The Use of Large Corpora to Train a New Type of Key-Finding Algorithm: An Improved Treatment of the Minor Mode

Author(s): Joshua Albrecht and Daniel Shanahan

Source: *Music Perception: An Interdisciplinary Journal*, Vol. 31, No. 1 (September 2013), pp. 59-67
Published by: University of California Press
Stable URL: http://www.jstor.org/stable/10.1525/mp.2013.31.1.59
Accessed: 01/02/2015 15:33

The Use of Large Corpora to Train a New Type of Key-Finding Algorithm: An Improved Treatment of the Minor Mode

Joshua Albrecht
The University of Mary Hardin-Baylor

Daniel Shanahan
Ohio State University

Computational models of key estimation have struggled to emulate the accuracy levels of human listeners, especially with pieces in the minor mode. The current study proposes a new key-finding algorithm, which utilizes Euclidean distance, rather than correlation, and is trained on the statistical properties of a large musical sample. A model was trained on a dataset of 490 pieces encoded into the Humdrum "kern" format, in which the key was known. This model was tested on a reserve dataset of 492 pieces, and was found to have a significantly higher overall accuracy than previous models. In addition, we determined separate accuracy ratings for major mode and minor mode works for the existing key-finding models and report that most existing models provide greater accuracy for major mode rather than minor mode works. The proposed key-finding algorithm performs more accurately on minor mode works than all of the other models tested, although it does not perform significantly better than the models created by Aarden (2003), Bellman (2005), or Sapp (2011). Finally, an algorithm that combines the Aarden-Essen model (2003) and the proposed algorithm is suggested, and results in significantly more accurate key assessments than all of the other extant models.

Received: October 17, 2012, accepted March 13, 2013.

Key words: key-finding, algorithm, computational methods, corpus study, minor mode

As **Chew (2000) writes, computational** music analysis is, at its core, an interdisciplinary effort. Computational models and algorithms are intended to serve as analogues for how we understand music, and their success is often judged by how close they are able to approximate the human understanding of music. For example, probabilistic models, such as those proposed by Temperley (2007, 2008), are meant to establish a more thorough understanding of cognitive processes that occur when one infers tonality from musical surface and structure. Yet, there is often a disparity between the elements of tonal induction that are most salient to listeners, and the ability of a computational model to "understand" these elements. This is rarely more evident than when analyzing models of key-finding. For decades, key-finding has been seen as a primary obstacle in computational musicology. In order to perform certain types of analyses involving scale-degree or function on large sets of musical data, keys must first be appropriately established. Unfortunately, however, using algorithms to assign keys to pieces often fails to emulate the ability of human analysts, especially with pieces in a minor mode. Although listeners are readily able to discern pieces in a minor key, computational models often misinterpret such pieces as being in the relative major.

According to Krumhansl, the ability to infer the tonality of a piece of music appears to rely on both learned cognitive reference points (familiar musical structures used to infer larger hierarchies), as well as "the sensitivity to the frequencies with which instances occur" (Krumhansl, 2000). In other words, like the perception of local dependencies in music (Altman, Dienes, & Good, 1995; Bigand, Perruchet, & Boyer, 1998; Loui, Wessel, & Kam, 2010; Saffran, Johnson, Aslin, & Newport, 1999), as well as more complex musical dependencies (Dienes & Longuet-Higgens, 2004), the learning of the statistical properties of key-structure would likely play an important role in the perception of tonality. Despite the importance of low-level statistical properties in influencing the perception of key, it is important to stress that understanding the tonal grammar of the idiom is a much more complicated phenomenon, including perception of the melodic, harmonic, rhythmic, and structural aspects of the style.

In order to examine the statistical properties that may influence the perception of tonality further, one might first compare various approaches to key-finding. Longuet-Higgins's *shape matching algorithm* (1976; Longuet-Higgins & Steedman, 1971) incorporated a harmonic network that focused on the process of elimination based on the pitch classes that occurred. Despite this

Music Perception, volume 31, issue 1, pp. 59–67, issn 0730-7829, electronic issn 1533-8312. DOI: 10.1525/mp.2013.31.1.59

TABLE 1. *The Accuracy Ratings for Key-Finding Methods Compared for Major, Minor, and Overall.*

	Entire Piece			1st and last 8 measures		
Algorithm	Major	Minor	Overall	Major	Minor	Overall
Krumhansl-Schmuckler	69.0%	83.2%	74.2%	85.3%	79.3%	83.1%
Temperley (Krumhansl-Schmuckler algorithm)	96.8%	74.3%	88.6%	94.6%	67.6%	84.8%
Bellman-Budge	94.9%	84.4%	91.1%	94.2%	86.6%	91.5%
Aarden-Essen	90.7%	93.3%	91.7%	94.9%	84.9%	89.8%
Sapp Simple Weightings	92.3%	87.2%	90.4%	95.2%	88.9%	92.9%
Proposed model (Krumhansl-Schmuckler algorithm)	92.7%	85.5%	90.0%	96.5%	83.8%	91.9%
Proposed model (Euclidean distance)	89.1%	95.0%	91.3%	94.2%	91.1%	93.1%

algorithm performing quite well on overtly tonal pieces (and correctly analyzing the keys of all 48 pieces of Bach's *Well-Tempered Clavier*), this type of model was less effective for pieces with any sort of non-diatonic tonicization or modulation (Temperley, 2007). Additionally, as it was largely concerned with monophonic excerpts, the shape matching algorithm was not able to take into account harmonies that might not be explicitly outlined in the melody (Steedman, 1984).

A more experimentally driven approach was taken by Krumhansl and Schmuckler (Krumhansl, 1990). In this algorithm, the pitch classes are counted for an entire work, and the resulting distribution is correlated with scale-degree distributions for all 24 major and minor works based on experimentally derived results (taken from the probe-tone studies in Krumhansl & Kessler, 1982). Aarden (2003) noted that the results of the probe-tone research was not necessarily reflective of the actual distribution of scale degrees. As such, he calculated the scale-degree distribution of folksong melodies (taken from the Essen folksong collection; Schaffrath, 1995) and from polyphonic music (250 non-modulating major key movements from the works of European composers primarily in the 18th and 19th centuries; MuseData database, CCARH, 2001). The Aarden model was a corpus-driven approach that provided more accurate results than the original Krumhansl-Schmuckler method. He argued that "the relationship between the 'key-profile' and musical structure is not as straightforward as is often assumed" (Aarden, 2003). Similarly, Bellman (2005) constructed a model that was derived from correlations calculated from Budge's (1943) study of chord frequencies of 18th and 19th century composers. Suggesting a different perspective, Sapp (2011) proposed a simple set of scale degree weights to be used with the Krumhansl-Schmuckler algorithm, consisting of only 2 (for tonic and dominant), 1 (for other diatonic tones), and 0 (for non-diatonic tones).

Temperley (2007, 2008) proposed an alternative approach in which the music was divided into segments, and the pitch-class distribution in each segment was matched with the scale-degree distributions for all 24 major and minor keys. Whichever probability distribution was most closely matched with the occurrence of pitch classes in the segment was taken as the key of the segment. In this way, Temperley's model offered an improvement over existing paradigms, in that it permitted different segments to be recognized as being in different keys, and identified ambiguous or modulating passages by revealing low probabilities of being in any particular key. While all of the approaches outlined above offer unique advantages in determining the key of a piece of music, they nevertheless contain their own inherent weaknesses. For example, the Krumhansl-Schmuckler method tends to skew toward the dominant. The Aarden approach tends to skew toward the subdominant, whereas the Temperley method tends to identify the relative major when analyzing minor key works. For a further analysis, see Sapp (2011).

An important observation about the existing key-finding models that has been somewhat neglected in the literature is the tendency for key-finding approaches to register much greater accuracy in determining the key of major mode works than of minor mode works. This disparity can be seen in Table 1 above. One approach to assessing a new key-finding algorithm might therefore be an examination of its ability to recognize pieces in the minor mode, rather than simply by improving the overall accuracy rating of key-finding. To anticipate our results, the method proposed here provides a more consistent approach to accurately analyzing the overall keys of pieces, whether in a minor or major mode.

Toward a New Key-Finding Method

Further discrepancies between human and computational models of perception arise when analyzing keys through correlational approaches. For example, very few music analysts would doubt that the final fugue from the second book of Bach's *Well-Tempered Clavier*

TABLE 2. *List of Composers and Compositions Used in Our Reserve and Testing Set.*

Composer	Composition	Number of Pieces (Movements Counted Separately)
Bach	The Brandenburg Concertos	19
	Chorales	323
	Sinfonias	15
	Inventions	15
	Well-Tempered Clavier	48
Beethoven	First and last movements of all string quartets	32
	First and last movements of all piano sonatas	61
Brahms	Op. 51	3
Chopin	Op. 28	24
	Mazurkas	49
Corelli	Trio Sonatas, Op. 1	57
Haydn	First and Last Movements of Each String Quartet	105
Hummel	Op. 67 *Préludes*	24
Kabalevsky	"Happy" Variations on a folksong	5
Miscellaneous	Barbershop Quartet Arrangements	33
Mozart	First and Last Movements of Each String Quartet	47
Telemann	3 Klavier Fantasies	3
Scarlatti, D.	58 Piano Sonatas	58
Vivaldi	Op. 8	36

is in B minor, yet a correlational approach (specifically, with the Krumhansl-Kessler weightings) would identify the piece as being in F# minor (and with a high correlation of .76). As a point of reference, the fourteenth fugue from the second book of The Well-Tempered Clavier (which is nominally in F# minor) is correctly identified, with a correlation of .82. The issue becomes even more complicated when analyzing pieces that are decidedly less tonal. For example, the Krumhansl-Schmuckler correlational approach would return a key of F# minor for Webern's Op.12, No.1 (Der tag ist vergangen, the first of the *Vier Lieder*), with a correlation of .69. While this piece is not the least tonal in Webern's oeuvre, it would be somewhat difficult to make the case that such a piece was in F# minor. Listeners are able to understand the multifaceted aspects of pitch hierarchy, harmonic rhythm, dynamics, metric accent, and extra-opus expectations (such as likely modulations in development sections), but correlational approaches depend largely on rank ordering of pitch classes. For example, if an atonal piece has 11 tones represented 100 times, and one tone represented 101 times, many correlational approaches will, with a high correlation, recommend that the latter pitch is the key of the piece. Pieces with an increased amount of variability among pitch classes might therefore be more susceptible to reduced accuracy with a correlational approach.

This can be clearly seen in the difference between major and minor mode pieces. In our sample of 625 major mode and 357 minor mode works from the common practice era (see Table 2 above), we determined the average standard deviation of pitch distributions by mode. Major mode works had a significantly higher standard deviation ($p < .001$) of pitch distributions than minor mode works, meaning that major mode pitch class usage tends to be less evenly distributed. In other words, pitch classes tend to either be used a great deal or almost not at all. By contrast, the minor mode pitch class usage tends to be more uniform, in that there are a greater variety of pitch classes used with relatively similar frequency.

One alternative approach to these correlational key-finding methods is one that employs Euclidean distance (see Albrecht & Huron, in press; Chew, 2000). The best way of visualizing this is to imagine that the proportion of the presence of each scale degree is a measurement along a separate axis. In a two-dimensional space, if there were 70% of pitch X and 30% pitch Y, the Cartesian location of the point representing this pitch-class distribution would be at $X = 0.7$ and $Y = 0.3$. In this case, we are examining the distribution of 12 pitch classes, resulting in a 12-dimensional Cartesian space. The pitch-class distribution of each piece is represented by a point in that 12-dimensional space. The distance is then measured between this point and the 24 points representing the 12 major and 12 minor key pitch-class distributions, and the key separated by the shortest distance is taken to be the key of the work. Although the underlying mathematics share similarities, in practice this approach produces much different results from an algorithm derived by correlation (see Table 1), in which

the pitch-class distribution of the piece is correlated with the 24 pitch-class distributions of each key and the highest correlation is taken to be the "home" key.

One difficulty with using works from the tonal era is the tendency for pieces to modulate to other keys within the work. In most cases, these modulations are to closely-related keys, but throughout the 19th century, works began to explore more distant key relationships. Necessarily, this sort of large-scale tonal progression involves an increased level of chromaticism across an entire work or movement. Approaches that simply amalgamate pitch-class distributions across an entire work fail to consider that different phrases within a piece may have different functions within the large-scale tonal framework of that piece. For example, VanHandel and Callahan (2012) found that key perception is largely dependent upon phrase location, with the result that beginning and ending phrases facilitated accurate key identification more often than middle sections. One approach to increasing the reliability of a key-finding algorithm would be to limit the music sampled to that music within the piece's "home key." We therefore decided to restrict our sample to music within the first and last eight measures, which are more likely to be indicative of the overarching tonality of the piece. For purposes of comparison, we also tested the other algorithms on just the first and last eight measures (see Table 1).

A further concern is what distributions are optimal for taking as the "ground truth" of the distribution of scale degrees for both the major and the minor mode. As mentioned above, the Krumhansl-Schmuckler algorithm used the Krumhansl-Kessler distributions taken from their probe-tone research, whereas Aarden used the Essen folksong database to calculate scale-degree distributions from a corpus of music. A fundamental assumption behind many of these approaches is that there is one "ideal" distribution for all pieces in the major mode and another distribution for all pieces in the minor mode. Albrecht and Huron (in press) conducted a study in which the assumption of stable scale-degree distributions for each mode was directly tested. They sampled fifty works from each of seven 50-year periods from 1400–1750 and used a clustering procedure to determine how many modes existed in each period. They then calculated scale-degree distributions for each mode in each period. Albrecht and Huron saw a pattern of development in which the distributions of several mode-clusters in earlier periods merged into the major and minor mode by 1550–1600, and continued to develop into the 18th century.

While it is therefore somewhat problematic to assume that only one form of scale-degree distribution for each mode was operative over the entire span of common-practice tonality, it has nevertheless been the practice of earlier key-finding approaches. While further research might develop a more nuanced approach to scale-degree distributions across the era of common-practice tonality, the present study will assume a similarly (albeit updated) uniform nature of scale-degree distributions over time.

While a large-scale, diachronic study examining how scale-degree profiles change across the common-practice era would be useful, it is not practical for the sake of this study. However, a way of mitigating the limiting aspect of using only one "ideal" distribution for major or minor is to derive the distributions from a cross-section of music covering the era. Assuming that statistical learning plays an important role in tonality perception, this approach also has the advantage of simulating such an inductive key-finding approach by using the statistical properties of real music. In other words, instead of simply co-opting existing key profiles for use with the Euclidean distance method, the current study sets out to use distributions of pitch classes as they exist in real music; namely, on a set of compositions from between 1700–1950. Unfortunately, in this dataset there is an emphasis on early and middle Classical music (due to the nature of the corpus that was readily available). Nevertheless, by employing this new methodology we hoped to improve upon the overall accuracy of previous key-finding models, especially when determining the key of a piece written in the minor mode.

The Dataset

In order to test the existing models against the one currently proposed, a large dataset with a ground truth of key assignments was used (see Table 2). This dataset was collected as a convenience sample from a series of scores encoded into the kern format and manually assigned keys by independent scholars not connected with this project (CCARH, 2001). In most of these cases, the key of the work was explicitly stated in the title (e.g., "Sonata in G minor"). In the case of the Bach chorales, although 371 chorales were encoded, a number were explicitly analyzed as being in a church mode, rather than a major or minor key. In order to limit our sample to works only in major and minor keys, these chorales were excluded. In the case of the Haydn and Mozart string quartets, no key was explicitly assigned to the movements. However, the key of the piece as a whole was indicated in the title. First and last movements of string quartets are much more likely than middle movements to be in the key indicated in the title, and so these

movements were retained while the middle movements were excluded.

By comparing the results from the key-finding models against the ground truth assigned by independent scholars, an estimate of the accuracy of the models was attained. We used 982 works from the Humdrum database (CCARH, 2001; Huron, 1995). This collection can be seen in Table 2. In total, our database used 492 pieces as a reserve dataset, and 490 as a training set. In the reserve dataset, there were 313 pieces in a major key, and 179 pieces in a minor key. In the dataset used to train the algorithm, there were 312 pieces in a major key, and 178 pieces in a minor key.

The model dataset distributions were used to train our algorithm. In other words, the pieces in the first dataset were segregated by mode, and pitch-class distributions were calculated for the first and last eight measures for each piece for the major and minor modes. Each of these pieces included a key identification, and this key identification was used to rotate the piece's pitch-class distribution so that the tonic pitch would be the same for every piece. The resulting scale-degree distributions for each piece were amalgamated across all of the other pieces in the same mode to create an averaged distribution for each mode. These averages were taken to be the scale-degree distributions for the major and the minor mode for use with the proposed algorithm. The weightings for the major and minor modes are given in Appendices A and B.

Results

The proposed model, which used a Euclidean distance approach and scale degree information for only the first and last eight measures, was tested against the existing models, which use a correlational approach and scale degree information for the entire piece. Additionally, we used our proposed weights with the correlational approach. However, it may be that using the first and last eight measures provides an unfair advantage to the proposed model. For this reason, the other models were run twice—on pitch class information from the whole piece, and pitch class information for just the first and last eight measures.

As seen in Table 1, the proposed algorithm resulted in more correct key assignments (93.1%) than the other models, though this difference was not significant for all models. When looking at success rates within either the major or minor mode, however, a more detailed story emerges. A number of algorithms had comparable or better accuracy ratings than the proposed algorithm for major mode works, but the proposed model had a higher accuracy than all of the other algorithms when examining pieces in a minor mode, with the exception of the Aarden-Essen model.

When comparing key-finding models, one approach would be to examine the number of correct calculations, and simply assume that the model with the highest percentage is the best. This, however, does not provide enough insight into the specific nature of the differences between models. In order to determine whether one model significantly outperforms another, we employed a method that examined the disparities between models, and provided a conservative approach to multiple comparisons. This was accomplished by examining the specific pieces that one model correctly analyzed that another incorrectly analyzed. The number of correct pieces in the best-performing model and the total number of differences between the two models to be compared were then used to perform a binomial test (for more on the benefits of this method, see Salzberg, 1997). The binomial test compares the results that are different between models and performs a significance test to see if one model outperforms the other on pieces in which they differ at levels significantly higher than chance. Although other tests might have more statistical power, a conservative approach such as this might serve to remove some possible analytical biases.

The results from the significance test are shown in Table 3. As seen in the table, the proposed model outperforms the Krumhansl-Kessler and the Temperley models overall, regardless of whether they use the entire piece as input or only the first and last eight measures. While the Temperley model is significantly more accurate on major mode pieces, it is significantly less accurate on minor mode pieces and overall. The proposed model accurately identifies more pieces overall than the proposed weights with the correlation method, the Aarden-Essen, the Bellman-Budge, and the Sapp models, but the difference is not significant. One reason for the lack of significance could be the relatively conservative test. As can be seen in the bottom row of Table 3, the proposed model performs significantly better on pieces in the minor mode than the proposed weights with the correlation method, the Temperley model, and the Krumhansl-Kessler models. There is no significant advantage for the proposed model over the other methods.

Common Misattributions in the Proposed Model

When examining a key-finding model, the incorrect key assignments often reveal more than the correct key assignments. As mentioned above, many of the other models tend to favor particular relationships as misattributions

TABLE 3. *Performance of the Krumhansl-Kessler, Bellman-Budge, Aarden-Essen, Temperley, and Sapp Simple Models, as Well as Our Proposed Weights with the Krumhansl-Schmuckler Algorithm (PKS) Compared to the Proposed Model.*

	Significantly Worse	No Significant Difference	Significantly Better
Overall	Krumhansl-Kessler-1 ($p < .001$)* Krumhansl-Kessler-2 ($p < .001$)* Temperley-1 ($p < .001$)* Temperley-2 ($p < .001$)*	Bellman-Budge-1 ($p = .17$) Bellman-Budge-2 ($p = .33$) Aarden-Essen-1 ($p = .34$) Aarden-Essen-2 ($p = .34$) Sapp Simple-1 ($p = 1.00$) Sapp Simple-2 ($p = .07$) PKS-1 ($p = .26$) PKS-2 ($p = .03$)	
Major	Krumhansl-Kessler-1 ($p < .001$)* Krumhansl-Kessler-2 ($p < .001$)*	Aarden-Essen-1 ($p = .12$) Aarden-Essen-2 ($p = .11$) Bellman-Budge-1 ($p =.80$) Bellman-Budge-1 ($p = .007$) Sapp Simple-1 ($p = .45$) Sapp Simple-2 ($p = .31$) PKS-1 ($p = .02$) PKS-2 ($p = .40$)	Temperley-1 ($p < .001$)* Temperley-2 ($p < .001$)*
Minor	Temperley-1 ($p < .001$)* Temperley-2 ($p < .001$)* Krumhansl-Kessler-1 ($p < .001$)* Krumhansl-Kessler-2 ($p < .001$)* PKS-1 ($p = .02$)*	Aarden-Essen-1 ($p = .40$) Aarden-Essen-2 ($p = .04$) Bellman-Budge-1 ($p = .08$) Bellman-Budge-2 ($p = .04$) Sapp Simple-1 ($p = .22$) Sapp Simple-2 ($p = .17$) PKS-2 ($p = .04$)	

Note: "1" models indicate those that were given only the first and last eight measures (similar to the AS model) while "2" models were given the entire piece. After employing a Bonferroni correction, significance was set at $p = .005$.

between the actual tonic and the assigned tonic. For the proposed algorithm, the errors in key assignments for pieces in the major key tend to favor the relative minor, the parallel minor, or the dominant. When analyzing pieces in a minor key, errors tended to favor the subdominant, the dominant, and the parallel major. See Table 4 for an outline of these issues. Note that, as our number of errors is quite small, it is difficult to see any statistically significant preference for one type of error over the others.

TABLE 4. *Errors When Using the Proposed Euclidean Distance Algorithm.*

Relationship Between Incorrect Key and Correct Key	Major Key	Minor Key
Dominant	3	5
Parallel key	4	4
Relative key	7	1
Subdominant	2	6
Other relationship	1	0

Discussion

Our proposed model performed better than both the Krumhansl-Schmuckler and the Temperley models when looking at both the major and minor modes together. Additionally, our model had a higher number of correct identifications than the Aarden-Essen and the Bellman-Budge model for both modes, although the difference was not significant. However, our proposed model was significantly more accurate for pieces in the minor mode than the Bellman-Budge model. The proposed model had more correct identifications in the minor mode than either the Aarden-Essen model or the simple weights that Sapp suggested, but this difference was also not significant. As described above, some of the advantage of the proposed model may come from the use of Euclidean distance as a measure rather than correlation. This is suggested by the comparison of the proposed model against the weights of the proposed model with the correlational algorithm, in which the Euclidean distance algorithm had more correct identifications.

Beyond the advantage arising from using Euclidean distance, there may be other reasons why our model performed well. First, it could be that our model performed better on the reserve dataset pieces because the model was trained on a set of pieces that were

stylistically very similar. As the original database was simply split in half, creating one training dataset and one reserve dataset, the distributions of scale degrees used could be isomorphic between the two sets in a stronger relationship than would exist in randomly sampled pieces. This advantage would not hold for the Aarden-Essen model (which used distributions from folk songs), the Temperley model (which used distributions from excerpts in a textbook), the Sapp model (which used simple weights based on traditional concepts of scale degree hierarchy), or the Krumhansl-Schmuckler model (which used experimentally derived probe-tone distributions). However, it could be argued that the data used to train the current model were more representative of the population of pieces in the classical repertoire, and therefore could be more appropriate for this specific type of application. This may explain why the Bellman-Budge model, which also took its distributions from classical music, also performed well.

Another reason our algorithm may have performed well is the fact that the model was only given the first and last eight measures or pieces. This advantage can be seen in that most of the key-finding models were more accurate on just the first and last eight measures than on the entire piece. On the one hand, it might seem like an unfair advantage to selectively analyze the excerpts most likely to adhere to tonal gestures (setting aside the role of off-tonic openings, such as in the works of Schumann and Chopin, among others; see Noden-Skinner, 1984; Rothstein, 1994). By ignoring the more ambiguous tonal sections, however, the algorithm is less likely to be confused by modulatory sections, transitions, and developments. While it could be argued that this provides a rather monochromatic view of tonality in the music, this approach is able to provide a clear, singular answer that might be utilized in computational research.

Of particular interest is the lack of overlap between the proposed algorithm and the other models using correlation. For example, the proposed model assigned the correct key to 458 of 492 pieces, while the Aarden-Essen model assigned the correct key to 451 pieces. Although the proposed model only correctly identified nine more pieces than the Aarden-Essen model, there were 41 pieces in which one of the two models was correct while the other was not. In fact, of the 492 pieces, 475 of them had correct key assignment by one of the two models (96.5%). This lack of overlap in correct assignments may be an artifact of the largely different approach that each model takes—namely, using a Euclidean distance measure and a correlation measure. If appropriate criteria could be determined to decide when to use the result from one model over the other, it might therefore be theoretically possible to design a heuristic involving both models that would approach this level of accuracy.

One such possible criterion in determining the reliability of a model's key assignment could be in assessing the confidence with which that model makes its assignment. In other words, although each model will assign the key that best matches the distribution in the piece analyzed, there are usually other possible keys that could be represented by the same distribution. A clear example of this is the tendency, mentioned above, of many models to incorrectly assign the key of the dominant, which closely shares the pitch class content and frequency of the tonic key. The more distance there is between the assigned key and the next closest key, the more confident one can be that the model is correct.

In the case of the proposed Euclidean distance algorithm, the confidence will be a ratio of the distance measure for the best key to the second best key, subtracted from 1. For example, if the best key had a Euclidean distance of .1 and the second best key had a distance of 1.0, the confidence would be $1 - (.1/1) = .9$, or a confidence rating of 90%. In a correlation algorithm, Huron (1995) has found that a confidence measure can be calculated by subtracting the correlation value for the second best key from the correlation value of the best key and multiplying by 3 (an empirically determined scaling factor). For example, if the best key had a correlation of .90 and the second best key had a correlation of .70, the confidence would be $(.90 - .70) * 3 = .60$, or a confidence rating of 60%. In this equation, if the confidence value exceeds 100%, a value of 100% is taken as the confidence level.

Using these confidence values, a meta-algorithm has been constructed in which both the Aarden-Essen and the proposed models are permitted to run on a piece and a decision is made between them based on the confidence levels for each model. The following algorithm is certainly to be understood as a post-hoc attempt to increase the accuracy of key-assignments as much as possible. In other words, various combinations of decisions were repeatedly attempted and tweaked to get the best result. For this reason, this meta-algorithm should be viewed cautiously, as the generalizability of it to other datasets has not been determined. It could be the case that the meta-algorithm has been overfitted to the data in our reserve dataset.

The algorithm can be considered a multi-level conditional statement, represented below:

1. If both algorithms assign the same key, take that key to be the key of the work.

2. If the key-assignments are different, then if the proposed algorithm has a confidence of 25% or higher, take that key to be the key of the work.
3. If not, but if the Aarden-Essen model has a confidence of 25% or higher, take that key to be the key of the work.
4. If the algorithms do not assign the same key, and neither algorithm has a confidence of 25% or higher, then determine if the final note in the bass matches the tonic of either key assignment, and if so, take that key to be the key of the work.
5. Should these steps all fail, take the key of the algorithm that has the higher confidence level.

Using this meta-algorithm, we reanalyzed the works in the reserve dataset. Of major mode works, the meta-algorithm correctly identified 297 of 313 pieces, or 94.9%. Of minor mode works, it correctly identified 171 of 179 pieces, or 95.5%. Overall, 468 of 492 works were correctly assigned keys, or 95.1%. For pieces in the major mode, the meta-algorithm performed significantly better than the Krumhansl-Schmuckler or Aarden-Essen models, although it did not perform significantly better than the proposed model, the Bellman-Budge, the Sapp, or the Temperley models. For the minor mode, however, the meta-algorithm is significantly better than all of the existing models, including the algorithm proposed in this paper. Overall, for both the major and minor modes, the results are significantly better than every other model. Perhaps most importantly, the meta-algorithm described above has consistently high ratings for both the major and minor modes.

Given the post-hoc nature of this meta-algorithm, future research is required to determine whether the proposed method of combining both the Aarden-Essen and the currently proposed model is reliable on the broader population of common-practice pieces of music. We plan on testing this on a future project using large databases acquired from the Internet. As there is a need for a key to be asserted from the outset in order for a number of computational tools to be used (such as the analysis of harmonic and melodic intervals using the Humdrum Toolkit or Music21), an improved key-finding method would prove to be beneficial for such research.

Author Note

We are grateful to David Huron, who first introduced us to the notion of Euclidean distance as a possibility in key finding, and modified the original version of the Humdrum "key" command to include a Euclidean distance option.

Correspondence concerning this article should be addressed to Joshua Albrecht, Music Department, The University of Mary Hardin-Baylor, 900 College Street, Belton, TX 76513. E-mail: jalbrecht@umhb.edu

References

Aarden, B. (2003) *Dynamic melodic expectancy* (Unpublished doctoral dissertation). Ohio State University, Columbus, OH.

Albrecht, J., & Huron, D. (in press). A statistical approach to tracing the historical development of major and minor pitch distributions, 1400-1750. *Music Perception*.

Altman, G. T. M., Dienes, Z., & Goode, A. (1995). Modality independence of implicitly learned grammatical knowledge. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 21*, 899-912.

Bellman, H. (2005). About the determination of key of a musical excerpt. In *Proceedings of Computer Music Modeling and Retrieval* (pp. 187-203). Springer: Pisa, Italy.

Bigand, E., Perruchet, P., & Boyer, M. (1998). Implicit learning of an artificial grammar of musical timbres. *Cahiers De Psychologie Cognitive – Current Psychology of Cognition, 17*, 577-600.

Budge, H. (1943). *A study of chord frequencies based on the music of representative composers of the eighteenth and nineteenth centuries* (Unpublished doctoral dissertation). Columbia University Teacher's College, New York.

CCARH. (2001). *The MuseData database* (http://www.musedata.org). Stanford, CA: Center for Computer Assisted Research in the Humanities.

Chew, E. (2000). *Towards a mathematical model of tonality* (Unpublished doctoral dissertation). Massachusetts Institute of Technology, Cambridge, MA.

Dienes, Z., & Longuet-Higgins, C. (2004). Can musical transformations be implicitly learned? *Cognitive Science, 28*, 531-558.

Huron, D. (1995). *The Humdrum toolkit: Reference manual.* Menlo Park, CA: Center for Computer Assisted Research in the Humanities.

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*, 334-368.

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

Krumhansl, C. L. (2000). Tonality induction: A statistical approach applied cross-culturally. *Music Perception*, *17*, 461-479.

Longuet-Higgins, H. C. (1976). Perception of melodies. *Nature*, *263*, 646-653.

Longuet-Higgins, H. C., (1979). Perception of music. *Proceedings of the Royal Society of London*, *205*, 307-322.

Longuet-Higgins, H. C., & Steedman M. J. (1971). On interpreting Bach. *Machine Intelligence*, *6*, 221-242.

Loui, P., Wessel, D. L., & Hudson Kam, C. L. (2010). Humans rapidly learn grammatical structure in a new musical scale. *Music Perception, 27*, 377-388.

Noden-Skinner, C. (1984) Tonal ambiguity in the opening measures of selected works by Chopin, *College Music Symposium, 24*, 28-34.

Rothstein, W. (1994). Ambiguity in the themes of Chopin's first, second, and fourth ballades. *Integral 8*, *1*, 1-50.

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*, 27-52.

Salzberg, S. (1997). On comparing classifiers: Pitfalls to avoid and a recommended approach. *Data Mining and Knowledge Discovery, 1*, 317-328.

Sapp, C. (2011). *Computational methods for the analysis of musical structure* (Unpublished doctoral dissertation). Stanford University, Stanford, CA.

Schaffrath, H. (1995). *The Essen folksong collection*. (D. Huron, Ed.). Stanford, CA: Center for Computer Assisted Research in the Humanities.

Steedman M. J. (1984). A generative grammar for jazz chord sequences. *Music Perception, 2*, 52-77.

Temperley, D. (1999). What's key for key? The Krumhansl-Schmuckler key-finding algorithm reconsidered. *Music Perception*, *17*, 65-100.

Temperley, D. (2007). *Music and probability*. Cambridge, MA: MIT Press.

Temperley, D. (2008). A probabilistic model of melody perception. *Cognitive Science*, *32*, 418-444.

VanHandel, L., & Callahan, M. (2012). The role of phrase location in key identification by pitch-class distribution. In E. Cambouropoulos, C. Tsougras, P. Mavromatis, K. Pastiadis (Eds.), *Proceedings of the 12th International Conference on Music Perception and Cognition and the 8th Triennial Conference of the European Society for the Cognitive Sciences of Music* (pp. 1069-1073). ICMPC: Thessaloniki, Greece.

Appendix

Appendix A. The weightings for the major and minor scale-degree distributions derived from the training set of 312 major mode and 178 minor mode works. Scale degree 0 is taken to be the tonic pitch, etc.

	0	1	2	3	4	5	6	7	8	9	10	11
Major	0.238	0.006	0.111	0.006	0.137	0.094	0.016	0.214	0.009	0.080	0.008	0.081
Minor	0.220	0.006	0.104	0.123	0.019	0.103	0.012	0.214	0.062	0.022	0.061	0.052

Appendix B. The weightings for the major and minor scale-degree distributions from the entire database of 625 major mode and 357 minor mode works. Scale degree 0 is taken to be the tonic pitch, etc.

	0	1	2	3	4	5	6	7	8	9	10	11
Major	0.238	0.006	0.111	0.006	0.137	0.094	0.016	0.214	0.009	0.080	0.008	0.081
Minor	0.220	0.006	0.104	0.123	0.019	0.103	0.012	0.214	0.062	0.022	0.061	0.052