AN ENERGY-BASED GENERATIVE SEQUENCE MODEL FOR TESTING SENSORY THEORIES OF WESTERN HARMONY

Peter M. C. Harrison
Queen Mary University of London
Cognitive Science Research Group

Marcus T. Pearce
Queen Mary University of London
Cognitive Science Research Group

ABSTRACT

The relationship between sensory consonance and Western harmony is an important topic in music theory and psychology. We introduce new methods for analysing this relationship, and apply them to large corpora representing three prominent genres of Western music: classical, popular, and jazz music. These methods centre on a generative sequence model with an exponential-family energy-based form that predicts chord sequences from continuous features. We use this model to investigate one aspect of instantaneous consonance (harmonicity) and two aspects of sequential consonance (spectral distance and voice-leading distance). Applied to our three musical genres, the results generally support the relationship between sensory consonance and harmony, but lead us to question the high importance attributed to spectral distance in the psychological literature. We anticipate that our methods will provide a useful platform for future work linking music psychology to music theory.

1. INTRODUCTION

Music theorists and psychologists have long sought to understand how Western harmony may be shaped by natural phenomena universal to all humans [13, 27, 36]. Key to this work is the notion of *sensory consonance*, describing a sound's natural pleasantness [32, 35, 38], and its inverse *sensory dissonance*, describing natural unpleasantness.

Sensory consonance has both *instantaneous* and *sequential* aspects. Instantaneous consonance is the consonance of an individual sound, whereas sequential consonance is a property of a progression between sounds.

Instantaneous sensory consonance primarily derives from *roughness* and *harmonicity*. Roughness is an unpleasant sensation caused by interactions between spectral components in the inner ear [8, 41], whereas harmonicity [1] is a pleasant percept elicited by a sound's resemblance to the harmonic series [4, 20].

---

[1] Related concepts include *tonalness* [27], *toneness* [15], *fusion* [14, 36], *complex sonorousness* [29], and *multiplicity* [29].

Sequential sensory consonance is primarily determined by *spectral distance* and *voice-leading distance*. Spectral distance [2] describes how much a sound's acoustic spectrum perceptually differs from neighbouring spectra [22–24, 27, 29]. Voice-leading distance [3] describes how far notes in one chord have to move to produce the next chord [2, 39, 40]. Consonance is associated with low spectral and voice-leading distance.

Many Western harmonic conventions can be rationalized as attempts to increase pleasantness by maximizing sensory consonance. The major triad maximizes consonance by minimizing roughness and maximizing harmonicity; the circle of fifths maximizes consonance by minimizing spectral distance; tritone substitutions are consonant through voice-leading efficiency [39].

This idea – that Western harmony seeks to maximize sensory consonance – has a long history in music theory [31]. Its empirical support is surprisingly limited, however. The best evidence comes from research linking sensory consonance maximization to rules from music theory [15, 27, 39], but this work is constrained by the subjectivity and limited scope of music-theoretic textbooks.

A better approach is to bypass textbooks and analyse musical scores directly. Usefully, large datasets of digitised musical scores are now available, as are many computational models of consonance. However, statistically linking them is non-trivial. One could calculate distributions of consonance features, but this would give only limited causal insight into how these distributions arise. Better insight might be achieved by regressing transition probabilities against consonance features, but this approach is statistically problematic because of variance heterogeneity induced by the inevitable sparsity of the transition tables.

This paper presents a new statistical model developed for tackling this problem. The model is generative and feature-based, defining a probability distribution over symbolic sequences based on features derived from these sequences. Unlike previous feature-based sequence models, it is specialized for continuous features, making it well-suited to consonance modelling. Moreover, the model parameters are easily interpretable and have quantifiable un-

---

[2] Spectral distance is also known by its antonym *spectral similarity* [23]. *Pitch commonality* [29] is a similar concept. Psychological models of harmony and tonality in the auditory short-term memory (ASTM) tradition typically rely on some kind of spectral distance measure [1, 7, 17].

[3] Voice-leading distance is termed *horizontal motion* in [2]. Parncutt's notion of *pitch distance* [28, 29] is also conceptually similar to voice-leading distance.

certainty, enabling error-controlled statistical inference.

We use this new model to test sensory theories of harmony as follows. We fit the model to corpora of chord sequences from classical, popular, and jazz music, using psychological models of sensory consonance as features. We then compute feature importance metrics to quantify how different aspects of consonance constrain harmonic movement. This work constitutes the first corpus analysis comprehensively linking sensory consonance to harmonic practice.

2. METHODS

2.1 Representations

2.1.1 Input

Chord progressions are represented as sequences of pitch-class sets. Exact chord repetitions are removed, but changes of chord inversion are represented as repeated pitch-class sets.

2.1.2 Pitch-Class Spectra

Some of our features use *pitch-class spectra* as defined in [22, 24]. A pitch-class spectrum is a continuous function that describes *perceptual weight* as a function of pitch class (p_c). Perceptual weight is the strength of perceptual evidence for a given pitch class being present. Pitch classes (p_c) take values in the interval $[0, 12)$ and relate to frequency (f, Hz scale) as follows:

$$p_c = \left[9 + 12 \log_2 \left(\frac{f}{440} \right) \right] \bmod 12. \quad (1)$$

Pitch-class sets are transformed to pitch-class spectra by expanding each pitch class into its implied harmonics. Pitch classes are modelled as harmonic complex tones with 12 harmonics, after [22]. The jth harmonic in a pitch class has level $j^{-\rho}$, where ρ is the roll-off parameter ($\rho > 0$). Partials are represented by Gaussians with mass equal to partial level, mean equal to partial pitch class, and standard deviation σ. Perceptual weights combine additively.

Formally, $W(p_c, X)$ defines a pitch-class spectrum, returning the perceptual weight at pitch-class p_c for an input pitch-class set $X = \{x_1, x_2, \ldots, x_m\}$:

$$W(p_c, X) = \sum_{i=1}^{m} T(p_c, x_i). \quad (2)$$

Here i indexes the pitch classes, and $T(p_c, x)$ is the contribution of a harmonic complex tone with fundamental pitch class x to an observation at pitch class p_c:

$$T(p_c, x) = \sum_{j=1}^{12} g\left(p_c, j^{-\rho}, h(x, j)\right). \quad (3)$$

Now j indexes the harmonics, $g(p_c, l, p_x)$ is the contribution from a harmonic with level l and pitch-class p_x to an observation at pitch-class p_c,

$$g(p_c, l, p_x) = \frac{l}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2} \left(\frac{d(p_c, p_x)}{\sigma} \right)^2 \right), \quad (4)$$

$d(p_x, p_y)$ is the distance between two pitch classes p_x and p_y,

$$d(p_x, p_y) = \min\left(|p_x - p_y|, 12 - |p_x - p_y| \right), \quad (5)$$

and $h(x, j)$ is the pitch class of the jth partial of a harmonic complex tone with fundamental pitch class x:

$$h(x, j) = (x + 12 \log_2 j) \bmod 12. \quad (6)$$

ρ and σ are set to 0.75 and 0.0683 after [22].

2.2 Features

2.2.1 Spectral Distance

Spectral distance is operationalised using the psychological model of [22, 24]. The spectral distance between two pitch-class sets X, Y is defined as 1 minus the continuous cosine similarity between the two pitch-class spectra:

$$D(X, Y) = 1 - \frac{\int_0^{12} W(z, X)W(z, Y)\,dz}{\sqrt{\int_0^{12} W(z, X)^2\,dz}\sqrt{\int_0^{12} W(z, Y)^2\,dz}} \quad (7)$$

with W as defined in Equation 2. The measure takes values in the interval $[0, 1]$, where 0 indicates maximal similarity and 1 indicates maximal divergence.

2.2.2 Harmonicity

Our harmonicity model is inspired by the template-matching algorithms of [21] and [29]. The model simulates how listeners search the auditory spectrum for occurrences of harmonic spectra. These inferred harmonic spectra are termed *virtual pitches*. High harmonicity corresponds to a strong virtual pitch percept.

Our model differs from previous models in two ways. First, it uses a pitch-class representation, not a pitch representation. This makes it voicing-invariant and hence more suitable for modelling pitch-class sets. Second, it takes into account the strength of all virtual pitches in the spectrum, not just the strongest virtual pitch.

The model works as follows. The *virtual pitch-class spectrum* Q defines the spectral similarity of the pitch-class set X to a harmonic complex tone with pitch class p_c:

$$Q(p_c, X) = D(p_c, X) \quad (8)$$

with D as defined in Equation 7. Normalising Q to unit mass produces Q':

$$Q'(p_c, X) = \frac{Q(p_c, X)}{\int_0^{12} Q(y, X)\,dy}. \quad (9)$$

Previous models compute harmonicity by taking the peak of this spectrum. In our experience this works for small

chords but not for larger chords, where several virtual pitches need to be accounted for. We therefore instead compute a spectral peakiness measure. Several such measures are possible, but here we use Kullback-Leibler divergence from a uniform distribution. $H(X)$, the harmonicity of a pitch-class set X, can therefore be written as follows:

$$H(X) = \int_0^{12} Q'(y, X) \log_2 \left(12 \, Q'(y, X)\right) dy. \quad (10)$$

Harmonicity has a large negative correlation with the number of notes in a chord. Some correlation is expected, but not to this degree: the harmonicity model considers a tritone (the least consonant two-note chord) to be more consonant than a major triad (the most consonant three-note chord). We therefore separate the two phenomena by adding a 'chord size' feature, corresponding to the number of notes in a given chord, and rescaling harmonicity to zero mean and unit variance across all chords with a given chord size.

2.2.3 Roughness

Roughness has traditionally been considered to be an important contributor to sensory consonance, though some recent research disputes its importance [20]. We originally planned to include roughness in our model, but then discovered that the phenomenon is highly sensitive to chord voicing. Since the voicing of a pitch-class set is undefined, its roughness is therefore unpredictable. Roughness is therefore not modelled in the present study.

2.2.4 Voice-Leading Distance

A *voice leading* connects the individual notes in two pitch-class sets to form simultaneous melodies [39]. Pitch-class sets of different sizes can be connected by allowing pitch classes to participate in multiple melodies. *Voice-leading distance* is an aggregate measure of the resulting melodic distance. We operationalise voice-leading distance using [39]'s geometric model.

Consider two pitch-class sets $X = \{x_1, x_2, \ldots, x_m\}$ and $Y = \{y_1, y_2, \ldots, y_n\}$. A voice-leading between X and Y can be written $A \to B$ where $A = (a_1, a_2, \ldots, a_N)$, $B = (b_1, b_2, \ldots, b_N)$, and the following holds: if $x \in X$ then $x \in A$, if $y \in Y$ then $y \in B$, if $a \in A$ then $a \in X$, if $b \in B$ then $b \in Y$, and $n \leq N$.

The distance of the voice leading $A \to B$ is denoted $V(A, B)$ and uses the *taxicab* norm:

$$V(A, B) = \sum_{i=1}^{N} d(a_i, b_i) \quad (11)$$

with $d(a_i, b_i)$ as defined in Equation 5.

The voice-leading distance between pitch-class sets X, Y is then defined as the smallest value of $V(A, B)$ for all legal A, B. This minimal distance can be efficiently computed using the algorithm described in [39].

2.2.5 Summary

This section defined three sensory consonance features. These included one instantaneous measure (harmonicity) and two sequential measures (spectral distance, voice-leading distance). Harmonicity correlated strongly with chord size, which could have confounded our analyses. We therefore controlled for chord size by normalising harmonicity for each chord size and including chord size as a feature.

2.3 Statistical Model

2.3.1 Overview

The statistical model is *generative*, defining a probability distribution over chord sequences (e.g. [12, 25, 33]). It is *feature-based*, using features of the chord and its context to predict chord probabilities (e.g. [12]). It is *energy-based*, defining scalar energies for each feature configuration which are then transformed and normalised to produce the final probability distribution (e.g. [3, 10, 30]). It is *exponential-family* in that the energy function is a linear function of the feature vector (e.g. [10, 30]). Informally, the model might be said to generalise linear regression to symbolic sequences.

2.3.2 Form

Let $\mathcal{A}$ denote the set of all possible chords, and let e_0^n denote a chord sequence of length n, where e_0 is always a generic start symbol. Let $e_i \in \mathcal{A}$ denote the ith chord and e_i^j the subsequence $(e_i, e_{i+1}, \ldots, e_j)$. Let $\mathbf{w}$ be the weight vector that parametrises the model.

The probability of a chord sequence is factorised into a chain of conditional chord probabilities.

$$P\left(e_0^n \mid \mathbf{w}\right) = \prod_{i=1}^{n} P\left(e_i \mid e_0^{i-1}, \mathbf{w}\right) \quad (12)$$

These are given energy-based expressions:

$$P\left(e_i \mid e_0^{i-1}, \mathbf{w}\right) = \frac{\exp\left(-E(e_0^{i-1}, e_i, \mathbf{w})\right)}{Z(e_0^{i-1}, \mathbf{w})} \quad (13)$$

where E is the *energy function* and Z is the *partition function*. Z normalises the probability distribution to unit mass:

$$Z(e_0^{i-1}, \mathbf{w}) = \sum_{x \in \mathcal{A}} \exp\left(-E(e_0^{i-1}, x, \mathbf{w})\right). \quad (14)$$

High E corresponds to low probability. E is defined as a sum of *feature functions*, f_j, weighted by $-\mathbf{w}$:

$$E(e_0^{i-1}, x, \mathbf{w}) = -\sum_{j=1}^{m} f_j(e_0^{i-1} :: x) w_j \quad (15)$$

where w_j is the jth component of $\mathbf{w}$, m is the dimensionality of $\mathbf{w}$, equalling the number of feature functions f_j, and $e_0^{i-1} :: x$ is the concatenation of e_0^{i-1} and $x \in \mathcal{A}$.

Feature functions measure a property of the last element of a sequence. Our feature functions are chord size,

harmonicity, spectral distance, and voice-leading distance. Chord size and harmonicity are context-independent, whereas spectral and voice-leading distance relate the last chord to the penultimate chord. When the penultimate chord is undefined, mean values are imputed for spectral and voice-leading distance, with the mean computed over all possible chord transitions.

2.3.3 Estimation

The model is parametrised by the weight vector **w**. This weight vector is optimised using maximum-likelihood estimation on a corpus of sequences, as follows.

Let $e_{0,k}^{n_k}$ denote the kth sequence from a corpus of size N, where n_k is the sequence's length. The negative log-likelihood of the weight vector **w** with respect to the corpus is then

$$C(\mathbf{w}) = -\sum_{k=1}^{N} \sum_{i=1}^{n_k} \log P(e_{i,k}|e_{0,k}^{i-1}, \mathbf{w}). \qquad (16)$$

After some algebra, the gradient can be written

$$\frac{dC}{d\mathbf{w}} = \sum_{k=1}^{N} \sum_{i=1}^{n_k} \frac{Z'(e_{0,k}^{i-1}, \mathbf{w})}{Z(e_{0,k}^{i-1}, \mathbf{w})} - \mathbf{f}(e_{0,k}^{i}) \qquad (17)$$

where

$$Z'(e_{0,k}^{i-1}, \mathbf{w}) = \sum_{x \in \mathcal{A}} \mathbf{f}(e_{0,k}^{i-1} :: x) \exp\left(-E(e_{0,k}^{i-1}, x, \mathbf{w})\right) \qquad (18)$$

and **f** is the vector of feature functions. This expression can be plugged into a generic optimiser to find a weight vector minimising the negative log-likelihood. The present work used the BFGS optimiser [37].

2.3.4 Feature Importance

This section introduces three complementary feature importance measures. These are *weight*, *explained entropy*, and *unique explained entropy*.

Weight describes a feature's relationship to chord probability. The weight for a feature function f_j is the parameter w_j, corresponding to (minus) the change in the energy function E in response to a one-unit change in the feature function f_j (Equation 15). Weight is a signed feature importance measure: the sign dictates whether the model prefers high (positive weight) or low (negative weight) feature values, and the magnitude dictates the strength of preference. To aid weight comparability between features, feature functions are scaled to unit variance over the set of all possible chord transitions.

Dividing the cost function (Equation 16) by the number of chords in the corpus ($\sum_{k=1}^{N} n_k$) gives an estimate of *cross entropy* in units of *nats*. Cross entropy measures chord-wise unpredictability with respect to a given model. From it we define two further measures: *explained entropy* and *unique explained entropy*.

Explained entropy for a feature f_j is computed by comparing cross entropy estimates for two models: a model trained using feature f_j and a null model trained with no features. Explained entropy is the difference between the two cross entropies. Higher values indicate that the feature explains a lot of structure in the corpus.

Unique explained entropy for a feature f_j is the amount that cross entropy changes when feature f_j is removed from the full feature set. It measures the unique explanatory power of a feature while controlling for other features.

2.3.5 Related Work

The literature contains several alternative approaches for feature-based modelling of chord sequences. One is the *multiple viewpoint method* [11, 12]. However, this method is specialised for discrete features, not the continuous features required for consonance modelling. A second alternative is the *maximum-entropy* approach of [10, 30]. This approach has some formal similarities with the present work, but its binary feature functions are incompatible with our continuous features. A third possibility is the *feature-based dynamic networks* of [33]; however, these networks would need substantial modification to represent the kind of feature dependencies required here.

2.4 Corpora

Our corpora represent three musical genres: classical music (1,022 movements/pieces), popular music (739 pieces), and jazz music (1,186 pieces). The classical corpus was compiled from *KernScores* [34], including ensemble music and keyboard music from several several major composers of common-practice tonal music (Bach, Haydn, Mozart, Beethoven, Chopin). Chord labels were obtained using the algorithm of [26] with an expanded chord dictionary, and with segment boundaries co-located with metrical beat locations as estimated from time signatures. Chord inversions were identified as the lowest-pitch chord tone in the harmonic segment being analysed. The popular and jazz corpora corresponded to publicly available datasets: the McGill Billboard corpus [6] and the *iRB* corpus [5].

2.5 Efficiency

Computation can be reduced by identifying repeated terms in the cost and cost gradient (Equations 16, 17). These repeated terms only need to be evaluated once. Our feature functions never look back further than the previous chord, and they are invariant to chord transposition; this means that repeated terms occur whenever a chord pair is repeated at some transposition. Collapsing over these repetitions reduces computation by a factor of 20–100 for our corpora.

2.6 Numeric Integration

The features related to pitch-class spectra all use integration. These integrals are numerically approximated using the rectangle rule with 1,200 subintervals, after [24].

2.7 Software

The statistical model was implemented in R and C++; source code is available from the authors on request.

3. RESULTS

3.1 Corpus level

Figure 1 plots feature importances for the three consonance measures: harmonicity (normalised by chord size), spectral distance, and voice-leading distance. Analyses are split by musical corpus, and confidence intervals are calculated using nonparametric bootstrapping [9].

3.1.1 Importance by Feature

All the consonance features contribute to harmonic structure in some way. The order of feature importance is fairly consistent between genres and importance measures. Broadly speaking, voice-leading distance is most important, followed by harmonicity, then spectral distance.

3.1.2 Importance by Corpus

Harmonicity is particularly important for popular music, less so for classical, and least for jazz. Spectral distance is most important for classical music, less so for popular, and unimportant for jazz.

The relative importance of voice-leading distance depends on the measure used: it scores highly on explained entropy, but less on weight and unique explained entropy. This may be because voice-leading distance and chord size capture some common information: moving from a small chord to a large chord typically involves a large voice-leading distance. If we wish to assess the unique effect of voice-leading distance, we can look at weight and unique explained entropy: these measures tell us that voice-leading distance is most important for jazz music, less for classical music, and least for popular music.

3.1.3 Signs of Weights

The sign of a feature weight determines whether the model prefers positive or negative values of the feature. The observed feature signs are all consistent with theory. Harmonicity has a positive weight for all genres, indicating that harmonicity is universally promoted. Spectral distance and voice-leading distance both have negative weights, indicating preference for lower values of these features.

3.2 Composition Level

We also explored the application of these techniques to individual compositions (Figure 2). While the composition-level analyses reflect the same trends as the corpus-level analyses (Figure 1), they also reveal substantial overlap between the corpora. We assessed the extent of this overlap by training a generic machine-learning classifier to predict genre from the complete set of feature importance measures. Our classifier was a random forest model trained using the randomForest package in R [18], with 2,000 trees and four variables sampled at each split. Performance was assessed using 10-fold cross-validation repeated and averaged over 10 runs, resulting in a classification accuracy of 86% and a kappa statistic of .79. This indicates that genre differences in sensory consonance are moderately salient, even at the composition level.

4. CONCLUSION

This paper introduces new methods for testing relationships between sensory consonance and Western harmony. The methods centre on a new statistical model that predicts symbolic sequences using continuous features. We demonstrate these methods through application to three corpora representing classical, popular, and jazz music.

The results strongly support theoretical relationships between sensory consonance and harmonic structure. The three aspects of sensory consonance tested – harmonicity, spectral distance, and voice-leading distance – all predicted harmonic movement. Not all aspects were equally important, however. Spectral distance performed poorly, particularly in jazz. This is interesting given the high importance attributed to spectral distance in recent psychological literature [1, 7, 22]. Harmonicity performed well in popular music, but less so in classical and jazz. In contrast, voice-leading distance performed consistently well.

The corpus analyses deserve further development. It would be worth probing the true universality of sensory consonance by exploring a broader range of styles and using more refined stylistic categories, possibly at the level of the composer. The validity of the classical analyses could also be improved through more principled sampling [19] and manual chord-labelling [16].

The three feature importance measures provide useful complementary perspectives, but it is unnecessary to plot each one every time. In future we recommend inspecting the weights to check whether a feature is promoted or avoided, but then plotting just unique explained entropy. Unique explained entropy is preferable to weight because its units are well-defined, and preferable to explained entropy because it controls for other features, thereby providing a better handle on causality.

We focused on interpreting the statistical model through feature importance measures, but an alternative strategy would be to use the model to generate chord sequences for subjective evaluation. This route lacks the objectivity of feature-importance analysis, but it would give a uniquely intuitive perspective on what the model has learned.

The modelling techniques could be developed further. An important limitation of the current model is the linearity of the energy function, which restricts it to monotonic feature effects. A polynomial energy function would address this problem. It would also be interesting to develop the psychological features further, perhaps adding echoic memory to the spectral distance measure [17], and introducing an octave-generalised roughness measure.

Despite these limitations, we believe that the current results have important implications for our understanding of Western tonal harmony. In particular, the results imply that voice-leading efficiency is a better candidate for a harmonic universal than spectral similarity. This result is important for music psychology, where voice-leading efficiency is relatively underemphasised compared to harmonicity and spectral similarity (though see [2, 27, 39]). Future psychological work may wish to re-examine the role of voice-leading efficiency in harmony perception.
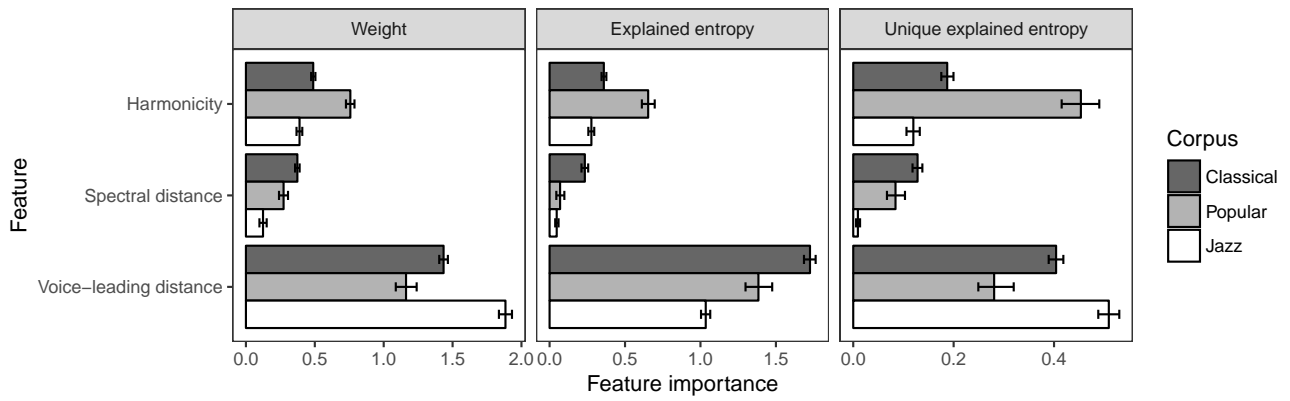
Figure 1. Measures of feature importance as a function of musical corpus. These measures are calculated from statistical models trained on the corpus level. Error bars represent 99% confidence intervals estimated by nonparametric bootstrapping [9]. Signs of feature weights are reversed for spectral distance and voice-leading distance, so that positive weights correspond to consonance maximisation.
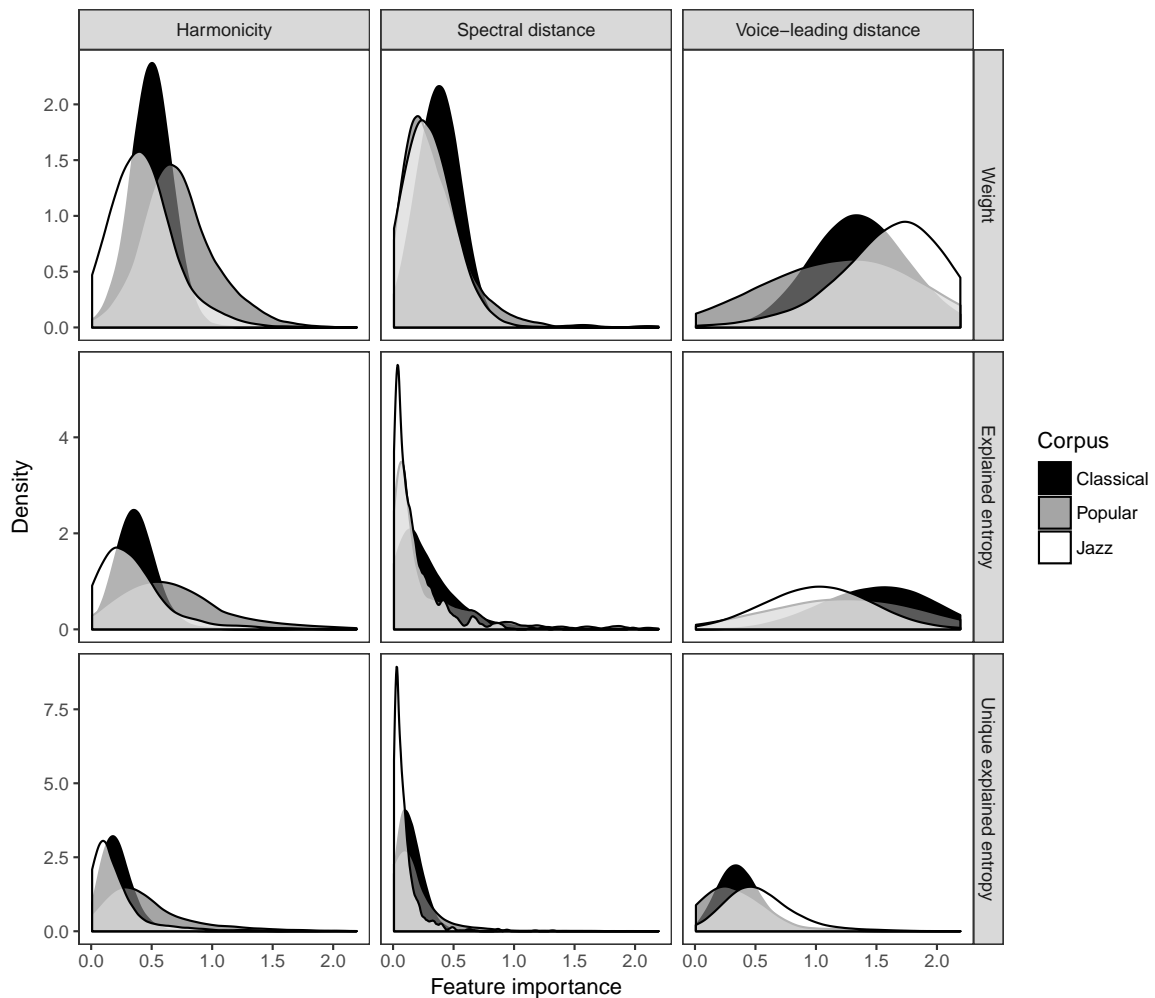
Figure 2. Distributions of feature importance measures as calculated for individual compositions within the three corpora. Distributions are represented by Epanechnikov kernel density functions. Signs of feature weights are reversed for spectral distance and voice-leading distance, so that positive weights correspond to consonance maximisation.

5. ACKNOWLEDGEMENTS

6. REFERENCES

[1] Emmanuel Bigand, Charles Delbé, Bénédicte Poulin-Charronnat, Marc Leman, and Barbara Tillmann. Empirical evidence for musical syntax processing? Computer simulations reveal the contribution of auditory short-term memory. *Frontiers in Systems Neuroscience*, 8, 2014.

[2] Emmanuel Bigand, Richard Parncutt, and Fred Lerdahl. Perception of musical tension in short chord sequences: The influence of harmonic function, sensory dissonance, horizontal motion, and musical training. *Perception & Psychophysics*, 58(1):124–141, 1996.

[3] Nicolas Boulanger-Lewandowski, Pascal Vincent, and Yoshua Bengio. Modeling temporal dependencies in high-dimensional sequences: Application to polyphonic music generation and transcription. In *Proc. of the 29th International Conference on Machine Learning (ICML-12)*, 2012.

[4] Daniel L. Bowling and Dale Purves. A biological rationale for musical consonance. *Proceedings of the National Academy of Sciences*, 112(36):11155–11160, 2015.

[5] Yuri Broze and Daniel Shanahan. Diachronic changes in jazz harmony: A cognitive perspective. *Music Perception*, 31(1):32–45, 2013.

[6] John Ashley Burgoyne. *Stochastic Processes & Database-Driven Musicology*. PhD thesis, McGill University, Montréal, Québec, Canada, 2011.

[7] Tom Collins, Barbara Tillmann, Frederick S. Barrett, Charles Delbé, and Petr Janata. A combined model of sensory and cognitive representations underlying tonal expectations in music: From audio signals to behavior. *Psychological Review*, 121(1):33–65, 2014.

[8] P. Daniel and R. Weber. Psychoacoustical roughness: Implementation of an optimized model. *Acta Acustica united with Acustica*, 83(1):113–123, 1997.

[9] B. Efron and R. J. Tibshirani. *An introduction to the bootstrap*. Chapman & Hall, Boca Raton, FL, 1993.

[10] Gaëtan Hadjeres, Jason Sakellariou, and François Pachet. Style imitation and chord invention in polyphonic music with exponential families. http://arxiv.org/abs/1609.05152, 2016.

[11] Peter M. C. Harrison and Marcus T. Pearce. A statistical-learning model of harmony perception. In *Proc. of DMRN+12: Digital Music Research Network One-Day Workshop*, page 15, London, UK, 2017.

[12] Thomas Hedges and Geraint A. Wiggins. The prediction of merged attributes with multiple viewpoint systems. *Journal of New Music Research*, 45(4):314–332, 2016.

[13] Hermann Helmholtz. *On the sensations of tone*. Dover, New York, NY, 1954. First published in 1863; translated by Alexander J. Ellis.

[14] David Huron. Tonal consonance versus tonal fusion in polyphonic sonorities. *Music Perception*, 9(2):135–154, 1991.

[15] David Huron. Tone and voice: A derivation of the rules of voice-leading from perceptual principles. *Music Perception*, 19(1):1–64, 2001.

[16] Nori Jacoby, Naftali Tishby, and Dmitri Tymoczko. An information theoretic approach to chord categorization and functional harmony. *Journal of New Music Research*, 44(3):219–244, 2015.

[17] Marc Leman. An auditory model of the role of short-term memory in probe-tone ratings. *Music Perception*, 17(4):481–509, 2000.

[18] Andy Liaw and Matthew Wiener. Classification and regression by randomForest. *R news*, 2(3):18–22, 2002.

[19] Justin London. Building a representative corpus of classical music. *Music Perception*, 31(1):68–90, 2013.

[20] Josh H. McDermott, Andriana J. Lehr, and Andrew J. Oxenham. Individual differences reveal the basis of consonance. *Current Biology*, 20(11):1035–1041, 2010.

[21] Andrew J. Milne. *A computational model of the cognition of tonality*. PhD thesis, The Open University, Milton Keynes, UK, 2013.

[22] Andrew J. Milne and Simon Holland. Empirically testing Tonnetz, voice-leading, and spectral models of perceived triadic distance. *Journal of Mathematics and Music*, 10(1):59–85, 2016.

[23] Andrew J. Milne, Robin Laney, and David Sharp. A spectral pitch class model of the probe tone data and scalic tonality. *Music Perception*, 32(4):364–393, 2015.

[24] Andrew J. Milne, William A. Sethares, Robin Laney, and David B. Sharp. Modelling the similarity of pitch collections with expectation tensors. *Journal of Mathematics and Music*, 5(1):1–20, 2011.

[25] Jean-Francois Paiement, Douglas Eck, and Samy Bengio. A probabilistic model for chord progressions. In *Proc. of the 6th International Conference on Music Information Retrieval*, London, UK, 2005.

[26] Bryan Pardo and William P. Birmingham. Algorithms for chordal analysis. *Computer Music Journal*, 26(2):27–49, 2002.

[27] Richard Parncutt. *Harmony: A psychoacoustical approach*. Springer-Verlag, Berlin, Germany, 1989.

[28] Richard Parncutt and Graham Hair. Consonance and dissonance in music theory and psychology: Disentangling dissonant dichotomies. *Journal of Interdisciplinary Music Studies*, 5(2):119–166, 2011.

[29] Richard Parncutt and Hans Strasburger. Applying psychoacoustics in composition: "Harmonic" progressions of "nonharmonic" sonorities. *Perspectives of New Music*, 32(2):88–129, 1994.

[30] Jeremy Pickens and Costas Iliopoulos. Markov random fields and maximum entropy modeling for music information retrieval. In *Proc. of the 6th International Conference on Music Information Retrieval*, pages 207–214, London, UK, 2005.

[31] Jean-Philippe Rameau. *Treatise on harmony*. Jean-Baptiste-Christophe Ballard, Paris, France, 1722.

[32] Pascaline Regnault, Emmanuel Bigand, and Mireille Besson. Different brain mechanisms mediate sensitivity to sensory consonance and harmonic context: Evidence from auditory event-related brain potentials. *Journal of Cognitive Neuroscience*, 13(2):241–255, 2001.

[33] Martin A. Rohrmeier and Thore Graepel. Comparing feature-based models of harmony. In *Proc. of the 9th International Symp. on Computer Music Modeling and Retrieval (CMMR)*, pages 357–370, London, UK, 2012.

[34] C. S. Sapp. Online database of scores in the Humdrum file format. In *Proc. of the 6th International Society for Music Information Retrieval Conference (ISMIR 2005)*, pages 664–665, 2005.

[35] E. Glenn Schellenberg and Laurel J. Trainor. Sensory consonance and the perceptual similarity of complex-tone harmonic intervals: Tests of adult and infant listeners. *Journal of the Acoustical Society of America*, 100(5):3321–3328, 1996.

[36] Carl Stumpf. *The Origins of Music*. Oxford University Press, Oxford, UK, 2012. First published in 1911; translated by David Trippett.

[37] Wenyu Sun and Ya-Xiang Yuan. *Optimization Theory and Methods: Nonlinear Programming*. Springer Science & Business Media, New York, NY, 2006.

[38] Laurel J. Trainor, Christine D. Tsang, and Vivian H. W. Cheung. Preference for sensory consonance in 2- and 4-month-old infants. *Music Perception*, 20(2):187–194, 2002.

[39] Dmitri Tymoczko. The geometry of musical chords. *Science*, 313(5783):72–74, 2006.

[40] Dmitri Tymoczko. *A Geometry of Music*. Oxford University Press, New York, NY, 2011.

[41] Václav Vencovský. Roughness prediction based on a model of cochlear hydrodynamics. *Archives of Acoustics*, 41(2):189–201, 2016.