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, July 23-28, 2012, Thessaloniki, Greece
Cambouropoulos E., Tsougras C., Mavromatis P., Pastiadis K. (Editors)

The Role of Phrase Location in Key Identification by Pitch Class Distribution

Leigh VanHandel,[*1] Michael Callahan,[*2]

[*] *College of Music, Michigan State University, USA*
[1]lvh@msu.edu, [2]mrc@msu.edu

ABSTRACT

This study extends prior research by investigating how pitch distribution differs at beginnings, middles, and ends of phrases, and by determining whether these differences impact key-finding. In the corpus of Haydn and Mozart string quartets used in Temperley and Marvin (2008), many phrases modulate to either the dominant or the relative major; this results in an overrepresentation of raised scale degree 4, as the leading tone to the dominant, and of lowered scale degree 7, as the dominant of III. The overrepresentation of these two scale degrees in the overall distribution may have contributed to the difficulties that Temperley and Marvin's subjects had with key finding. This study corrects the problem of overrepresentation by limiting the corpus to non-modulating phrases. A behavioral study indicates that subjects have better success with the distributional view of key finding with this modified distribution of pitches. In addition, melodies were constructed using independent pitch distributions for the beginnings, middles, and ends of phrases. Preliminary results show that subjects improve at identifying the key of a melody when the pitch distributions within its beginning, middle, and end follow those of the three sections of the original phrases.

I. INTRODUCTION AND BACKGROUND

In their 2008 article, Temperley and Marvin distinguish between *distributional* and *structural* views of key identification. Their study thoroughly discusses previous work on key-finding, including both behavioral studies and mathematical or computer-based models of key-finding. Their behavioral experiment tests whether listeners are capable of using a distributional strategy to identify key when presented with stochastically-generated melodies using a pitch distribution derived from a corpus of string quartet beginnings.

This study refines the methodology used in Temperley and Marvin (2008) and discusses the resulting changes to the pitch distribution, as well as a behavioral experiment designed to determine whether the revised pitch distributions enable participants to identify key with greater accuracy.

II. CORPUS MODIFICATION

Temperley and Marvin generate melodies for their key-finding experiment according to probabilistic distributions that track the relative commonness of each scale degree in both major and minor keys. To establish these scale-degree profiles, they rely on a corpus that includes measures 1-8 of each movement in the complete string quartets of Haydn and Mozart (310 movements total). The implicit assumption is that the opening eight measures of a movement are a good representation of that movement's key, and that these opening passages in the aggregate constitute a reliable basis for calculating the frequency of each scale degree in major and in minor. In fact, however, 58 of these movements (19%) modulate or begin to modulate to another key within the first eight measures; thus, not all of the music in these phrases is a good representation of the movement's tonic key. As a result, Temperley and Marvin's statistical profile of major and minor keys is skewed slightly because they assess each of these 58 passages in terms of a key other than the one in which some of the music actually functions.

The present study approaches the same corpus more selectively by refining it according to two criteria. First, all movements that modulate or begin to modulate prior to the end of the eighth measure have been eliminated.[1] Secondly, we hypothesize that the location within a phrase (i.e. beginning, middle, or end) is relevant to listeners' determination of key. In order to establish independent distributional scale-degree profiles for phrase beginnings, phrase middles, and phrase endings, we must define these regions consistently across all phrases in the corpus. Therefore, only those eight-measure openings that consist of a complete phrase with a half or authentic cadence in measure eight are included in the corpus. The following discussion elaborates on the rationale and the consequences of each of the two selection criteria, focusing only on major-key movements.[2]

A. Modulating Phrases.

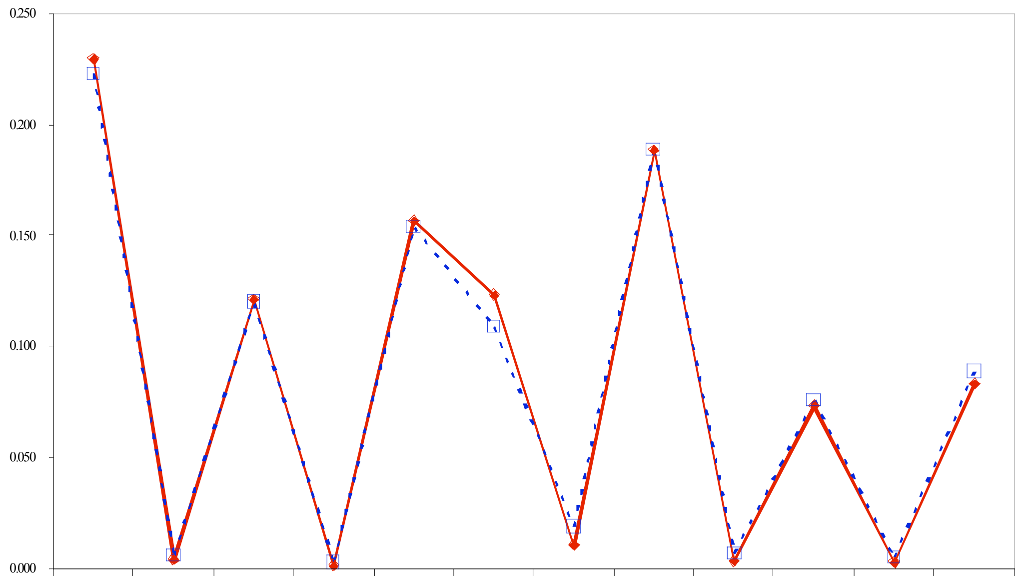
In the 58 movements that modulate prior to the end of measure 8, the music between the onset of the modulation and the end of measure 8 is *in* one key, but it is *counted* by Temperley and Marvin as being in a different key. When these early-to-modulate movements are included in the corpus, the presence of music outside the movement's tonic key skews the scale-degree frequencies in predictable ways. Of the 58 modulating openings, the vast majority (47) are major-key movements that modulate to the dominant key (e.g. from C major to G major).[3] The two keys differ by one diatonic pitch: ^4 of the original key does not belong to the dominant key, so its frequency wanes as the modulation takes place while that of ^#4 (i.e. ^7 of the dominant key) increases. In a corpus that includes major key movements

---

[1] A phrase was disqualified on the basis of modulation if it reached a cadence in a key other than the initial tonic. Phrases that include transient tonicizations of non-tonic scale degrees remain in the corpus as long as they end in a half or authentic cadence in the original key.

[2] The overwhelming majority of movements in the Haydn and Mozart string quartets are in major keys. Given the small number of minor-key movements in the total corpus, and the even smaller number that remained eligible after our two selection criteria were applied, the number of minor-key data points is too small to be statistically valid for the present experiment.

[3] Of the other eleven movements that modulate within the first eight measures, only one is a major-key movement; unusually, it reaches a half cadence in the submediant key in m. 8.



	1	#1/b2	2	#2/b3	3	4	#4/b5	5	#5/b6	6	#6/b7	7
Revised distribution	0.230	0.004	0.121	0.002	0.157	0.123	0.011	0.189	0.003	0.074	0.003	0.083
Temperley & Marvin (2008)	0.223	0.006	0.12	0.003	0.154	0.109	0.019	0.189	0.007	0.076	0.005	0.089

Figure 1. Revised major pitch-class distribution vs. major distribution from Temperley & Marvin (2008).

modulate to the dominant, we would expect ^#4 to be overrepresented in the major-key distribution and ^4 to be underrepresented. Figure 1 shows exactly this discrepancy, comparing the major-key distribution in our revised corpus to that of Temperley and Marvin's less selective corpus; these two distributions are significantly different (χ^2=212.07, df = 6, p < .0001).[4] Among the differences in relative frequency of scale degrees across the two distributions, ^4 is 1.4% more common in our corpus than in theirs (0.123 vs. 0.109) and ^#4 is 0.8% less common in ours (0.011 vs. 0.019). In addition, our higher frequency of ^1 and lower frequency of ^7 indicates a further reduction of emphasis on the dominant due to removal of modulating phrases.

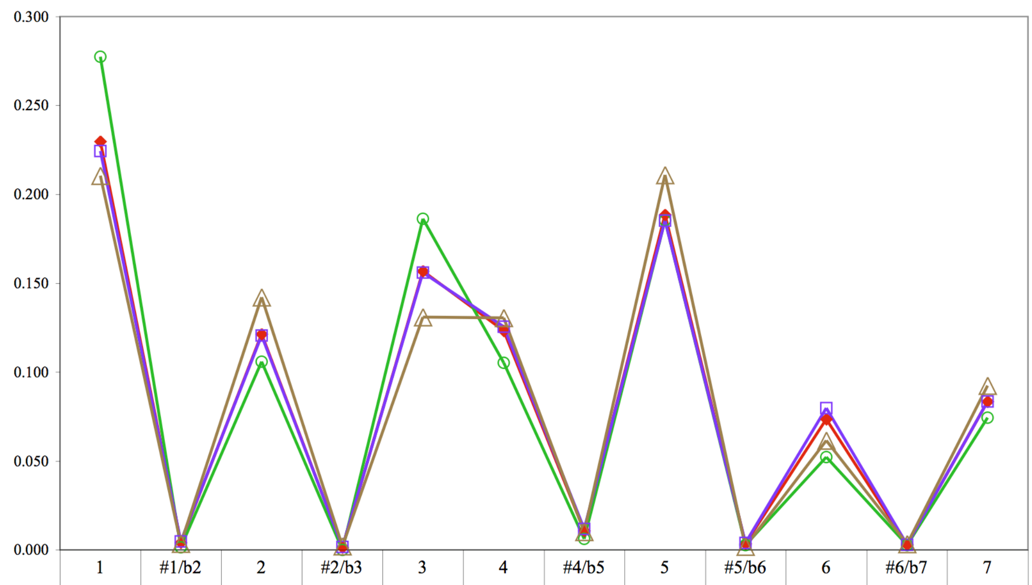
B. Phrase Beginnings, Middles, and Ends.

We hypothesize that the different scale-degree distributions within phrase beginnings, middles, and ends factor into listeners' determination of key. At the beginning of a phrase, the clarity of the key and the prevalence of stable scale degrees may help to contextualize the more mixed mid-phrase distribution. Listeners accustomed to phrases that begin with high concentrations of tonic scale degrees may be biased toward perceiving phrase beginnings in this way. Moreover, a listener's expectation to hear a cadence at the end of a phrase may affect how the distribution of scale degrees is interpreted at the end of a melody. That is, it may be easier to decide upon a key if the melody ends with scale degrees compatible with a half or authentic cadence (e.g. ^1, ^2, ^3, ^5, or ^7) than if it ends with an undifferentiated continuation of the same distribution that it contained all along. In particular, scale-degrees ^4 and ^6 are incompatible with the arrival on either a tonic triad or a dominant triad, and may undermine the determination of key even in spite of a preceding distribution that was unambiguous.

For simplicity, we define the beginning of a phrase as its first quarter and the end of a phrase as its last quarter; thus, for an eight-measure phrase, measures 1-2 are considered the beginning, measures 3-6 the middle, and measures 7-8 the end. In order to create three different scale-degree distributions (i.e. for mm. 1-2, mm. 3-6, and mm. 7-8), we restrict our corpus to eight-measure openings that constitute complete phrases ending in half or authentic cadences. Of Temperley and Marvin's original 310-movement corpus, 175 movements (133 of Haydn, 42 of Mozart) were disqualified on the basis of phrase length. Many began with longer phrases that did not reach a cadence until after m. 8, while some reached the end of their first phrase prior to m. 8 and were underway in the second phrase by the time the eight-measure window elapsed.

Figure 2 compares the scale-degree distribution of our entire revised corpus (mm. 1-8) to that of just our phrase beginnings (mm. 1-2), phrase middles (mm. 3-6) and phrase endings (mm. 7-8). Each of these four distributions is significantly different from each of the other three.

[4] The revised corpus was created in *Humdrum,* an open-source software package for music research, which was also used for calculating the distributions.



	1	#1/b2	2	#2/b3	3	4	#4/b5	5	#5/b6	6	#6/b7	7
Revised overall distribution	0.230	0.004	0.121	0.002	0.157	0.123	0.011	0.189	0.003	0.074	0.003	0.083
Distribution for mm. 1-2 (beginning)	0.277	0.001	0.106	0.000	0.186	0.105	0.006	0.186	0.003	0.052	0.003	0.074
Distribution for mm. 3-6 (middle)	0.224	0.005	0.121	0.002	0.156	0.125	0.012	0.185	0.004	0.080	0.003	0.084
Distribution for mm. 7-8 (end)	0.211	0.004	0.142	0.002	0.131	0.131	0.010	0.211	0.002	0.062	0.003	0.092

Figure 2. Revised overall distribution vs. distribution for mm. 1-2 (beginning), mm. 3-6 (middle), and mm. 7-8 (end).

These four pitch-class distributions exhibit several clear differences that correspond intuitively to their location in the phrase. Phrases generally begin with high concentrations of tonic triads, which explains the higher incidence of the two uniquely tonic scale degrees in the *beginning* distribution: ^1 and ^3 are more frequent in mm. 1-2 than overall. Correspondingly, the non-tonic scale degrees are less frequent in mm. 1-2 than overall, particularly ^6, which is also absent from the dominant and dominant-seventh harmonies that typically prolong tonic at the outset of a phrase.

In the *ending* distribution, there is a higher representation of the dominant seventh chord (^5, ^7, ^2, and ^4) than in the aggregate distribution, reflecting the importance of the dominant as both the penultimate harmony of authentic cadences and the last one of half cadences. Our revised corpus still includes phrases that end with half cadences in the original key, so the uniquely tonic scale degrees of ^1 and ^3 are not as strongly represented as in the *beginning* distribution. However, as expected, the cadentially absent ^6 appears less frequently.

We hypothesize that a heightened sensitivity to the changes in scale-degree distribution across the span of a phrase may assist listeners in identifying the key. Our contention is that melodies generated with attention to the distinct roles and scale-degree distributions of *beginning*, *middle*, and *end* will be parsed more successfully by participants than those generated by the single, global scale-degree distribution utilized in Temperley and Marvin's study.

III. EXPERIMENT

We report here on the preliminary results from a behavioral experiment testing melodies constructed from both the revised overall pitch distribution and melodies constructed from the pitch distribution based on the beginning, middle, and ends of phrases.

A. Method

1) Participants. To date, we have data from nineteen participants (15 male, 4 female) with an average age of 21 years (SD=2.65). Participants were undergraduates (17) or graduate students (2) in the College of Music at Michigan State University. Participants began musical study at a mean age of 7.47 years (SD=3.22) and, at the time of the experiment, had completed an average of 3.15 years (SD=1.97) of collegiate music theory. Seventeen participants identified as Caucasian, one as Hispanic, and one as Asian or Asian-American. None self-reported as having absolute pitch, although two self-reported as "not sure."

Participants were divided at random into groups that completed either the *unweighted* or *weighted* condition described below; there was no significant difference in current age, age at which they began music study, or years of college-level theory between groups.

2) Equipment. The experiment was administered in a small room. The melodies were presented using a randomized iTunes playlist on an iMac computer. Participants listened to stimuli through Sony MDR–7502 headphones, and determined note names on a Kawai K–1 keyboard using the default piano sound.

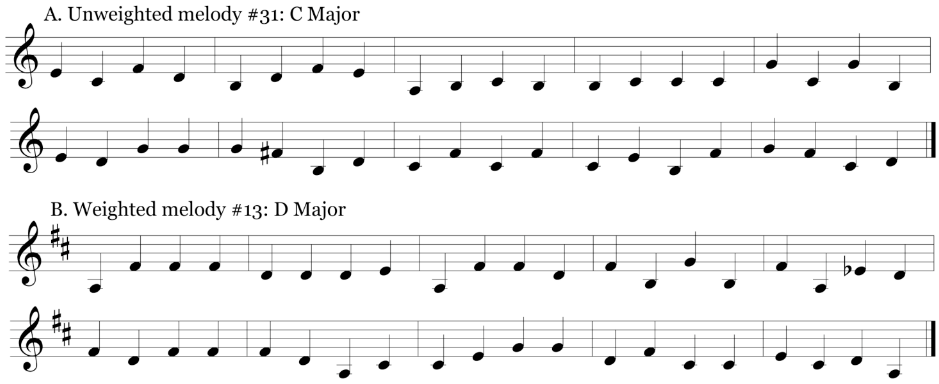


Figure 3. Examples of melodic stimuli.

Participants had the opportunity to adjust the loudness of the stimuli and the keyboard to a comfortable listening level before and during the experiment.

3) Stimuli. Two sets of randomly generated major-mode melodies were created using a method similar to that in Temperley and Marvin (2008). Each set contained 40 melodies, each 40 notes in length and randomly selected from all 12 major keys. One set of melodies used our revised overall pitch distribution as a generating profile (the *unweighted* melodies); the other (*weighted*) set used the three pitch distributions for phrase beginnings, middles, and ends, with the first ten notes (mm. 1-2) drawn from the beginning distribution, the last ten notes (mm. 7-8) drawn from the end distribution, and the remaining twenty notes (mm. 3-6) drawn from the middle distribution.

As in the Temperley & Marvin (2008) study, each melody was assigned a randomly chosen range of 12 semitones within a range of A3 to G5; this ensured that the tonic of the key was not always the lowest pitch. The melodies were isochronous (400 ms per note) and were played using the QuickTime 7.7 piano timbre. Figure 3 reproduces two of the melodies; melody A uses the unweighted revised overall distribution and has the generating key of C Major, and melody B uses the weighted distribution and has the generating key of D Major.

4) Procedure. Participants were randomly assigned to hear either the unweighted melodies or the weighted ones within a single session. Of the forty melodies generated for each condition, four from each set were selected quasi-randomly to be used as sample stimuli, and the remaining 36 were used as stimuli.[5] The melodies were randomized for each participant by shuffling an iTunes playlist, and each melody was started by the experimenter. The four practice melodies allowed participants to adjust volumes and ask questions, but no feedback was given on performance. The subjects then heard the 36 melodies as described above. Participants were instructed to determine tonic, singing and/or humming if they wished, and to use the keyboard to identify the key name. They then communicated their answer verbally to the experimenter, who recorded their response and began the next melody.

B. Results

As in the Temperley and Marvin (2008) study, our primary interest was to determine whether participants' key judgments matched the keys used to generate the melodies. Each participant judged the keys of 36 melodies, either unweighted (our overall distribution) or weighted (beginning-middle-end).

For the unweighted melodies, 71% (SE=.25) of participant judgments matched the generating key. Performance was much better than an assumed chance performance of 8.3% (since there were 12 possible major keys), $t(9)=14.9$, $p < .0001$; performance was also significantly better than the 51% accuracy in Temperley and Marvin, $t(9)=4.7$, $p = .001$.

For the weighted melodies, 77% (SE=.21) of participant judgments matched the generating key. Again, performance was much better than chance, at $t(8)=17.9$, $p < .0001$, and better than performance than in the Temperley and Marvin study at $t(8)=6.8$, $p = .0001$.

As in the Temperley and Marvin study, we examined the agreement between participants by finding the key that was chosen by the largest number of participants for each melody (Temperley and Marvin's "most popular key", or MPK). The MPK judgments matched the generating keys in 34 out of the 36 unweighted melodies, and 33 out of 36 of the weighted melodies (compared to 50 out of 60 in Temperley and Marvin). The MPK judgments accounted for 72.5% of the 360 judgments for the unweighted melodies, and for 80.5% of the 324 judgments for the weighted melodies, compared to an overall 56.1% in Temperley and Marvin.

C. Discussion

Participants were more successful in identifying the keys of melodies generated according to our overall unweighted distribution than of those generated according to Temperley and Marvin's distribution. The primary difference between their distribution and our overall distribution is the removal of eight-measure openings that either modulate or do not

[5] The four melodies selected as sample stimuli were in major keys that ended up being over-represented by the random generating process.

consist of a single phrase with a cadence in measure 8. It appears that limiting the corpus to music that functions entirely within the key of reference helped participants significantly with the key identification task, since they were able to identify 71% of the generating keys correctly in the unweighted overall distribution, compared to 51% in the Temperley and Marvin study.

Our weighted distribution also seems to have helped participants; participants who heard the weighted melodies were slightly better (77%) at correctly identifying the generating key than those who heard the unweighted melodies (71%).

In addition to identifying the generating key more often, participants agreed with each other to a greater extent in our study than in Temperley and Marvin's, with MPK judgments of 72.5% (unweighted) and 80.5% (weighted) compared to 56.1% in the prior study. This indicates that, in the case of the five melodies for which the MPK did not match the generating key, our participants more consistently identified the same incorrect tonic. Table 1 shows the five melodies where the MPK did not match the generating key. In each case, the MPK was either a fifth higher or lower than the generating key, and therefore one of the two closest keys in terms of shared pitch-class content. Moreover, the generating key was the second most popular key in each case.

Table 1. Generating key and most popular key for the five melodies with a mismatch.

Melody	Generating key	Most popular key	Second most popular key
Unweighted Melody #17	A	D	E/A (tie)
Unweighted Melody #35	A	D	A
Weighted Melody #8	Ab	Eb	Bb/Ab (tie)
Weighted Melody #14	A	D	A
Weighted Melody #23	D	G	D

Temperley and Marvin were surprised by the relatively low agreement among participants in their experiment, which, to them, "casts serious doubt on the distributional view of key perception" (p. 209). Our revised overall distribution, which more accurately reflects the pitch-class distribution of Haydn and Mozart string quartet openings that function entirely within their tonic keys, appears to have helped participants identify keys both more accurately and more consistently. It may be that the participants in Temperley and Marvin's study struggled in part because the corpus included passages that did not accurately represent the putative major and minor keys of reference.

Temperley and Marvin consider the relationship between distributional and structural key-finding methods and models; they conclude that "structural cues of some kind—cues relating to the ordering and temporal arrangement of pitches—play a role in key perception" (p. 210). While our weighted distribution is not structural in the same sense as the structural cues discussed in Vos (1999), Butler (1989), Brown, Butler, & Jones (1994), or Farbood et al. (2010), the significant differences between distributions at the beginning, middle, and end of eight-measure phrases may permit musically informed listeners to develop different expectations for pitch-class distribution for different locations in a phrase. In other words, phrase location may itself be a structural cue for listeners to consider when key-finding.

Based on the results of our weighted melodies, which improved upon both Temperley and Marvin's results and our own unweighted melodies, it may be that sensitivity to phrase location is an important factor in distributional key-finding behavior.

D. Future directions

Temperley and Marvin (2008) also study mathematical models of key-finding and compare the results of those models to the results from their behavioral study. It remains to be seen how our revised overall distribution and our beginning-middle-end distribution would respond to the various models used in their study.

IV. CONCLUSION

Temperley and Marvin (2008) investigate the role of distributional methods of key-finding both in comparison to mathematical models and in a behavioral study. This study offers a refinement to the corpus that leads to better success in the pilot behavioral study, and also introduces a hybrid distributional/structural model based on phrase location that improves accuracy in participants' key-finding. The role of distributional and structural methods for key-finding is one that deserves continued study.

ACKNOWLEDGMENT

We thank Phillip Sink, Joseph Vanderstel, and Daniel Alt for their valuable assistance in this project. We also thank Devin McAuley and members of the Timing, Attention and Perception lab at Michigan State University for their suggestions and assistance.

REFERENCES

Brown, H., Butler, D., & Jones, M. R. (1994). Musical and temporal influences on key discovery. *Music Perception, 11,* 371–407.

Butler, D. (1989). Describing the perception of tonality in music: A critique of the tonal hierarchy theory and a proposal for a theory of intervallic rivalry. *Music Perception, 6,* 219–242.

Farbood, M. M., Marcus, G., Mavromatis, P., Poeppel, D. (2010). The effect of structure and rate variation on key-finding. In *Proceedings of the 11th International Conference of Music Perception and Cognition*, pp. 36-40. Seattle, Washington, USA.

Huron, D. (1999). *Music research using Humdrum: A user's guide.* Stanford, California: Center for Computer Assisted Research in the Humanities.

Temperley, D., & Marvin, E. W. (2008). Pitch-class distribution and the identification of key. *Music Perception, 25/3,* 193–212.

Vos, P. G. (1999). Key implications of ascending fourth and descending fifth openings. *Psychology of Music, 27,* 4–18.