Automatic Generation of Melodic Accompaniments for Lyrics

Kristine Monteith, Tony Martinez, and Dan Ventura
Computer Science Department
Brigham Young University
Provo, UT 84602 USA
kristinemonteith@gmail.com, martinez@cs.byu.edu, ventura@cs.byu.edu

Abstract

Music and speech are two realms predominately species-specific to humans, and many human creative endeavors involve these two modalities. The pairing of music and spoken text can heighten the emotional and cognitive impact of both - the complete song being much more compelling than either the lyrics or the accompaniment alone. This work describes a system that is able to automatically generate and evaluate musical accompaniments for a given set of lyrics. It derives the rhythm for the melodic accompaniment from the cadence of the text. Pitches are generated through the use of *n*-gram models constructed from melodies of songs with a similar style. This system is able to generate pleasing melodies that fit well with the text of the lyrics, often doing so at a level similar to that of human ability.

Introduction

Programmers and researchers have often attempted to endow machines with some form of intelligence. In some cases, the end goal of this is purely practical; a machine with the capacity to learn could provide a multitude of useful and resource-saving tasks. But in other cases, the goal is simply to make machines behave in a more creative or more "human" manner. As one author explains, "Looked at in one way, ours is a history of self-imitation...We are ten times more fascinated by clockwork imitations than by real human beings performing the same task." (McCorduck 2004).

One major area of human creativity involves the production of music. Wiggins (2006) states that, "...musical behavior is a uniquely human trait...further, it is also ubiquitously human: there is no known human society which does not exhibit musical behaviour in some form." Naturally, many computer science researchers have turned their attention to musical computation tasks. Researchers have attempted to classify music, measure musical similarity, and predict the musical preferences of users (Chai and Vercoe 2001; McKay and Fujinaga 2004). Others have investigated the ability to search through, annotate, and identify audio files (Dannenberg et al. 2003; Dickerson and Ventura 2009). More directly in the realm of computational creativity, researchers have developed systems that can automatically arrange and compose music (Oliveira and Cardoso 2007; Delgado, Fajardo, and Molina-Solana 2009).

Like music, speech is an ability that is almost exclusively human. While species such as whales or birds may communicate through audio expressions, and apes may even be taught simple human-like vocabularies and grammars using sign language, the complexities of human language set us apart in the animal kingdom. Major research efforts have been directed toward machine recognition and synthesis of human speech (Rabiner 1989; Koskenniemi 1983) Computers programs have been designed to carry on conversations, some of them doing so in a surprisingly human-like manner (Weizenbaum 1966; Saygin, Cicekli, and Akman 2000). More creative programming endeavors have involved the generation of poetry (Gervás 2001; Rahman and Manurung 2011) or text for stories (Riedl 2004; Pérez y Pérez and Sharples 2004; Gervás et al. 2005; Ang, Yu, and Ong 2011).

Gfeller (1990) points out the similarities between speech and music: "Both speech and music are species specific and can be found in all known cultures. Both forms of communication evolve over time and have structural similarities such as pitch, duration, timbre, and intensity organized through particular rules (i.e. syntax or grammar) that result in listener expectations." Studies show that music and the spoken word can be particularly powerful when paired together. For example, in one study, researchers found that a sung version of a story was often more effective at reducing an undesirable target behavior than a read version of the story (Brownell 2002). Music can help individuals with autism and auditory processing disorders more easily engage in dialog (Wigram 2002). The pairing of music with language can even help individuals regain lost speech abilities through a process known as Melodic Intonation Therapy (Gfeller 1990; Schlaug, Marchina, and Norton 2008). On the other hand, lyrics have the advantage of being able to impart discursive information where the more abstract nature of music makes it less fit to do so (Kreitler and Kreitler 1972). Lyrics can also contribute to the emotional impact of a song. One study found that lyrics enhanced the emotional impact of a selection with sad or angry music (Ali and Peynircioglu 2006). Another found that lyrics tended to be a better estimator of the overall mood of a song than the melody when the lyrics and the melody disagree (Wu et al. 2009).

This work describes a system that can automatically com-

pose melodic accompaniments for any given text. For each given lyric, it generates hundreds of different possibilities for rhythms and pitches and evaluates these possibilities with a number of different metrics in order to select a final output. The system also incorporates an awareness of musical style. It learns stylistic elements from a training corpus of melodies in a given genre and uses these to output a new piece with similar elements. In addition to self-evaluation, the generated selections are further evaluated by a human audience. Survey feedback indicates that the system is able to generate melodies that fit well with the cadence of the text and that are often as pleasing as the original accompanying tunes. Colton, Charnley, and Pease (2011) suggest a number of different measures that can be used to evaluate systems during the creative process. We direct particular attention to two of these–precision and reliability–and demonstrate that, for simpler styles, our system is able to perform well with regard to these metrics.

Related Work

Conklin (2003) summarizes a number of statistical models which can be used for music generation, including random walk, Hidden Markov Models, stochastic sampling, and pattern-based sampling. These approaches can be seen in a number of different studies. For example, Chuan and Chew (2007) use Markov chains to harmonize given melody lines, focusing on harmonization in a given style. Cope (2006) also uses statistical models to generate music in a particular style, producing pieces indistinguishable from human-generated compositions. Pearce and Wiggins (2007) provide an analysis of a number of strategies for melodic generation, including one similar to the generative model used in this paper.

Delgado, Fajardo, and Molina-Solana (2009) use a rule-based system to generate compositions according to a specified mood. Oliveira and Cardoso (2007) describe a wide array of features that contribute to emotional content in music and present a system that uses this information to select and transform chunks of music in accordance with a target emotion.

Researchers have also directed efforts towards developing systems intended for accompaniment purposes. Dannenberg (1985) presents a system of automatic accompaniment designed to adapt to a live soloist. Lewis (2000) also details a "virtual improvising orchestra" that responds to a performer's musical choices.

While not directly related to generating melodic accompaniment for lyrics, a number of studies have looked at aligning musical signals to textual lyrics (the end result being similar to manually-aligned karaoke tracks). For example, Wang and associates (2004) use both low-level audio features and high-level musical knowledge to find the rhythm of the audio track and use this information to align the music with the corresponding lyrics.

Methodology

In order to generate original melodies, a set of melodies is compiled for each different style of composition. These melodies were isolated from MIDIs obtained from the Free MIDI File Database[1] and the "I Love MIDIs" website[2]. These selections help determine both the rhythmic values and pitches that will be assigned to each syllable of the text. The system catalogs the rhythmic patterns that occur for each of the various numbers of notes in a given measure. The system also creates an n-gram model representing what notes are most likely to follow a given series of notes in a given set of melodies. Models were developed for three stylistic categories: nursery rhymes, folk songs (bluegrass), and rock songs (Beatles).

For each lyric, the system first analyzes the text and assigns rhythms. It determines where the downbeats will fall for each given line of the text. One hundred different downbeat assignments are generated randomly, and evaluated according to a number of aesthetic measures. The system selects the random assignment with the highest score for use in the generated melody. The system then determines the rhythmic values that will be assigned to each syllable in the text by counting the number of syllables in a given measure and finding a rhythm that matches that number of syllables in one of the songs of the training corpus. Once rhythmic values are assigned, the system assigns pitches to each value using the n-gram model constructed from the training corpus. Once again, one hundred different assignments are generated and evaluated according to a number of metrics. Further details on the rhythm and pitch generation are provided in the following subsections.

Rhythm Generation

Rhythms are generated based on patterns of syllabic stress in the lyrics. Each word of the text is located in the CMU Pronunciation Dictionary[3] to determine the stress patterns of the constituent phonemes. (Each phoneme in the dictionary is labeled 0, 1, or 2 for "No Stress," "Primary Stress," or "Secondary Stress.") The system also looks up each word to determine if it occurs on a list of common articles, prepositions, and conjunctions.

The system then attempts to find the best positions for downbeats. For each given line of text, the system generates 100 possible downbeat assignments. The text of each line is distributed over four measures, so four syllables are randomly selected to carry a downbeat. Each assignment is then scored, and the system selects the assignment receiving the highest score for use in the melodic accompaniment. Downbeat assignments that fall on stressed syllables are rated highly, as are downbeats that fall on the beginning of a word and ones that do not fall on articles, prepositions, or conjunctions. Downbeat assignments that space syllables more evenly across the allotted four measures are also rated more highly (i.e. assignments that have a lower standard deviation for number of syllables per measure receive higher scores). See Figure 4 for further details on the precise downbeat scoring metrics. Figure 1 illustrates a possible downbeat assignment for a sample lyric.

[1] http://www.mididb.com/

[2] http://www.ilovemidis.com/ForKids/NurseryRhymes/

[3] http://www.speech.cs.cmu.edu/cgi-bin/cmudict

Lyrics:	Pat	a	cake	pat	a	cake
Phonemes:	PAET	AH	KEYK	PAET	AH	KEYK
Stress:	1	0	1	1	0	1
Downbeats:	true	false	false	true	false	false

Lyrics:	ba-	ker's	man
Phonemes:	BEY	KERZ	MAEN
Stress:	1	0	1
Downbeats:	true	false	true

Figure 1: Sample downbeat assignments for *Pat-A-Cake* lyrics

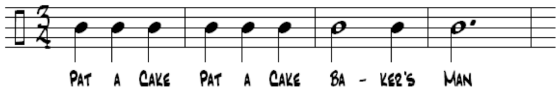

Figure 2: Default rhythm assignments for *Pat-A-Cake* lyrics

Once the downbeats are assigned, a rhythmic value is assigned to each syllable. The system randomly selects a piece in the training corpus to provide rhythmic inspiration. This selection determines the time signature of the generated piece (e.g. three beats or four beats to a measure). For each measure of music generated, the system looks to the selected piece and randomly chooses a measure that has the necessary number of notes. For example, if the system needs to generate a rhythm for a measure with three syllables, it randomly chooses a measure in the training corpus piece that has three notes in it and uses its rhythm in the generated piece. If no measures are available that match the number of syllables in the lyric, the system arbitrarily assigns rhythmic values, with longer values being assigned to earlier syllables. For example, in a measure with three syllables using a three-beat pattern, each syllable would be assigned to a quarter note. In a measure with four syllables, the first two syllables would be assigned to quarter notes and the last two syllables to eighth notes. Figure 2 illustrates the default rhythms assignment for a sample lyric.

Pitch Generation

Once the rhythm is determined, pitches are selected for the various rhythmic durations. Selections from a given style corpus are first transposed into the same key. Then an *n*-gram model with an *n* value of four is constructed from these original melodic lines. The model was created simply from the original training melodies, with no smoothing. For the new, computer-generated selections, melodies are initialized with a series of random notes, selected from a distribution that models which notes are most likely to begin musical selections in the given corpus. In order to foster song cohesion, each line of the song is initialized with the same randomly generated three notes. Additional notes for each line are randomly selected based on a probability distribution of what note is most likely to follow the given three notes as indicated by the *n*-gram model of the style corpus.

The system generates several hundred possible series of pitches for each line. Each possible pitch assignment is then scored. To encourage melodic interest, higher scores are given to melodic lines with a higher number of distinct

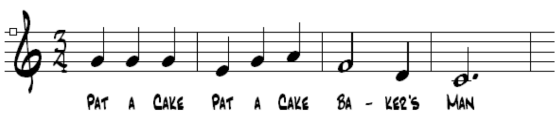

Figure 3: Sample pitch assignments for *Pat-A-Cake* lyrics

pitches and melodies featuring excessive repeated notes are penalized. Melodies with a range greater than an octave and a half or with interval jumps greater than an octave are penalized since these are less "sing-able." Melodic lines that do not end on a note in a typical major or minor scale and final melodic lines that do not end on a tonic note are given a score of zero. More precise details about the scoring of pitch assignments are given in Figure 4. Possible pitch assignments for a sample lyric are shown in Figure 3.

Results

Accompaniments were generated for lyrics in three stylistic categories: nursery rhymes, folk songs (bluegrass), and rock songs (Beatles). In each case, an attempt was made to find less commonly known melodies, so that the generated music could be more fairly compared to the original melodic lines. Melodic lines were generated for the following:

Nursery rhymes:

- Goosey Goosey Gander
- Little Bo Peep
- Pat-a-Cake
- Rub-a-Dub-Dub
- The Three Little Kittens

Folk songs:

- Arkansas Traveler
- Battle of New Orleans
- Old Joe Clark
- Sally Goodin
- Wabash Cannonball

Rock songs:

- Act Naturally
- Ask Me Why
- A Taste of Honey
- Don't Pass Me By
- I'll Cry Instead

Three melodies were generated for each of the fifteen lyrics considered. One was generated using a corpus of songs that matched the style of the lyrics (e.g. to generate a melody for *Goosey Goosey Gander* the four other nursery

MelodicAccompaniment($Lyric$,$StyleCorpus$)
for all $LINE_i$ in $Lyric$ **do**
  $STR_i \leftarrow$ patterns of syllabic stress in $LINE_i$
  $POS_i \leftarrow$ parts of speech for each syllable in $LINE_i$
  $BEG_i \leftarrow$ boolean values indicating that a syllable in $LINE_i$ begins a word
  for $i = 1 \rightarrow 100$ **do**
    $DB_j \leftarrow$ randomly assign downbeats to four syllables
    $score_j \leftarrow$ ScoreDownbeats(DB_j,STR_i,POS_i,BEG_i)
  end for
  $DB_i \leftarrow DB_j$ that coincides with the largest $score_j$
  $RHYTHM_i \leftarrow$ SelectRhythms(DB_i)
  for $i = 1 \rightarrow 100$ **do**
    $PITCHES_j \leftarrow$ assign pitches using *n*-gram model from $StyleCorpus$
    $score_j \leftarrow$ ScorePitches($PITCHES_j$)
  end for
  $PITCHES_i \leftarrow PITCHES_j$ that coincides with the largest $score_j$
  $MELODY_i \leftarrow$ combine $RHYTHM_i$ and $PITCHES_i$
  $MELODY += MELODY_i$
end for
return $MELODY$

ScoreDownbeats(DB_j,STR_i,POS_i,BEG_i)
for $k = 1 \rightarrow j$ **do**
  If DB_{jk} and $STR_{ik} = 1$ then $score += 1$
  If DB_{jk} and $POS_{ik}! = Art|Prep|Conj$ then $score += 0.5$
  If DB_{jk} and BEG_{ik} then $score += 0.5$
  $x \leftarrow maxSyllablesPerMeasure$
  $score += (x - stdDevSyllablesPerMeasure) * 0.5$
  $score += (x - numPickupSyllables) * 0.25$
  $score += (x - numSyllablesLastMeasure) * 0.25$
end for
return score

SelectRhythms(D_i, S_i)
$M \leftarrow$ divide S_i into measures based on D_i
$C \leftarrow$ randomly select a song in $StyleCorpus$
$R \leftarrow 0$
for all M_j in M **do**
  $R_j \leftarrow$ randomly selected measure from C with the same # of notes as syllables in M_j
  R += R_j
end for
return R

ScorePitches($PITCHES_j$)
$score \leftarrow uniquePitches(PITCHES_j)/size(PITCHES_j)$
If $MaxRepeatPitches(PITCHES_j) < maxRepeatPitches$ then $score += 1$
If $Range(PITCHES_j) < maxRange$ then $score += 1$
If $MaxInterval(PITCHES_j) < maxInterval$ then $score += 1$
If $!EndsOnScaleNote(PITCHES_j)$ then $score = 0$
If $LastLine(j)$ and $!EndsOnTonic(PITCHES_j)$ then $score = 0$
return score

Figure 4: Algorithm for automatically generating melodic accompaniment for text

	bluegrass	nursery	rock	average
bluegrass	1.34	3.09	1.19	1.87
nursery	1.14	3.32	1.19	1.88
rock	1.25	3.28	1.11	1.88
original	1.50	3.50	1.47	2.16

Table 1: Average responses to the question "How familiar are you with these lyrics?" Each row represents a compositional style and each column a category of lyrics.

rhyme songs were used to build the *n*-gram model) and two more were generated in the remaining two creative styles[4].

Study participants were divided into four groups. Each group was asked to listen to versions of songs for each of the fifteen lyrics, with selections for each group being a mixture of lyrics with the original human-composed melodies and lyrics with the three types of computer-generated melodies. Subjects were not informed that any of the melodies were computer-generated until after data collection. Fifty-two subjects participated in the study, and each melodic version was played for thirteen people.

After each selection, subjects were asked to respond to the following questions (1=not at all, 5=very much):

- How familiar are you with these lyrics? 1 2 3 4 5
- How familiar are you with this melody? 1 2 3 4 5
- How pleasing is the melodic line? 1 2 3 4 5
- How well does the music fit with the lyrics? 1 2 3 4 5
- Is this the style of melody you would have expected to accompany these lyrics? 1 2 3 4 5
- Are you familiar with any other melodies for these lyrics? YES NO

Table 1 shows the average responses to the question about familiarity of lyrics for each of the three categories. In each case, lyrics were rated as more familiar when they were paired with their original melodies as opposed to the computer-generated melodies. However, none of these differences were significant at the $p < 0.05$ level. The majority of subjects were relatively unfamiliar with the bluegrass and rock lyrics. The nursery rhyme lyrics were slightly more familiar, but in many cases, subjects were familiar with the lyrics but not any specific tune.

Table 2 shows the average responses to the question about familiarity of melody for each of the three categories. On average, subjects were slightly more familiar with the original melodies in the bluegrass and rock categories than they were with the lyrics. The original nursery rhymes melodies were rated as slightly less familiar on average than the lyrics. System-generated melodies received an average score of less than two for familiarity in each of the three categories (significantly lower than original melodies with a statistical significance of $p < 0.01$).

Subjects were likely to be less receptive to new melodies if they were very familiar with the old ones. (One respondent

[4]Selections generated for these experiments are available at http://axon.cs.byu.edu/emotiveMusicGeneration

	bluegrass	nursery	rock	average
bluegrass	1.62	1.49	1.40	1.50
nursery	1.53	2.17	1.34	1.68
rock	1.41	1.39	1.24	1.35
original	2.31	2.94	1.81	2.35

Table 2: Average responses to the question "How familiar are you with this melody?" Each row represents a compositional style and each column a category of lyrics.

	bluegrass	nursery	rock	average
bluegrass	3.50	3.50	3.56	3.52
nursery	3.37	3.24	3.09	3.23
rock	2.70	2.17	2.16	2.34
original	3.79	3.79	2.95	3.51

Table 3: Average responses to the question "How pleasing is the melodic line?" Each row represents a compositional style and each column a category of lyrics.

mentioned that hearing a new melody to a familiar childhood song was a little "unnerving".) Tables 3 through 7 report only the responses where subjects indicated that they were not familiar with an alternate melody for a given set of lyrics.

As shown in Table 3, the system was able to generate melodies that received the same average ratings for pleasing melodic lines as the original melodies. The average rating for songs in the bluegrass style was almost identical to that of the original melodies. The average ratings for pleasantness of generated nursery rhythm melodies was not significantly different than the original tunes.

For over a third of the lyrics, a computer-generated melody in at least one style was rated as more pleasing than the original melody. These tunes are listed in Table 4 along with their average ratings. For example, the original melody for *Battle of New Orleans* received a rating of 3.33 for average melodic pleasantness. The computer-generated melody for this lyric in a nursery rhyme style received a rating of 3.92. The original melody for *Little Bo Peep* received an average melodic pleasantness rating of 3.22. The bluegrass-styled computer-generated melody received a rating of 3.80, and the nursery-rhyme-styled generated melody received a rating of 3.43.

Table 5 shows that the original melodies were rated on average as fitting a little better with the lyrics (although the difference between the original melodies and the songs composed in the bluegrass style is not statistically significant). However, as shown in Table 6 a number of the individual computer-generated melodies were still rated as fitting better with the lyrics than the original melodies. For example, the rock version of *Old Joe Clark* received a rating of 3.00 from this metric while the original version received a rating of 2.75. Both the bluegrass and nursery-rhyme versions of *Ask Me Why* received higher ratings than the original version.

Table 7 reports responses to the question "Is this the style of melody you would have expected to accompany these lyrics?" Not surprisingly, the original melodies were

	Battle of New Orleans	Little Bo Peep	Rub A Dub Dub	Act Naturally	Ask Me Why	I'll Cry Instead
bluegrass	3.23	**3.60**	**3.80**	**3.50**	**4.23**	**3.79**
nursery	**3.92**	**3.43**	3.17	2.91	**3.14**	**2.92**
rock	2.83	2.60	2.13	2.54	2.00	**2.36**
original	3.33	3.22	3.50	2.70	2.83	2.12

Table 4: Average responses to the question "How pleasing is the melodic line?" for six songs where system-generated melody in one or more styles scored higher than the original melody.

	bluegrass	nursery	rock	average
bluegrass	3.59	3.20	3.18	3.32
nursery	3.35	3.36	2.71	3.14
rock	3.23	2.18	2.26	2.56
original	3.88	4.27	2.90	3.68

Table 5: Average responses to the question "How well does the music fit with the lyrics?" Each row represents a compositional style and each column a category of lyrics.

more "expected" on average than melodies composed in new styles. The computer-generated melodies composed in the style of the original melodies were also generally more expected with one exception: bluegrass melodies for rock lyrics tended to receive higher expectation ratings.

In a number of cases, the system was able to compose an unexpected melody that still received high ratings for pleasing melodies and a lyric/note match. Two such examples are shown in Table 8. In both cases, the songs received above average ratings for melodic pleasantness and average ratings for music/lyric match, but below average ratings for style expectedness.

Discussion

The original nursery rhymes were composed predominantly with notes of the major scale, and the rhythms in these songs were similarly simple. (Songs generated with corpus-inspired rhythms were quite similar to songs generated with the system's default rhythms.) With the exception of a flat seventh introduced by the mixolydian scale of *Old Joe Clark*, the bluegrass melodies also feature pitches exclusively from the major scale. Bluegrass rhythms also tended to be similarly straightforward. With simpler rhythms and fewer accidentals, more of the melodies generated in these two styles are likely to "work." The original bluegrass melodies tended to have more interesting melodic motion, and this appears to have translated into more interesting system-generated melodies. In contrast, the rock songs featured a much wider variety of scales and accidentals. These

	Arkansas Traveler	Old Joe Clark	Three Little Kittens	Ask Me Why	A Taste of Honey	I'll Cry Instead
bluegrass	**4.08**	2.71	**4.25**	**3.54**	2.57	**3.43**
nursery	3.08	2.75	3.80	**3.07**	**2.85**	**2.38**
rock	3.08	**3.00**	2.18	1.77	2.08	**2.27**
original	3.91	2.75	4.17	2.75	2.79	2.15

Table 6: Average responses to the question "How well does the music fit with the lyrics?" for six songs where system-generated melody in one or more styles scored higher than the original melody.

	bluegrass	nursery	rock	average
bluegrass	3.47	2.85	2.91	3.08
nursery	3.22	3.46	2.44	3.04
rock	3.12	1.82	2.14	2.36
original	3.69	4.27	2.79	3.58

Table 7: Average responses to the question "Is this the style of melody you would have expected to accompany these lyrics?"

extra tones do add color to the generated selections, but further refinements may be necessary to select which more complicated melodies are "fresh" or "original" instead of just "weird."

Wiggins (2006) proposes a definition for computational creativity as "The performance of tasks which, if performed by a human, would be deemed creative." The task of simply composing any decent new melody for an established tune could be considered creative. Composing one that improved on the original constitutes an even greater degree of creative talent. By this metric, our system fits the definition of "creative."

	Pat-A-Cake (bluegrass)	Act Naturally (bluegrass)
How pleasing is the melodic line?	3.80	3.50
How well does the music fit with the lyrics?	3.20	3.17
Is this the style of melody you would have expected?	2.60	2.50

Table 8: Average responses to questions for two songs where the melodic accompaniment was surprising but still worked.

Colton (2008) suggests that, for a computational system to be considered creative, it must be perceived as possessing skill, appreciation, and imagination. A basic knowledge of traditional music behavior allows a system to meet the "skillful" criteria. Our system takes advantage of statistical information about rhythms and melodic movement found in the training songs to compose new melodies that behave according to traditional musical conventions. A computational system may be considered "appreciative" if it can produce something of value and adjust its work according the preferences of itself or others. Our system addresses this criterion by producing hundreds of different possible rhythm and pitch assignments and evaluating them against some basic rules for pleasantness and singability. The "imaginative" criterion can be met if the system can create new material independent of both its creators and other composers. Since all of the generated melodies can be distinguished from songs in the training corpora, this criterion is met at least on a basic level. Our system further demonstrates its imaginative abilities by composing melodies in alternate styles that still manage to demonstrate an acceptable level of melodic pleasantness and synchronization with the cadence of the text.

Boden (1995) argues that unpredictability is also a critical element of creativity, and a number of researchers have investigated the role of unpredictability in creative systems (Macedo 2001; Macedo and Cardoso 2002) Our system meets the requirement of unpredictability with its ability to compose in various and sometimes unexpected styles. It is able to generate melodies that surprise listeners but still achieve high ratings for pleasantness.

Colton, Charnley, and Pease (2011) propose a number of different metrics in conjunction with their FACE and IDEA models that can be used to assess software during a session of creative acts. Equations for calculating these metrics are listed in Figure 5, were S is the creative system, (c_i^g, e_i^g) is a concept/expression pair generated by the system, $\overline{a^g}$ is an aesthetic measure of evaluation, and t is a minimum acceptable aesthetic threshold. Two of the measures suggested are precision (obtained by dividing the number of generated works by the number that met a minimum acceptable aesthetic level) and reliability (obtained from taking the system's best creation as calculated by some aesthetic measure and subtracting the system's worst). Table 9 reports the results of these calculations for the system's compositions in each of the three styles and compares them to the same metrics calculated for the original songs using responses to the question "How pleasing is the melodic line?" as the scoring metric. In order to calculate precision, we consider the worst score obtained by an original, human-composed melody to be the minimum acceptable threshold value. While the prize for most pleasing melody still goes to a human-composed song, all of the songs composed in a bluegrass and nursery style and two-thirds of the rock songs meet the basic criteria of being better than the worst original melody. The system is generating original melodies that are better than some established, human-generated songs a remarkable percentage of the time. The reliability of the system in generating bluegrass and nursery-style melodies is also worth mentioning. The reliability measures for these two categories are

$$average(S) = \frac{1}{n}\sigma_{i=1}^{n}\overline{a^g}(c_i^g, e_i^g)$$
$$best\_ever(S) = max_{i=1}^{n}(\overline{a^g}(c_i^g, e_i^g))$$
$$worst\_ever(S) = min_{i=1}^{n}(\overline{a^g}(c_i^g, e_i^g))$$
$$precision(S) = \frac{1}{n}|\{(c_i^g, e_i^g) : 1 < i < n \wedge \overline{a^g}(c_i^g, e_i^g) > t\}|$$
$$reliability(S) = best\_ever(S) - worst\_ever(S)$$

Figure 5: Assessment metrics proposed by Colton, Charnley, and Pease (2011)

	bluegrass	nursery	rock	original
average	3.52	3.23	2.34	3.51
best ever	4.23	3.92	3.83	4.50
worst ever	2.93	2.58	1.73	2.12
precision	1.00	1.00	0.67	1.00
reliability	1.30	1.33	2.11	2.38

Table 9: Assessment metrics calculate on average responses to the question "How well does the music fit with the lyrics?"

1.30 and 1.33 as compared to the 2.38 reliability measure for original songs. (Note that, for reliability, smaller scores are more desirable.) While the system probably shouldn't quit its day job to become a classic rock songwriter quite yet, it is considerably reliable at producing reasonable and pleasing melodies in the other two genres.

Similar results can be seen in Table 10 where responses to the question "How well does the music fit with the lyrics?" are used as the aesthetic measure. As with the previous calculations, the "worst ever" score for an original melody was used as a minimum aesthetic threshold for the generated melodies. Again, all of the nursery rhyme and bluegrass-styled compositions meet this threshold, as do two-thirds of the rock-styled songs. A song generated in the nursery rhyme or bluegrass style also more reliably matches the lyrics than an arbitrarily selected human-generated song.

Previous versions of our system analyzed each melody in a given training corpus according to a number of different metrics and used the results in the construction of neural networks designed to evaluate generated melodies (Monteith, Martinez, and Ventura 2010). For the sake of simplicity and computational speed, the most pertinent of these findings were distilled into rules for use by the system in these experiments. In other words, the information gathered by the system to date about melody generation has been simplified and codified so that more focus could be directed towards matching rhythms to text. However, the system could likely benefit from the use of additional metrics and further "observation" of human-generated and approved tunes in its attempts to create pleasing melodies. A similar process of evaluation could be applied to the process of rhythm generation, particularly in the assignment of downbeats. Currently, the system relies on a small set of arbitrary, pre-coded rules to determine downbeat placement. It would likely require a much larger training corpus than we currently have available, but perhaps more natural-sounding placements could be obtained if the system could learn from a corpus of "good" lyric/melody pairings the types of words

	bluegrass	nursery	rock	original
average	3.32	3.14	2.56	3.68
best ever	4.25	3.86	4.23	4.75
worst ever	2.57	2.36	1.63	2.15
precision	1.00	1.00	0.67	1.00
reliability	1.68	1.49	2.61	2.60

Table 10: Assessment metrics calculate on average responses to the question "How well does the music fit with the lyrics?"

and syllables best suited for supporting downbeats. Audience feedback could help determine an optimal weighting of the various evaluation criteria.

References

Ali, S. O., and Peynircioglu, Z. F. 2006. Songs and emotions: are lyrics and melodies equal partners? *Psychology of Music* 4(4):511–534.

Ang, K.; Yu, S.; and Ong, E. 2011. Theme-based cause-effect planning for multiple-scene story generation. In *Proceedings of the International Conference on Computational Creativity*, 48–53.

Boden, M. 1995. Creativity and unpredictability. *Stanford Humanities Review* 4.

Brownell, M. D. 2002. Musically adapted social stories to modify behaviors in students with autism: four case studies. *Journal of Music Therapy* 39:117–144.

Chai, W., and Vercoe, B. 2001. Folk music classification using hidden markov models. In *Proceedings of the International Conference on Artificial Intelligence*.

Chuan, C., and Chew, E. 2007. A hybrid system for automatic generation of style-specific accompaniment. In *Proceedings International Joint Workshop on Computational Creativity*, 57–64.

Colton, S.; Pease, A.; and Charnley, J. 2011. Computational creativity theory:the FACE and IDEA descriptive models. In *Proceedings of the 2nd International Conference on Computational Creativity*, 90–95.

Colton, S. 2008. Creativity versus the perception of creativity in computational systems. In *Creative Intelligent Systems: Papers from the AAAI Spring Symposium*, 14–20. Stanford, CA: AAAI Press.

Conklin, D. 2003. Music generation from statistical models. In *Proceedings of the AISB Symposium on Artificial Intelligence and Creativity in the Arts and Sciences*, 30–35.

Cope, D. 2006. *Computer Models of Musical Creativity*. Cambridge, Massachusetts: The MIT Press.

Dannenberg, R. B.; Birmingham, W. P.; Tzanetakis, G.; Meek, C.; Hu, N.; and Pardo, B. 2003. The MUSART testbed for query-by-humming evaluation. In *Proceedings of the International Conference on Music Information Retrieval*, 41–51.

Dannenberg, R. B. 1985. An on-line algorithm for real-

time accompaniment. In *Proceedings of the International Computer Music Conference*, 279–289.

Delgado, M.; Fajardo, W.; and Molina-Solana, M. 2009. Inmamusys: Intelligent multi-agent music system. *Expert Systems with Applications* 36(3-1):4574–4580.

Dickerson, K., and Ventura, D. 2009. Music recommendation and query-by-content using self-organizing maps. In *Proceedings of the International Joint Conference on Neural Networks*, 705–710.

Gervás, P.; Díaz-Agudo, B.; Peinado, F.; and Hervás, R. 2005. Story plot generation based on CBR. *Journal of Knowledge-Based Systems* 18(4–5):235–242.

Gervás, P. 2001. An expert system for the composition of formal spanish poetry. *Journal of Knowledge-Based Systems* 114(3-4):181–188.

Gfeller, K. 1990. Music, the language of emotions. In Unkefer, R., ed., *Music Therapy in the Treatment of Adults with Mental Disorders; Theoretical Basis and Clinical Interventions*. New York: Schirmer Books.

Koskenniemi, K. 1983. *Two-level morphology: A general computational model of word-form recognition and production*. University of Helsinki: Department of General Linguistics.

Kreitler, H., and Kreitler, S. 1972. *Psychology of the arts*. Durham, NC: Duke University Press.

Lewis, G. 2000. Too many notes: Computers, complexity and culture in voyager. *Leonardo Music Journal* 10:33–39.

Macedo, L., and Cardoso, A. 2002. Assessing creativity: the importance of unexpected novelty. In *Proceedings of the ECAI'02 Workshop on Creative Systems: Approaches to Creativity in Artificial Intelligence and Cognitive Science*, 31–38.

Macedo, L. 2001. Creativity and surprise. In *Proceedings of the AISB Symposium on Artificial Intelligence and Creativity in Arts and Science*, 84–92.

McCorduck, P. 2004. *Machines Who Think*. Natick, MA: A. K. Peters, Ltd., 2nd edition.

McKay, C., and Fujinaga, I. 2004. Automatic genre classification using large high-level musical feature sets. In *Proceedings of the Fifth International Symposium on Music Information Retrieval*, 525–530.

Monteith, K.; Martinez, T.; and Ventura, D. 2010. Automatic generation of music for inducing emotive response. In *Proceedings of the International Conference on Computational Creativity*, 140–149.

Oliveira, A., and Cardoso, A. 2007. Towards affective-psychophysiological foundations for music production. In *Proceedings of the 2nd International Conference on Affective Computing and Intelligent Interaction*, 511–522.

Pearce, M. T., and Wiggins, G. A. 2007. Evaluating cognitive models of musical composition. In *Proceedings of the 4th International Joint Workshop on Computational Creativity*, 73–80.

Pérez y Pérez, R., and Sharples, M. 2004. Three computer-based models of storytelling: BRUTUS, MINSTREL and MEXICA. *Knowledge-Based System* 17(1):15–29.

Rabiner, L. R. 1989. A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE* 77(2):257–286.

Rahman, F., and Manurung, R. 2011. Multiobjective optimization for meaningful metrical poetry. In *Proceedings of the International Conference on Computational Creativity*, 4–9.

Riedl, M. 2004. Narrative generation: Balancing plot and character. *Ph.D. Dissertation, North Carolina State University*.

Saygin, A. P.; Cicekli, I.; and Akman, V. 2000. Turing test: 50 years later. *Minds and Machines* 10(4):463–518.

Schlaug, G.; Marchina, S.; and Norton, A. 2008. From singing to speaking: Why patients with Broca's aphasia can sing and how that may lead to recovery of expressive language functions. *Music Perception* 25:315–323.

Wang, Y.; Kan, M.-Y.; Nwe, T. L.; Shenoy, A.; and Yin, J. 2004. LyricAlly: automatic synchronization of acoustic musical signals and textual lyrics. In *Proceedings of the 12th Annual ACM International Conference on Multimedia*, 212–219. New York, NY, USA: ACM Press.

Weizenbaum, J. 1966. ELIZA - a computer program for the study of natural language communication between man and machine. *Communications of the ACM* 9(1):36–45.

Wiggins, G. 2006. A preliminary framework for description, analysis and comparison of creative systems. *Journal of Knowledge Based Systems* 19(7):449–458.

Wigram, T. 2002. Indications in music therapy: evidence from assessment that can identify the expectations of music therapy as a treatment for autistic spectrum disorder (ASD): meeting the challenge of evidence based practice. *British Journal of Music Therapy* 16:11–28.

Wu, Y.-S.; r. Chu, W.; Chi, C.-Y.; Wu, D. C.; T.-H.Tsai, R.; and j Hsu, J. Y. 2009. The power of words: Enhancing music mood estimation with textual input of lyrics. In *Proceedings of the International Conference on Affective Computing and Intelligent Interaction*, 1–6.