., & Oxenham, A. J. (2008). Music perception, pitch, and the auditory system. *Current opinion in neurobiology*, *18*(4), 452–463.

Meddis, R., & Hewitt, M. J. (1991). Virtual pitch and phase sensitivity of a computer model of the auditory periphery. i: Pitch identification. *The Journal of the Acoustical Society of America*, *89*(6), 2866–2882.

Newell, A., & Simon, H. A. (1972). *Human problem solving*. Englewood Cliffs, NJ: Prentice-Hall.

Ohlsson, S., & Langley, P. (1985). *Identifying solution paths in cognitive diagnosis* (Tech. Rep. No. CMU-RI-TR-85-2). Pittsburgh, PA: Carnegie Mellon University.

Pearce, M. T. (2005). *The construction and evaluation of statistical models of melodic structure in music perception and composition*. Unpublished doctoral dissertation, City University London.

Pinker, S. (1994). *The language instinct: The new science of language and mind* (Vol. 7529). Penguin UK.

Saffran, J. R., Johnson, E. K., Aslin, R. N., & Newport, E. L. (1999). Statistical learning of tone sequences by human infants and adults. *Cognition*, *70*(1), 27–52.

Saffran, J. R., & Wilson, D. P. (2003). From syllables to syntax: Multilevel statistical learning by 12-month-old infants. *Infancy*, *4*(2), 273–284.

Shepard, R. N. (1987). Toward a universal law of generalization for psychological science. *Science*, *237*, 1317-1323.

Shrager, J., & Langley, P. (Eds.). (1990). *Computational models of scientific discovery and theory formation*. San Mateo, CA: Morgan Kaufmann.

Smith, N. A., & Schmuckler, M. A. (2004). The perception of tonal structure through the differentiation and organization of pitches. *Journal of experimental psychology: human perception and performance*, *30*(2), 268.

Tillmann, B., Bharucha, J., & Bigand, E. (2000). Implicit learning of tonality: a self-organizing approach. *Psychol. Rev.*, *107*, 885.

van Noorden, L. (1982). Two channel pitch perception. In *Music, mind, and brain* (pp. 251–269). Springer.