A Probabilistic Model of Melody Perception

David Temperley
Eastman School of Music
26 Gibbs St.
Rochester, NY 14604
dtemperley@esm.rochester.edu

Abstract

This study presents a probabilistic model of melody perception, which infers the key of a melody and also judges the probability of the melody itself. (A "melody" is defined here as a sequence of pitches, without rhythmic information.) The model uses Bayesian reasoning. A generative probabilistic model is proposed, based on three principles: 1) melodies tend to remain within a narrow pitch range; 2) note-to-note intervals within a melody tend to be small; 3) notes tend to conform to a distribution (or "key-profile") that depends on the key. The model is tested in three ways: on a key-finding task, on a melodic expectation task, and on an error-detection task.

Keywords: Music cognition, key identification, probabilistic modeling, expectation, error detection

1. Introduction

In recent years, methods of Bayesian probabilistic modeling have been widely used in a variety of areas in information processing and cognitive modeling, such as natural language processing, vision, and knowledge representation; they have also been applied to musical problems such as transcription [1] and metrical analysis [2, 3]. In general, we can frame the Bayesian approach as a way of recovering some kind of underlying *structure* from some kind of *surface* representation. Bayesian logic tells us that

$$P(\text{structure} \mid \text{surface}) \propto P(\text{surface} \mid \text{structure})P(\text{structure}) \quad (1)$$

Thus we may determine the most probable structure given a surface if we know, for all structures, the probability of the surface given the structure and the prior probability of the structure. A further consequence of the Bayesian approach is the possibility of calculating the probability of the surface pattern itself:

$$P(\text{surface}) = \sum_{\text{structure}} P(\text{structure}, \text{surface}) = \sum_{\text{structure}} P(\text{surface} \mid \text{structure})\, P(\text{structure}) \quad (2)$$

It can be seen, then, that calculating the most probable structure given a surface and calculating the probability of a surface are very closely related problems.

In the current case, we define a surface as a sequence of pitches (without rhythmic information); the structure is a key. The problem, then, is to determine the most probable key given a note sequence—this is the familiar and well-studied "key-finding" problem. The probability of a surface is then the probability of a sequence of pitches. I will argue that this concept of "surface probability" is of relevance to a variety of processes in music information processing, such as expectation, error detection, and transcription. (More information about the model presented here can be found in [4]).

Like most Bayesian models, our approach begins with a generative model, which generates a key and then a series of pitches. The model also considers two other important factors in melodic construction, range and pitch proximity. To set the model's parameters, we use a corpus of over 6,000 computationally-encoded European folk melodies, the Essen Folksong Collection [5]. We will then show how this model can be used in Bayesian fashion to perform key identification as well as "surface-level" processes such as expectation and error detection.

2. The Generative Model

We begin with a very basic question: What kind of pitch sequence makes a likely melody? Perhaps the first principle that comes to mind is that a melody tends to be confined to a fairly limited range of pitches. In the Essen corpus, the average range of a melody (from the highest to the lowest pitch, inclusive) is 13.6 semitones. We can model this situation in a generative way by first choosing a central pitch *c* for the melody; this is randomly chosen from a normal distribution, which we call the *central pitch profile*. We then create a second normal distribution centered around *c*, the *range profile*, which is used to actually generate the notes. A melody can then be constructed as a series of notes generated from the range profile.

A further important principle in melodic construction is that intervals between adjacent notes in a melody tend to be small [6]. In the Essen corpus, more than half of all melodic intervals are 2 semitones or less. We can approximate this distribution of melodic intervals with a *proximity profile*—a normal distribution centered around a given pitch, indicating the pitch probabilities for the following note. We then create a new distribution which is the product of the proximity profile and the range profile. In effect, this "range × proximity" profile favors melodies which maintain small note-to-note intervals, but also remain within a fairly narrow global range.

Our third and final principle of melodic construction is that melodies (at least in the Western tradition) tend to adhere to the scale of a particular key. We incorporate this using the concept of *key-profiles*. A key-profile is a twelve-valued vector, representing the stability or appropriateness of pitch-classes in relation to a key [8]. In this case, the key-profiles are based on the actual distribution of scale-degrees (pitch-classes in relation to the key) in the Essen corpus. We count the occurrences of each scale-degree in each song; we sum these counts over all songs (grouping major-key and minor-key songs separately), and express the totals as proportions. The resulting key-profiles are shown in figure 1. The profiles show that, for example, 18.4% of notes in major-key melodies are scale-degree 1. The profiles reflect conventional musical wisdom, in that pitches belonging to the major or minor scale of the key have higher values than other pitches, and pitches of the tonic chord (the 1, 3, and 5 degrees in major or the 1, b3, and 5 degrees in minor) have higher values than other scalar ones.

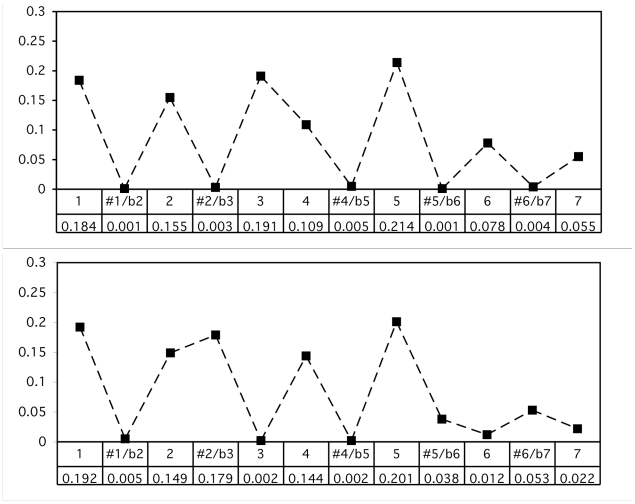

Figure 1. Key-profiles for major keys (above) and minor keys (below).

To combine all three of our principles together, we duplicate the key-profiles over many octaves; we can then multiply a key-profile together with the range and proximity profiles. We will call the resulting distribution the *RPK profile*. (To be interpretable as probabilities, the RPK profile values must be normalized to sum to 1.) In generating a melody, we must construct the RPK profile anew at each point, since the proximity profile depends on the previous pitch. (For the first note, since there is no previous pitch, we simply use the product of the range and key profiles.)

Figure 2 shows an RPK profile, assuming a key of C major, a central pitch of 68 (Ab4), and a previous note of C4. One can discern a roughly bell-shaped curve to this profile (though with smaller peaks and valleys). The proximity profile pulls the center of the curve towards C4, but the range profile pulls it towards Ab4; the result is that the actual center is in between the two. The key-profile gives higher values to pitches that are within the C major scale, thus accounting for the local peaks and valleys.

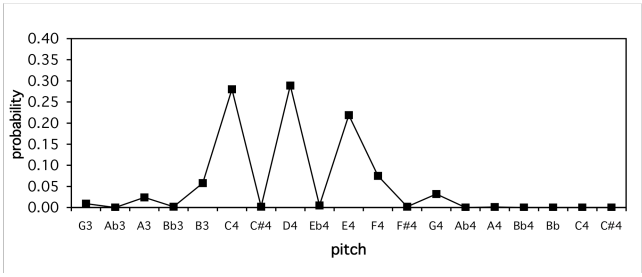

Figure 2. An RPK profile, assuming a key of C major, a central pitch of Ab4, and a previous pitch of C4.

The generative process thus operates by choosing a key and a central pitch, and then generating a series of pitches. (The process does not decide how many pitches to generate; this is assumed to be given.) The probability of a pitch occurring at any point is given by its RPK profile value: the normalized product of its range-profile value (given the central pitch), its proximity-profile value (given the previous pitch), and its key-profile value (given the chosen key). We express the joint probability of a pitch sequence of n notes with a key k and a central pitch c as follows:

$$P(\text{pitch sequence}, k, c) = P(k)\, P(c) \prod_n RPK_n \qquad (3)$$

where $P(k)$ is the probability of a key being chosen, $P(c)$ is the probability of a central pitch being chosen, and RPK_n are the RPK-profile values for the pitches of the melody given the key, central pitch, and previous pitch. We assume that all keys of the same mode are equal in prior probability, since most listeners—lacking "absolute pitch"—are incapable of identifying keys in absolute terms. However, we assign major keys a higher probability than minor keys, reflecting the higher proportion of major-key melodies in the Essen collection.

How do we calculate the overall probability of a pitch sequence? For the moment, let us think of the structure as the combination of a key and a central pitch; the surface is a sequence of pitches. From equations 2 and 3 above:

$$P(\text{pitch sequence}) = \sum_{k,\,c} (P(k)\,P(c) \prod_{n} RPK_n) \qquad (4)$$

This can be calculated quite easily by considering each (k, c) pair and calculating the joint probability of the pitch sequence with that pair.

3. Testing the Model on Key-Finding

We now consider how the generative process described above might be incorporated into a key-finding model. The task is simply to choose a single key for a given melody. (We do not allow the possibility of modulations—changes of key.) Using Bayesian logic, the most probable key given the melody will be the one maximizing $P(k,$ pitch sequence). Using equations 2 and 4 above:

$$P(k, \text{ pitch sequence}) = \sum_{c} (\text{pitch sequence}, k, c)$$

$$= \sum_{c} (P(k)\,P(c) \prod_{n} RPK_n) \qquad (5)$$

Thus the most probable key given a pitch sequence is the one maximizing this expression.

Our key-finding process thus proceeds as follows. For each key, we calculate the joint probability of the melody with that key and each central pitch, and sum this over all central pitches. The probability of a pitch at a given point in the melody depends only on its value in the RPK profile at that point; the RPK profile can be recreated at each note, just as it was in the generative process. We perform this process for all keys, and choose the key yielding the highest value; this is the most probable key given the melody.

To test the model's key-finding ability, we use a sample of 65 songs from the Essen collection. (This sample was not included in the corpus used for setting the model's parameters.) Each song in the corpus is annotated with a single key label; these labels provide a set of "correct" judgments against which the model can be evaluated. The model judged the key correctly for 57 of the 65 melodies (87.7%). By way of comparison, two other well-known key-finding algorithms were also tested on the corpus (using my own implementations). The model of Longuet-Higgins and Steedman [7] guessed the correct key on 46 out of 65 melodies, or 70.8% correct; the model of Krumhansl and Schmuckler [8] guessed the correct key on 49 out of 65, or 75.4% correct.

4. Expectation and Error Detection

It is well known that in listening to a melody, listeners form expectations as to what note will occur next. Melodic expectation has been the subject of a great deal of research in music psychology and music theory. Of particular interest here is an experimental study by Cuddy and Lunney [9]. In this study, subjects were played a context of two notes played in sequence (the "implicative interval"), followed by a third note (the "continuation tone"), and were asked to judge the third note given the first two on a scale of 1 ("extremely bad continuation") to 7 ("extremely good continuation"). Eight different contexts were used: ascending and descending major second, ascending and descending minor third, ascending and descending major sixth, and ascending and descending minor seventh. Each two-note context was followed by 25 different continuation tones, representing all tones within an octave above or below the second tone of the context (which was always either C4 or F#4). For each condition (context plus continuation tone), Cuddy and Lunney reported the average rating, thus yielding 200 data points in all.

We tested the current model's ability to predict melodic expectation, using Cuddy and Lunney's data. To do this, it was necessary to interpret their data probabilistically. Specifically, each rating was taken to indicate the log probability of the continuation tone given the previous two-tone context. Under the current model, the probability of a pitch p_n given a previous context ($p_0 \ldots p_{n-1}$) can be expressed as

$$P(p_n \mid p_0 \ldots p_{n-1}) = P(p_0 \ldots p_n) \,/\, P(p_0 \ldots p_{n-1}) \qquad (6)$$

where $P(p_0 \ldots p_n)$ is the overall probability of the context plus the continuation tone, and $P(p_0 \ldots p_{n-1})$ is the probability of just the context. An expression indicating the probability of a sequence of tones was given in equation 4 above; this can be used here to calculate both $P(p_0 \ldots p_{n-1})$ and $P(p_0 \ldots p_n)$. For example, given a context of (Bb4, C4) and a continuation tone of D4, the model's expectation judgment would be $\log(P(\text{Bb4, C4, D4}) \,/\, P(\text{Bb4, C4})) = -1.955$.

The model was run on the 200 test items in Cuddy and Lunney's data, and its outputs were compared with the experimental ratings for each item. Using the parameters gathered from the Essen Folksong Collection, the model yielded the correlation $r = 0.664$. It seemed reasonable, however, to adjust the parameters to achieve a better fit to the data. (This is analogous to what is done in most other tests of expectation models—such as those in [9] and [10]—in which multiple regression is used to fit a set of factors to the data in an optimal way.) With adjusted parameters, the model achieved a score of $r = .822$. This score is slightly better than that of Cuddy and Lunney's own model (.80), though not quite as good as that of Schellenberg's model [10] on the same data (.851).

Another kind of phenomenon that is illuminated by the current model could be described as "pitch error detection." It seems uncontroversial that most human listeners have some ability to detect errors—"wrong notes"—even in an unfamiliar melody. The ability of the current model to detect errors was tested using the 65-song Essen test set described above. The model was given the original melodies as well as randomly distorted versions of the same melodies; the question was whether it could reliably assign a higher probability to the correct versions

as opposed to the distorted versions. The deformed version of a melody was produced by randomly choosing one note and replacing it by a random pitch within the range of the melody (between the lowest and highest pitch). The process was repeated 10 times for each of the 65 melodies, yielding a total of 650 trials. In each trial, the model's analyses for the correct version and the deformed version were compared simply with regard to the total probability given to each melody (as defined in equation 4), to see which version was assigned higher probability. In effect, then, the model simply judged which of a pair of melodies was more likely to contain an error, without expressing any opinion as to exactly where the error was.

The model assigned the correct version of the melody higher probability than the deformed version in 573 out of 650 trials (88.2%). This level of performance seems promising. Probably, not all random "errors" of this type would be identifiable as errors even by humans; whether the model's ability is comparable to that of human listeners remains to be tested.

5. Further Issues

We have presented a probabilistic model which performs key identification as well as the surface-level tasks of expectation and error detection. On balance, where comparison is possible, the model is at least competitive with other models in its level of performance. Beyond the issue of performance, however, the current model has important advantages over others that have been proposed. In particular, the current model is able to perform both the structural task of key identification and the surface-level tasks of expectation and error detection within a single framework. This sets it apart from prior models, which have addressed these problems separately. The connection between expectation and key-finding is indirectly reflected in some earlier work—notably in the fact both expectation models [9, 10] and key-finding models [8] have made use of key-profiles. But this connection is brought out much more clearly in the current approach. The current model also provides a natural way of calculating the overall probability of a melody, which earlier key-finding and expectation models do not.

A further aspect of melody perception deserving brief mention here is the actual identification of notes. The extraction of note information from an auditory signal is a complex process, involving the grouping of partials (individual frequencies) into complex tones, and the correct categorization of these complex tones into pitch categories. (See [11] for a review of recent research on this problem.) It seems likely that the model proposed above could contribute to this task, by evaluating the probability of different possible note patterns (as in the error-detection task above). These judgments could then be used in a "top-down" fashion—in effect, bringing to bear musical considerations such as key, pitch proximity, and range on the transcription process. Some efforts have been made to apply Bayesian methods to transcription [1, 12], but much more could be done in this area.

One obvious question that arises here is whether the current model could be extended to handle polyphonic music. I have argued elsewhere [13] that a rather different approach to key identification is required in polyphonic music. Briefly, there is so much use of doubled and repeated pitch-classes in polyphonic music that counting every event gives too much weight to such pitch-classes; a better approach is to simply judge each pitch-class as "present" or "absent" within a small segment of music. (See [13] for a polyphonic key-finding model based on this idea.) However, this model cannot be used to calculate the probability of a pitch pattern, and is therefore not well-suited to modeling expectation and error detection. The way these problems might be addressed in polyphonic music remains an open question.

References

[1] K. Kashino, K. Nakadai, T. Kinoshita, and H. Tanaka, "Application of Bayesian probability networks to musical scene analysis," in *Computational Auditory Scene Analysis*, D.F. Rosenthal and H.G. Okuno, Eds. Mahwah, NJ: Lawrence Erlbaum, 1998, pp. 115-137.

[2] A. T. Cemgil, B. Kappen, P. Desain, and H. Honing, "On Tempo tracking: Tempogram representation and Kalman filtering," *Journal of New Music Research*, vol. 29, pp. 259–273, 2000.

[3] C. Raphael, "A hybrid graphical model for rhythmic parsing," *Artificial Intelligence,* vol. 137, pp. 217–238, 2002.

[4] D. Temperley, *Music and Probability*, Cambridge: MIT Press, forthcoming.

[5] H. Schaffrath, *The Essen Folksong Collection*, David Huron, Ed. Stanford, CA: Center for Computer–Assisted Research in the Humanities, 1995.

[6] P. von Hippel and D. Huron, "Why do skips precede reversals? The effect of tessitura on melodic structure," *Music Perception*, vol. 18, pp. 59–85, 2000.

[7] H. C. Longuet–Higgins and M. J. Steedman, "On interpreting Bach," *Machine Intelligence*, vol. 6, pp. 221–241, 1971.

[8] C. L. Krumhansl, *Cognitive Foundations of Musical Pitch*, New York: Oxford University Press, 1990.

[9] L. L. Cuddy and C. A. Lunney, "Expectancies generated by melodic intervals: Perceptual judgments of melodic continuity," *Perception & Psychophysics*, vol. 57, pp. 451–62, 1995.

[10] E. G. Schellenberg, "Simplifying the implication-realization model of melodic expectancy," *Music Perception*, vol. 14, pp. 295–318, 1997.

[11] A. P. Klapuri, "Automatic music transcription as we know it today," *JNMR*, vol. 33, pp. 269–282, 2004.

[12] S. A. Abdallah, and M. D. Plumbley, "Polyphonic transcription by non-negative sparse coding of power spectra," in *ISMIR 2004*.

[13] Temperley, D, "Bayesian models of musical structure and cognition," *Musicae Scientiae*, vol. 8, pp. 175–205, 2004.