EVALUATING AND VISUALIZING EFFECTIVENESS OF STYLE EMULATION IN MUSICAL ACCOMPANIMENT

Ching-Hua Chuan* **and Elaine Chew**†
University of Southern California Viterbi School of Engineering
*Department of Computer Science and †Epstein Department of Industrial and Systems Engineering
†Radcliffe Institute for Advanced Study at Harvard University
{chinghuc, echew}@usc.edu

ABSTRACT

We propose general quantitative methods for evaluating and visualizing the results of machine-generated style-specific accompaniment. The evaluation of automated accompaniment systems, and the degree to which they emulate a style, has been based primarily on subjective opinion. To quantify style similarity between machine-generated and original accompaniments, we propose two types of measures: one based on transformations in the neo-Riemannian chord space, and another based on the distribution of melody-chord intervals. The first set of experiments demonstrate the methods on an automatic style-specific accompaniment (ASSA) system. They test the effect of training data choice on style emulation effectiveness, and challenge the assumption that more data is better. The second set of experiments compare the output of the ASSA system with those of a rule-based system, and random chord generator. While the examples focus primarily on machine emulation of Pop/Rock accompaniment, the methods generalize to music of other genres.

1 INTRODUCTION

Automatic generation of music in a specific style has been a focal topic in computational music research since the early days of computing. Many researchers have designed systems that emulate music styles in compositions. The evaluation of the musical output, the degree to which it achieves its goal of emulating a particular style remains a challenge. Evaluation is often in the form of subjective opinion. It is our goal to fill this gap in quantitative evaluation of style emulation in automatic accompaniment systems.

In [1], Meyer states that "style is a replication of patterning, . . . , that results from a series of choices made within some set of constraints." Accompaniment can be considered the outcome of a series of choices over possible chords under certain contextual constraints, such as melody, phrase, and key. For example, in four-part harmonization, the composition must follow the counterpoint and voice-leading rules in music theory. In contrast, modeling of the accompaniment style in Pop/Rock tends to be vague and difficult because, according to Stephenson [2], "in rock, . . . , melody is allowed to interact with harmony more freely," and the process becomes more complex because, as Everett states [3], "many different tonal systems are now practiced by the same artist, on the same album."

In this paper we propose methods to visualize, measure, and quantify the quality of an accompaniment generated with the goal of emulating an original accompaniment, the "ground truth". These measures are designed to evaluate automatic style-specific accompaniment (ASSA) systems. We examine the style as captured by the musical relations between melody and harmony, and between adjacent chords. We develop six metrics based on these two types of measures.

Using these quantitative methods, we design experiments to explore training set selection strategies that further the ASSA system's style emulation goals. For machine learning tasks, it is often the case that more training data guarantee better results. For ASSA goals, more training songs may not necessarily improve the output quality if the training set is not consistent with the desired style. The first set of experiment explores the factors impacting style emulation success.

A second set of experiments compare the degree of style-specificity between the best ASSA systems, a rule-based harmonization system, and a random chord generator. We conduct the experiments on five well known Pop/Rock albums by Green Day, Keane, and Radiohead, providing detailed statistics and case studies for the resulting accompaniments. The findings we report generalize to genres outside of Rock music, and to some extent to simulation of style in music in general.

The paper is organized as follows: Section 2 describes related automatic accompaniment systems, with and without style requirements, and their evaluations. It concludes with a brief description of an ASSA system, which forms the basis of much of our evaluations. We present the evaluation methods in Section 3. Intra-system comparisons are shown in Sections 4, and inter-system results in 5, followed by the conclusions.

2 RELATED WORK

Baroque style four-part harmonization has been a popular accompaniment generation application since the earliest days of computing. Recent examples include [4, 5]. Many rules govern these Chorale-type harmonizations, which can be applied in their synthesis and evaluation. Sixteenth century compositions in the style of Palestrina have also be emulated using Markov models [6]. Such compositions are similarly and strictly circumscribed by a host of rules, which can be used in the evaluation of their correctness [7]. Temperley and Sleator proposed a set of preference rules to harmonizing melodies in the Western classical style [8]; these rules are implemented in their Harmonic Analyzer [9].

In the popular realm, the i-Ring system [10] generates an accompaniment for any eight-measure melody. The accompaniment is based on state transition probabilities calculated from a training set of 150 songs. For evaluation, 10 participants were asked to rate the accompaniment as Good, Medium, or Bad.

More recently, MySong [11] uses Hidden Markov Models, training on 298 popular songs in genres including Pop, Rock, R&B, Jazz, and Country Music. Users can choose between two style-related modes: 'Jazz,' which uses less common triads, or 'Happy,' which selects more major-mode chord transitions. 26 sample accompaniments by MySong were evaluated subjectively by 30 volunteer musicians.

While the output of the above systems fit the melody, because the systems learn from general training sets, the style captured by the output lacks specificity. Thus, these systems do not address the emulation of specific accompaniment styles, as embodied in a distinctive song or a particular band's output. Evaluations take the form of subjective opinion, and lack an objective or quantitative component.

In [12], Chuan & Chew proposed an ASSA system that can generate style-specific accompaniment to a melody given only a few examples. The system consists of a chord tone determination and a chord progression module. The chord tone determination module applies machine learning to determine which notes in a melody are likely chord tones based on the specified style. The system then constructs possible chord progressions using neo-Riemannian transforms, representing them in a tree structure, and selects the highest probability path based on the learned probabilities in a Markov chain. This system was rated informally by subjectively judgement via a Turing test.

3 QUANTIFYING ACCOMPANIMENT DISTANCE

This section first presents ways to quantify and visualize distance between two chords, and metrics to evaluate distance between two different accompaniments to the same melody. The examples are generated by the ASSA system of [12].

3.1 Visualization

Two visualization models are described as follows. The first considers musical distance between two chords based on neo-Riemannian transforms, the second addresses the difference between chord choices in relation to the melody.

3.1.1 Neo-Riemannian Distance in Chord Space

Music theorists have used neo-Riemannian transformations to analyze harmonic diversity in Pop/Rock music [13]. In the neo-Riemannian chord space, chords are connected by three types of neo-Riemannian operations (NROs), P (parallel), L (leading-tone exchange), and R (relative).

Figure 1 shows the three types of NROs in chord space, where chords are represented in Roman numerals. Vertical neighbors are connected by P/R operations, and horizontal neighbors have Dominant/Subdominant (D/S) relationships.

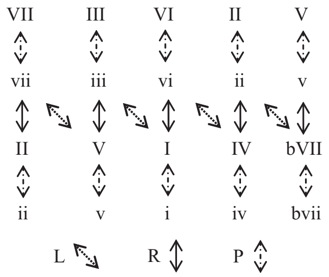


Figure 1. Neo-Riemannian transforms in chord space.

To quantify the distance between two chords – for example, to determine how well a generated chord compares to the original – we can compute the shortest distance between the two chords in the neo-Riemannian space. Two chords connected by an NRO share two tones, with the third being one or two half steps apart. The fewer NROs between the chords, the closer they are perceptually.

Chord Space Distance-Time Plot: To visualize the distance between a machine-generated and an original accompaniment, we can graph this distance in chord space over time. Figure 2 shows one such graph for comparing the generated and original accompaniments.

The x-axis marks the chord count; the yz-axes represent chord space. The plot charts a given accompaniment's distance from the baseline (original), according to chord space distance. If the two accompaniments are identical, the graph would be a horizontal line defined by $y = z = 0$. The advantage of this visualization is that we can see the quality of each generated chord by observing its neo-Riemannian distance from the original, and we can observe their chord relation by examining the direction of the deviation.

Chord Map Distribution: The second visualization employs the key map idea proposed in [14]. A chord map (Fig-

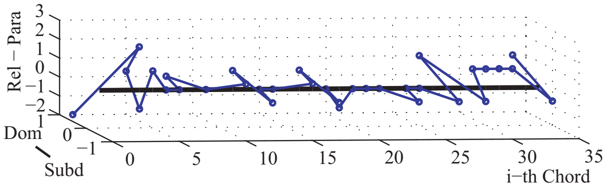


Figure 2. Chord space distance-time plot for Keane's *Somewhere Only We Know* (K1) trained on *Bedshaped* (K11) in the album *Hopes and Fears*.

ure 3) presents a nine-cell grid in which each cell presents a closely related chord with respect to the center chord. Apart from D/S and P/R, the nine cells include the I (Identity) and the combination operations: DR, SR, DP, and SP.

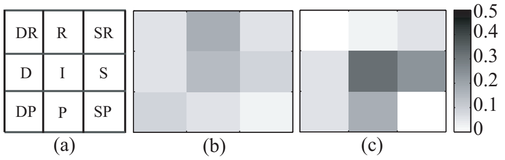


Figure 3. Chord map distributions: (a) the nine close chords; (b) result for K1 trained on K5 (*She Has No Time*); and, (c) result for K1 trained on K11

We count the number of chords in these nine categories, with respect to the original, in a generated accompaniment, and divide the number by the total number of chords. Two resulting grayscale plots are shown in Figures 3 (b) and (c). The plots show that the accompaniment for K1 trained on K11 is markedly better than that trained on K5.

3.1.2 Melody-Chord Interval Distribution

Another way to compare two accompaniments is by examining the relation between the chords and the melody they harmonize, to capture the degree of consonance-dissonance.

Suppose a subsequence of melodic notes, with pitch-duration values $\{(a_i, d_i), i = 1, \ldots, m\}$, is accompanied by a chord $\mathbf{x}$ containing the pitches $\{x_1, \ldots, x_n\}$. Assume that the a_i's and the x_j's have been normalized by the key of the melody so that 0 represents the tonic. The weight of a particular interval, k, between the melody pitches, $\mathbf{a}$, and the chord, $\mathbf{x}$, is defined as follows:

$$W_{\mathbf{a},\mathbf{x}}(k) = \sum_{\{i,j:(a_i - x_j) mod 12 = k\}} d_i, \qquad (1)$$

where $k = 0, \ldots, 11$. For example, the interval weight vector for a melodic note C, with duration d, and a C major triad is $[d, 0, 0, 0, 0, d, 0, 0, d, 0, 0, 0]$. We then normalize the interval weight vector to obtain the interval distribution:

$$Dist_{\mathbf{a},\mathbf{x}}(k) = \frac{W_{\mathbf{a},\mathbf{x}}(k)}{\sum_{i=0}^{11} W_{\mathbf{a},\mathbf{x}}(i)}. \qquad (2)$$

To compare two accompaniments, we take the difference between their melody-chord distributions. Suppose we have a sequence of melodic fragments $\{\mathbf{a}_1, \ldots, \mathbf{a}_T\}$, their original accompanying chords, $\{\mathbf{x}_1, \ldots, \mathbf{x}_T\}$, and chords in the generated accompaniment, $\{\mathbf{y}_1, \ldots, \mathbf{y}_T\}$. We calculate the difference between the distributions as follows:

$$\Delta_i = Dist_{a_i, y_i} - Dist_{a_i, x_i}, \text{ where } i = 1, \ldots, T. \qquad (3)$$

Melody-Chord Interval Distribution Difference: Figure 4 shows a visualization of the difference between two melody-chord interval distributions: the original accompaniment to the melody of K1, and one by the system trained on K11. The three-dimensional map of interval (half-step) distribution over time is shown as a two-dimensional grayscale plot. The horizontal axis shows the chord count, while the vertical axis the number of half steps.

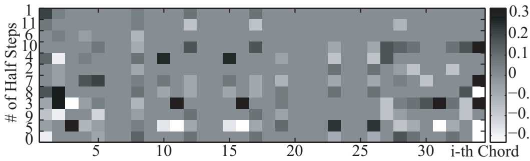


Figure 4. Ordered melody-chord interval distribution difference between the original K1 accompaniment and one trained on K11.

A predominance of darker gray indicates that the generated accompaniment contains extra types of half steps not in the original, while lighter colors mean that the generated chords lack certain types of half steps.

For visual coherence, we order the y-axis by the distribution of the original accompaniment, $Dist_{\mathbf{a},\mathbf{x}}$, in descending order, up the vertical axis. In this way, we can expect a generated accompaniment to be more stylistically distant from the original (having more frequent rare intervals) when more darker cells appear near the top of the graph, and less stylistically consistent (having fewer of the popular intervals) if more lighter cells are present near the bottom.

3.2 Metrics

We propose six metrics, three percentages and three distance metrics, to quantitatively assess a generated accompaniment, as shown in Table 1.

The first metric, *correct rate*, calculates the percentage of melodic notes that are correctly classified as chord tones and non-chord tones in the generated accompaniment, where chord and non-chord tones are given by examining the original accompaniment. This measure pertains to the ASSA system in [12], and to other systems that classify melody notes into chord and non-chord tones.

Table 1. Metric values' names, meanings and ranges

name	meaning	max	min
correct rate	% correct chord/non-chord tone classification	100	0
same chords	% of generated chords identical with original	100	0
chords-in-grid	% of generated chords in chord map	100	0
NR distance	ave NR distance bet generated/original chords	7	0
HS distance	MSE between interval (half-step) dist	2	0
wHS dist	weighted MSE bet interval distributions	2	0

The second metric, *same chords*, records the percentage of chords generated that are identical to the original. Note that this is also the value in the center cell of the chord maps shown in Figure 3. The third metric, *chords in grid*, gives the percentage of chords generated that are closely related to the original on the chord map; this value can be computed by summing all values in the cells of the chord map.

The *NR distance* metric shows the average shortest neo-Riemannian distance between the generated chord and the original, i.e., the average minimum neo-Riemannian distance between each data-point and the center line in the chord space distance-time plot shown in Figure 2. Note that the maximum NR distance between any two chords is 7.

The last two metrics are generated from the melody-chord interval distribution described in Section 3.1.2. The *HS distance* is the mean squared difference between the interval (half-step) distributions of the generated and the original accompaniments:

$$HS = \frac{1}{T}\sum_{i=1}^{T}\sum_{k=1}^{11}\Delta_i(k)^2. \tag{4}$$

The worst case occurs when only one type of half step dominates the generated and original accompaniments' distributions, but the types are different, causing a distance of 2.

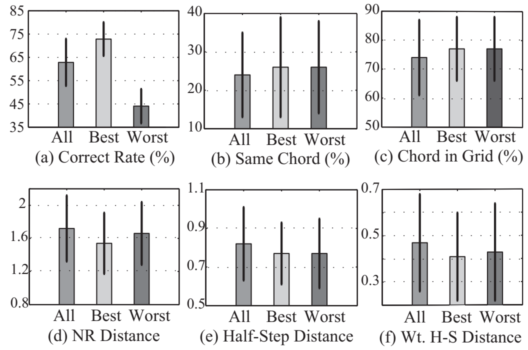
The final metric *wHS distance* is like the HS distance, except that Δ_i is multiplied by the inverse of the half-step distribution value for the original accompaniment. The weighted HS distance increases the weight for infrequent half steps, reflecting the fact that rare, dissonant intervals can significantly impact the perceived accompaniment style.

Let $Dist_{\mathbf{a},\mathbf{x}}$ be the half-step distribution of the original accompaniment for melody fragment $\mathbf{a}$. The weighting factor for each half step, $k = 0, \ldots, 11$, can be calculated as:

$$w(k) = \begin{cases} 1/Dist_{\mathbf{a},\mathbf{x}}(k), & \text{if } Dist_{\mathbf{a},\mathbf{x}}(k) > 0 \\ 0.001, & \text{if } Dist_{\mathbf{a},\mathbf{x}}(k) = 0 \end{cases} \tag{5}$$

The weighted HS distance is then calculated as:

$$wHS = \frac{1}{T}\sum_{i=1}^{T}\sum_{k=1}^{11} w(k) \times \Delta_i(k)^2, \tag{6}$$



(a) Correct Rate (%)
(b) Same Chord (%)
(c) Chord in Grid (%)
(d) NR Distance
(e) Half-Step Distance
(f) Wt. H-S Distance

Figure 5. Summary statistics for generated accompaniments over five albums

4 INTRA-SYSTEM COMPARISONS

This section describes experiments involving only the ASSA system of [12]. Recall that this system has a chord tone determination (CTD) module, and a chord progression module. We examine the general quality of the results generated by the system, focussing on the impact of training data on the quality of the machine-generated accompaniment, so as to improve the choice of training songs.

The data set consists of songs from the five albums: Green Day's *Dookie* and *American Idiots*, Keane's *Hopes and Fears*, and Radiohead's *Pablo Honey* and *Hail to the Thief*. We use commercial lead sheets for the ground truth.

4.1 One-Song v.s. All-Except-One Training Policy

To test whether more data is indeed better in ASSA, these experiments compare two training song choices. The first selects one song from an album for training, and tests the model on another song in the same album. The second uses all except one song for training, and tests the model on the held out song.

Figure 5 shows the statistics. "Best" and "Worst" report the performance by the model that is trained on a single-song in the same album, and that has the highest and lowest correct CTD rates, respectively. "All" refers to results where all other songs in the same album form the training set.

Note that except for the correct CTD rate, Figure 5(a), the statistics for Worst-CTD consistently outperform those for All. The results indicate that more data is not always better, and that the use of neo-Riemannian transforms offer a robust way to generate chord progressions, even when chord tone information is poor.

4.2 Correct Chord Tones vs. Best Chord Overlap

To test the ASSA system's sensitivity to different parameters, we compare the Best-CTD performance in the previous section with two other training data selection policies.

The first uses the three songs in the same album with the best CTD correct rates; the second uses the three songs that share the most common chords (CC) with, and having the least extra chords over, the test song.

Figure 6 shows the summary statistics for these tests. We observe that the CC training set achieves the highest *same chord* and *chords in grid* percentages, and reports the lowest half-step distance metrics.

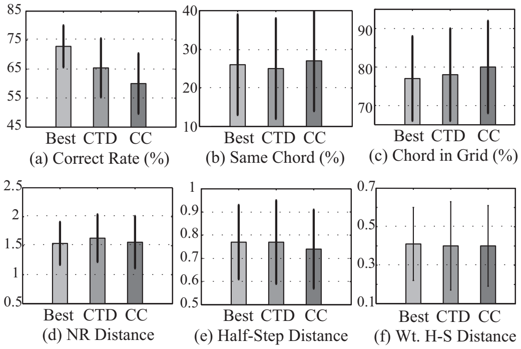


Figure 6. Summary statistics for accompaniments generated using Best, CTD and Common Chord training data sets

4.3 Case Study: Visualization of Details of Two Results

In this section we examine two accompaniments by visualizing their chord space distance-time plots, and their melody-chord interval distribution difference graphs. We compare the output of systems trained on the CC set (the best training set), and on all other songs in the same album (the worst training set.) Figure 7 shows the first 12 bars of the melody and the original accompaniment on Keane's *Your Eyes Open*, and the chords generated by ASSA with the CC and All training sets.

Figure 8 shows the neo-Riemannian distance plots over time. Observe that the accompaniment generated by All contains larger deviations from the original than that by the CC training set. Figure 9 shows the melody-chord interval distribution visualizations of the two generated accompaniments. Note that Figure 9(b) has more dark cells near the top than Figure 9(a). These darker cells indicate that the accompaniment generated by the All training set contains more melody-chord intervals that are rare in the original.

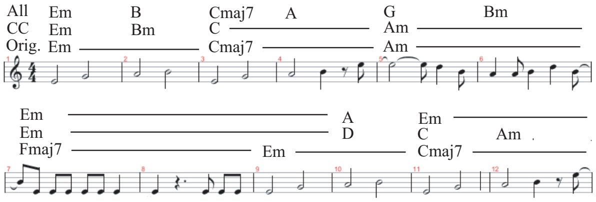


Figure 7. The first 12 bars of Keane's *Open Your Eyes*

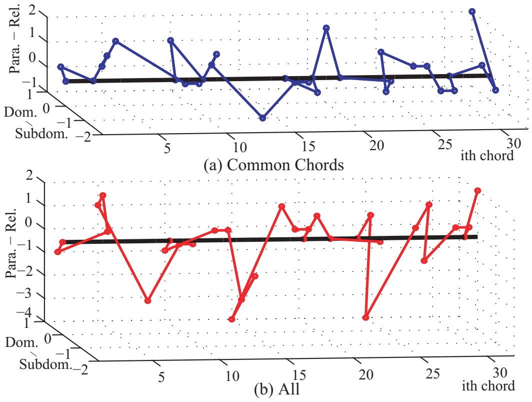


Figure 8. Chord space distance-time plot for Keane's *Open Your Eyes*

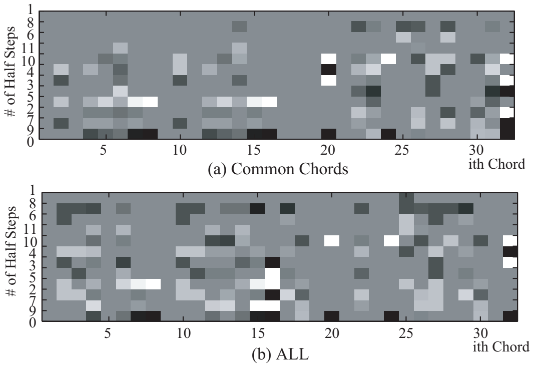


Figure 9. Interval distribution distance graphs for Keane's *Open Your Eyes*

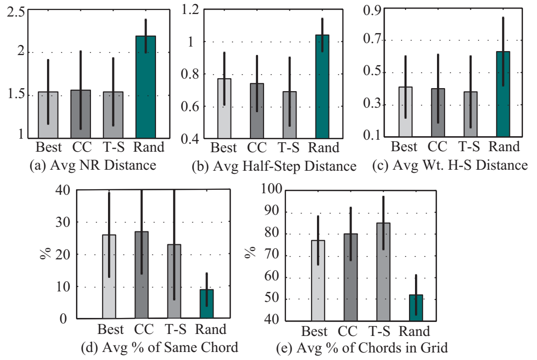
5 INTER-SYSTEM COMPARISONS

In this section we compare the ASSA system [12] with two rule-based harmonization systems, Temperley and Sleator's (T-S) Harmonic Analyzer [8], and a random harmonizer with only one constraint.

The Harmonic Analyzer [9] generates the root of the chords for harmonization without indicating their modes (major or minor). For comparison, we interpret the chords as being the common ones as described in [15]. For example, when G is reported by the Harmonic Analyzer in a song in C major, we assign a G major instead of a G minor chord.

We further design a simple random chord selector as the base case for the comparisons. In order to construct a reasonable accompaniment system, we add one constraint to the random chord generator: an accompaniment chord must contain at least one of the melody notes it harmonizes, and are randomly assigned if the bar has no melody note.

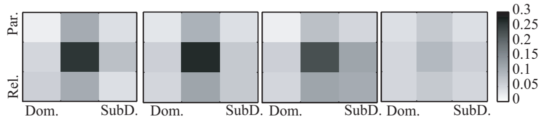
Figure 10 shows summary statistics comparing the ASSA with the Best-CTD and the CC training sets, the T-S Harmonic Analyzer, and the random chord selector (Rand). The figure shows that (a) the ASSA and T-S systems report similar neo-Riemannian distances, (b) and (c) T-S generates shorter half-step distances, and (e) achieves a higher percentage of chords in grid; however, (d) ASSA with both training sets outperforms T-S on same chord percentage.



(a) Avg NR Distance (b) Avg Half-Step Distance (c) Avg Wt. H-S Distance

(d) Avg % of Same Chord (e) Avg % of Chords in Grid

Figure 10. Overall statistics over systems.



(a) Best (b) Com. Chord (c) T-S (d) Random

Figure 11. Chord map distributions of systems.

Figure 11 shows the chord map distributions for the four systems. (a) and (b) show that more chords generated by the ASSA system match the original exactly. They also show that ASSA tends to generate chords that are parallel or relative to the original. In (c), we observe that the T-S system generates almost equal numbers of chords in R, S, and SR. In (d), the chords generated by the random chord selector are relatively evenly distributed among the close chords.

6 CONCLUSIONS

We have proposed quantitative methods for evaluating and visualizing machine-generated style-specific accompaniments. Using these methods, we showed that a training set with more chords in common with original leads to better style emulation. In the inter-system comparisons, we showed that the ASSA system produces more chords identical to those in the original song than the T-S system.

7 ACKNOWLEDGEMENTS

This material is based in part upon work supported by a USC Digital Dissertation Fellowship and NSF Grant No. 0347988. Any opinions, findings, and conclusions or recommendations expressed herein are those of the authors, and do not necessarily reflect the views of USC or NSF.

8 REFERENCES

[1] Meyer, L.B. *Style and Music: Theory, History, and Ideology*, University of Pennsylvania Press, 1990.

[2] Stephenson, K. *What To Listen For In Rock: A Stylistic Analysis*, pp. 75, New Haven: Yale U. Press, 2002.

[3] Everett, W. “Making Sense of Rock’s Tonal Systems,” *Music Theory Online*, Vol. 10, No. 4, 2004.

[4] Allan, M. and Williams, C.K.I. “Harmonising Chorales by Probabilistic Inference,” *Proc. of the Neural Information Processing Systems Conf.*, Vancouver, 2004.

[5] Phon-Amnuaisuk, S., Tuwson, A., and Wiggins, G. “Evolving Music Harmonisation,” *Artificial Neural Nets and Genetic Algorithms: Proc. of the 4th Intl. Conf.*, Portoroz, 1999.

[6] Farbood, M. and Schoner, B. “Analysis and Synthesis of Palestrina-Style Counterpoint Using Markov Chains,” *Proc. of the Intl. Computer Music Conf.*, Havana, 2001.

[7] Huang, C.Z.A. and Chew, E. “Palestrina Pal: A Grammar Checker for Music Compositions in the Style of Palestrina,” *Proc. of the 5th Conf. on Understanding and Creating Music*, Caserta, 2005.

[8] Temperley, D. and Sleator, D. “Modeling Meter and Harmony: A Preference Rule Approach,” *Computer Music Journal*, Vol. 15, No. 1, pp. 10 - 27, 1999.

[9] Temperley – Sleator Harmonic Analyzer, *www.cs.cmu.edu/ sleator/harmonic-analysis*.

[10] Lee, H.R. and Jang, J.S. “i-Ring: A System for Humming Transcription and Chord Generation,” *Proc. of the IEEE Intl. Conf. on Multimedia and Expo*, Taipei, 2004.

[11] Morris, D., Simon, I., and Basu, S. “MySong: Automatic Accompaniment Generation for Vocal Melodies,” *Proc. of Computer-Human Interaction*, Florence, 2008.

[12] Chuan, C.H. and Chew, E. “A Hybrid System for Automatic Generation of Style-Specific Accompaniment,” *Proc. of the 4th Intl. Joint Workshop on Computational Creativity*, London, 2007.

[13] Capuzzo, G. “Neo-Riemannian Theory and the Analysis of Pop-Rock Music,” *Music Theory Spectrum*, Vol. 26, No. 2, pp. 177-199, 2004.

[14] Chuan, C.H. and Chew, E. “Audio Key Finding: Considerations in System Design, and Case Studies on 24 Chopin’s Preludes,” *EURASIP Journal on Applied Signal Processing*, Vol. 2007, Article ID 56561, 15 pages.

[15] Kostka, S. and Payne, D. *Tonal Harmony*, McGraw-Hill: New York, 2003.