Soundwriter: Real-Time Music Generation for Oral Storytelling through Emotion Mapping

Toshihisa Tsuruoka
Department of Music and Performing Arts Professions
New York University
tt1694@nyu.edu

Ashley Erika Muniz
Department of Music and Performing Arts Professions
New York University
aem767@nyu.edu

ABSTRACT

Soundwriter is a real-time emotion detection system implemented in MATLAB that takes a spoken story and analyzes its emotional dimensions via two primary processes: 1) semantic analysis and 2) prosodic analysis. The former process classifies emotions by labeling either an incoming stream of detected words or previously typewritten sentences on three emotional dimensions—arousal, valence, and dominance—while the latter process gathers emotional qualities by analyzing low-level audio features of a speaker's voice, such as global F0 (pitch contour), speech rate, and intensity (loudness). In this paper, we demonstrate a creative application using Max/MSP, where Soundwriter dictates the behavior of the music generation algorithm built in Max/MSP for live performance applications. In particular, we target children as the primary audience and explore ways in which Soundwriter can reinforce the traditional experience of oral storytelling through the lens of child development. Simultaneously, we introduce an interactive scoring system that stimulates a new way of composing music and enables composers to transform poems into interactive music notations.

Keywords

Poetry and Music, Max/MSP, Generative Audio, Real-Time Notation, Child Development

1. INTRODUCTION

1.1 A Brief Introduction to *Soundwriter*

Soundwriter is developed with the goal of enriching the oral storytelling experience. While there have been a few attempts to create an offline automated scoring system for audio stories, nearly none have investigated a real-time system that is able to classify emotions on both semantic and prosodic features of a narrated story, let alone dynamically generating original music that is catered to the story's detected emotions. Upon this notion, *Soundwriter* implements a real-time hybrid system in which the emotional state of a given story is analyzed from its lexical and auditory features. The lexical features of the story is analyzed based on the Affective Norms for English Words database [1] where emotionally salient words are given ratings on their arousal, valence, and dominance.

The auditory features are analyzed via the detection the emotionally charged prosodic features of the reader's voice such as pitch contour, speech rate, and intensity. These two analyses together influence the final classification of the story's emotional state. To do this, *Soundwriter* implements a real-time speech recognition system via machine learning using convolutional neural networks (CNN) as well as a low-level audio feature detection system.

Soundwriter is part of the ongoing project called Bookscape—a project that aims to augment the traditional book-reading experience as a whole by employing machine learning techniques via music information retrieval methods and a soundscape synthesis system. Further details will be published in future publications.

1.2 Children's Literature and Child Development

The practice of oral storytelling has long been commonplace among families and schools to entertain and educate children. By taking advantage of this tradition, where children's stories are narrated expressively and often comprise of clear transitions and a wide range of emotions [2], children's literature was an appropriate subject through which to demonstrate *Soundwriter*'s potential. For this demonstrative work, Muniz constructed the poem "A Naughty Bug Strolls…" in pursuit of optimizing children's enjoyment as well as showcasing *Soundwriter*'s ability to generate musical accompaniment according to the emotional state of the poem. We have constructed an immersive piece as a proof-of-concept where the poem and music co-exist to deliver cohesive emotional changes.

2. RELATED WORKS

2.1 Algorithmic Composition Based on Film Scripts

Kirke and Miranda's system, *TRAC*, attempts to automatically analyze film scripts and generate notated "musical sketches" in order to accelerate the compositional process of a film composer under time pressure. This analyzation process entails stemming script text (removing suffixes) as well as removing stop words such as conjunctions. The script is then scouted for emotionally salient words (on a character-by-character, section-by-section, or entire script basis) which are ascribed emotion ratings based on the *ANEW* database. The emotional ratings of the selected segment are then linked to musical ideas according to the "affective composition rules" which attempt to communicate the detected emotion ratings via musical expressions. For

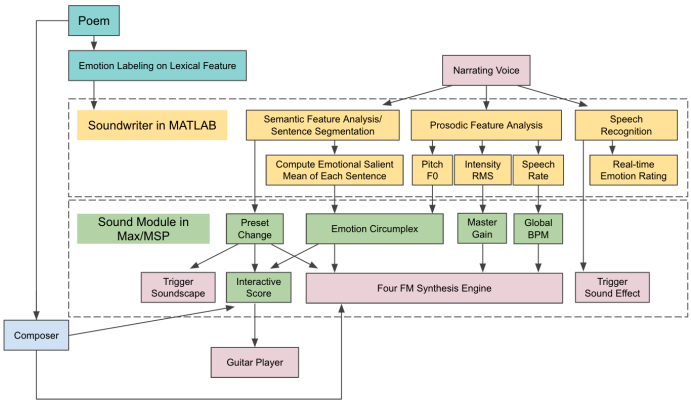

Figure 1. Overall flowchart of the proposed system.

instance, low dominance ratings correspond to low tempo in the music. Through this process, musical phrases are derived, eventually forming sections of music which provide the composer with a musical sketch to work off of [3].

2.2 Children's Perception of Emotions from Auditory Cues

From a child development perspective, [4] indicates that it takes more time for children to learn how to decipher emotions from prosodic expressions and musical expressions (rhythm and harmony) compared to other stimuli. By the age of 11, children are able to detect emotions from facial expressions as well as adults can, but their ability to decode emotions embedded within prosodic and musical expressions continues to emerge past this age. The study also reveals that these abilities emerge in conjunction with one another during a child's sound cognition development. The proposed system attempts to explore a new medium of entertainment for children, whereby prosodic and musical expressions come together to inform emotional changes alongside semantic cues from the story itself. In short, we hope to cultivate children's sound cognition skill to infer emotions by offering an emotive experience through the interweaving of three stimuli: 1) semantic meanings, 2) prosodic expressions, and 3) musical expressions.

2.3 Real-Time Notation Systems for Live Performance

Composers have been implementing real-time music notation systems for live performance, unveiling the creative possibilities within a computer-human relationship. Karlheinz Essl's *Champ d' Action* (1998), for example, utilizes stochastic algorithms that generate graphic notations (images, symbols, and text) for each performer to view individually on computer screens throughout the performance. The conductor triggers cues within the software which informs the performers to be active, resulting in an improvisational piece. Similarly, Kevin Baird's *No Clergy* (2005) employs stochastic algorithms to generate conventional staff-based notations for each performer and requires the audience to vote on parameter values (such as dynamics and rest duration) that affect the algorithm via web browsers. Wulfson, Barrett, and Winter's *LiveScore* (2007) also implements stochastically generated music notation that is affected by audience participation (adjusting knobs on MIDI controllers) and is wirelessly sent to each musician's laptop. Nick Didkovsky's *Zero Waste* for piano and computer (2002) takes on a more dynamic real-time notation experience, stochastically generating a score in which notes emerge, disappear, form melodies, and combine into chords in response to MIDI data generated by the pianist's sight-reading of the generated score [5].

3. IMPLEMENTATION

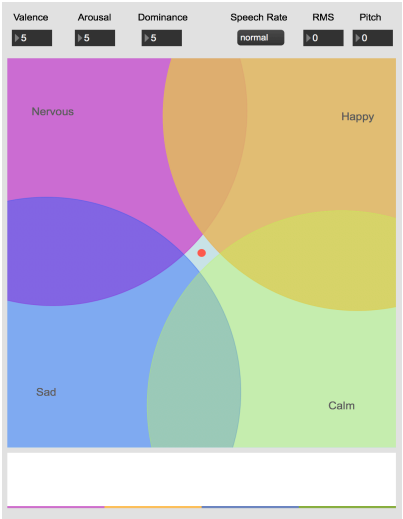

Figure 2. Max/MSP interface inspired by Russell's two-dimensional emotion circumplex [6].

3.1 Concept Behind Sound Generation

The goal of the sound generation is twofold: 1) to enhance the experience for children, and 2) to stimulate compositional possibilities when composers design and organize sounds that accompany narrative storytelling. This demonstrative work is not a pursuit to build a fully automated scoring machine. Rather, the collection of data acquired from *Soundwriter* is to be used freely through composers' visions and skills. And ultimately, this system provides composers a way to transform poems and stories into musical ideas capable of aiding a live performance.

In the proposed system, there are four sources of sound, in addition to the narrating voice: 1) live guitar, 2) four FM (frequency modulation) synthesis engines, 3) soundscape ambience, and 4) sound effects. This section illustrates

how the *Soundwriter*'s emotion classification and speech recognition algorithms influence the behaviors on each of these sound source in Max/MSP. Figure 1 summarizes the overall signal flow and data sonification directories.

3.2 Four Representative Emotions

As shown in Figure 2, Russell's two-dimensional emotion space is transformed into what is essentially a graphic representation of musical possibilities. And this emotion circumplex dictates the way in which the interactive guitar score is generated as well as the behavior of the four FM synthesis engines. The red dot represents the current emotional state of the spoken story, moving through the space as different sentences arrive and dictate their *valence* (along the x-axis) and *arousal* values (along the y-axis) from *Soundwriter*'s emotion classification algorithm. Four representative areas (*Happy*, *Sad*, *Calm*, and *Nervous*) convey general emotions that represent each area's surrounding emotions as described in Russell's model. For instance, the *Nervous* area respectively represents emotions such as *angry*, *stressed*, *frustrated*, etc. In Russell's model, these four representative emotions are located across the space equally distanced from each other, hence being the ideal choices for representation in this system [7].

The rationale behind this generalization is that the semantic and prosody based emotion classification systems such as *Soundwriter* still embodies inherent difficulty of classifying emotions into specific emotion categories. This is due to multiple reasons such as the lack of available training dataset that is appropriately labeled, the lack of complex semantic understanding in computational models, and an inevitable disagreement among our perception of emotions [8]. Though for the future implementation we strive to develop a more comprehensive emotion classification system, we find that the way our poem reacts with the four representative emotions is fairly accurate to our natural perception, especially thanks to the nature of our poem which was written to suit children, leaving out ambiguous qualities such as emotion negation and complex metaphor.

3.3 F0 Pitch Analysis

Among many studies such as [8] and [9], it is well-observed that the contribution of a global F0 (fundamental frequency) contour over utterance level (duration of one utterance) represents the arousal magnitude of perceived emotions. The overall increase in the global F0 mean and range corresponds to high arousal emotions such as *happy*, *anger*, and *fear*, while the overall decrease corresponds to low arousal emotions such as *sad* and *tenderness* [10]. However, the association between the F0 contour and a specific emotion category can hardly be defined. Nonetheless, existing studies such as [11] indicate that pitch analysis, independent from linguistic examination, can strongly suggest the magnitude of emotional arousal and is the strongest indicator of emotion amongst other audio features.

Based on this premise, *Soundwriter* calculates F0 mean and range value on a sentence level during oral reading. This measurement influences the current position of the red dot along the y-axis (*arousal* level). When the narrator finishes reading a sentence, he/she clicks the "New Sentence" button on the *Soundwriter* interface (see Figure 3). This action informs *Soundwriter* of the duration of that sentence segment, enabling *Soundwriter* to measure F0 mean and range values of a sentence. The red dot's current location, which represents the current sentence's ratings, is then briefly re-positioned to account for the narrator's expression. Once this re-positioning is completed, the red dot finally settles into the new sentence's emotional ratings. The magnitude of the re-positioning is determined by amount that the F0 values deviates from the reader's predetermined neutral F0 values derived from his/her non-emotional vocal expression.

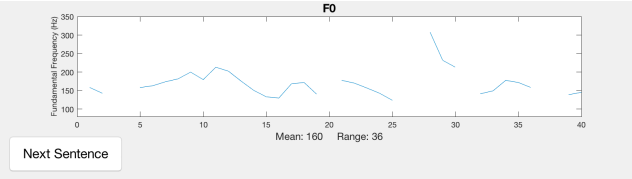

Figure 3. Real-time F0 pitch measurement and "New Sentence" button in *Soundwriter*

3.4 Interactive Real-Time Interactive Notation for Guitar

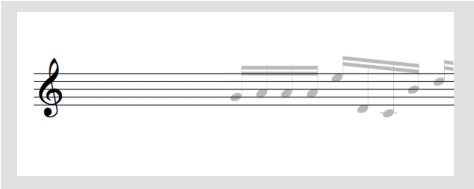

Figure 4. Real-time interactive notation for guitar.

In this demonstrative work, Tsuruoka composed four distinct musical ideas for guitar, each corresponding to one of the four representative emotions as well as reflecting the emotional fluctuation embedded within the poem. As the story progresses, the red dot moves, reflecting the emotional state of the story. Simultaneously, this movement triggers musical figures to flow across the computer screen from right to left (see Figure 3). The stronger one emotion is (represented by the red dot nearing the center of one emotion area), the brighter and slower the corresponding music figure emerges, and vice versa. When two emotion areas collide, the performer sees two musical figures on screen—one flowing across faster and dimmer (weak emotion) and one flowing across slower and brighter (strong emotion). In this way, the interactive music notation foreshadows the information available on the emotion circumplex, enabling the performer to react on conventional music notations rather than requiring him/her to interpret designated musical figures by looking at the circumplex.

3.5 Four FM Synthesis Engines

Four identical FM synthesis engines with built-in sequencers were made to represent emotion areas (*Happy*, *Sad*, *Calm*, and *Nervous),* each with a designated sound design

and sequencer pattern saved as a preset. Similar to the guitar music figures, these presets are made in advance so that the composer's interpretation of the poem is best represented through this sound synthesis. The computational efficiency and timbral versatility of FM synthesis design makes it the ideal choice for this system.

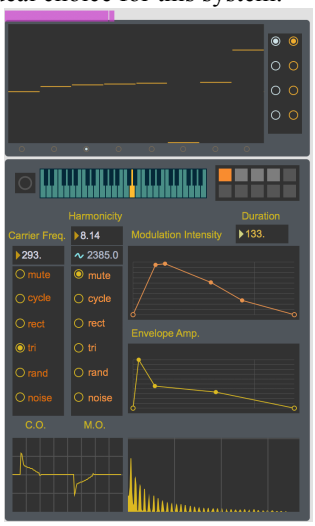

Figure 5. FM synthesis engine in Max/MSP with a sequencer representing *Nervous* emotion area.

The movement of the red dot dictates the loudness of each FM engine. As the red dot nears the center of one emotion area, the output volume of the corresponding FM engine increases. When the dot recedes from one emotion area, its FM engine becomes silent. In addition, a root mean squared (RMS) level extracted from the narrator's voice is mapped to the master gain stage of all FM engines. Therefore, the sound synthesis system parallels the loudness of the narrator's voice.

In addition, the four engines receive various divisions of the master tempo as follows:

Master Tempo 'M'	50 BPM (variable)
Happy Area	'M' x 2 BPM
Sad Area	'M'
Calm Area	'M' x 1/3 BPM
Nervous Area	'M' x 1/5 BPM

Table 2. Tempo mapping on each FM engine.

Several studies such as [12] and [13] indicate that one of the earliest auditory cues from which children infer emotions is tempo. Fast tempi inform happiness while slow tempi suggest sadness. This correlation develops before the ability to hear musical modes (specific series of pitches) and their affiliated emotions.

When two emotional areas collide, the tempi from each area are layered to create polyrhythmic textures. This creates an interesting musical effect. For instance, when *Nervous* and *Happy* overlap, a hemiola rhythm (2 against 3) is produced. Furthermore, the master tempo is manipulated by the detected speech rate which is categorized as either *normal*, *slow*, or *silent*. When *normal*, the master tempo is static; when *slow,* the master tempo begins to incrementally decrease; and when *silent*, the master tempo decreases in a large increment. Although this metric configuration creates interesting rhythmic patterns, one could argue that it produces a robotic feel due to its mechanical nature. Future investigation is needed in order to translate the detected emotion ratings into appropriate musical motives and ideas as previously explored in research such as [3]. At this early stage, we focus on the real-time detection of emotions from a given story while most of the compositional procedures are in the hands of the composer. By allowing him/her to create the FM synthesis sequences (including timbre, dynamics, and pitch) as well as guitar figures from his/her free interpretation of the emotions within the story, we circumvent the music from becoming mechanically dull.

3.6 Speech Recognition for Sound Effects

Lastly, *Soundwriter* utilizes a CNN-based speech recognition system. In this demonstrative work, this is used to detect cue words that trigger sound effects. For instance, whenever the narrator reads the word "cut", the corresponding sound effect of a scissor noise is triggered, creating an interactive storytelling experience. In the future, we plan to investigate in building a more robust speech recognition system that can transcribe the spoken story completely on the fly, enabling us to dismiss the necessity of inputting a story as text.

4. DISCUSSION

The compositional methods discussed in this demonstrative work brings forth a notable difficulty concerning the linear structure of the resulting music. Particularly, we observe that the lack of rests (or pauses) causes the resulting music to have little sense of structure. This is due to the fact that *Soundwriter* only takes into account the momentary emotional state of a sentence segment, each associated with pre-composed FM synthesis patterns and guitar figures that always carry sound. Although the detected loudness of the voice is mapped to the overall loudness of the music, this dynamic change does not suffice when it comes to creating transitional moments in music that would contribute to a better sense of linear form and structure. In addition, it is difficult for the composer to anticipate which emotional transition (therefore musical transition) will be attained by *Soundwriter*. This constrains compositional possibilities by limiting the composer to musical figures that are closely related to each other and therefore transition smoothly regardless of the transition pattern. In other words, four pieces of music representing four emotions do not create four distinct musical units but rather create monothematic music as a whole with four components. This is another factor that prohibits the resulting music from being structurally rich. Perhaps creating a more comprehensive algorithm that is capable of analyzing the changes in a story's scenes or setting, beyond analyzing the emotional state, will help provide music with transitional cues from the story.

Although we employed four generalized emotions to represent the entire emotion circumplex (originally containing 28 emotion categories [6]), more resolution in emotion categorization may promote a better representation of the

story's varying emotions. Simultaneously, implementing a regime that translates detected emotion ratings into appropriate musical figures (emotionally relevant modes, tempi, timbre, etc.) can result in a better musical representation of the detected emotions [3]. Also, there is room for improvement when it comes to the interplay between speech and music. Developing a system that implements proven practices such as musical underlays [14], which glue the gap between a spoken story and its musical accompaniment, can be an interesting avenue.

Also, it is worth mentioning the practical difficulty of preventing the generated music from masking the speech recognition performance on a device where the speakers and microphone are located near one another.

5. FUTURE WORKS

Many exciting fields of application are expected for this type of speech-to-sound synthesis system. First and foremost, we are interested in developing a similar system that is specifically suited for parents who read stories to their children at bedtime. In addition, there are many potential applications as a non-pharmacological approach to child development. For instance, this technology can be developed into a sedative music creation device for infants, or into an alternative methodology to speech therapy for children via a modified speech recognition system employing interactive musical feedback.

At its core, the adaptable nature of this system stimulates various areas of interest thanks to its ability of extracting the expressional dimensions from semantic and prosodic expressions in real-time. Some of the prospective applications in the arts include: generative audio compositions, installation works, digital poetry, and automated film scoring.

6. CONCLUSION

A real-time software system named *Soundwriter* was introduced with the goal of enriching the oral storytelling experience by integrating music that is in accord with the emotional fluctuations of a story. This system is implemented in MATLAB and takes a written and spoken story in order to detect its emotional state. The lexical features of the story are analyzed for their affective qualities (*arousal*, *valence*, and *dominance*) with the *ANEW* database [1] while the auditory features (pitch, speech rate, and dynamics) are analyzed via a low-level audio feature detection system. Simultaneously, a compositional method that utilizes *Soundwriter*'s data was introduced using Max/MSP. The system organizes the detected emotions into four emotion areas on a circumplex which then dictates the way in which the pre-composed musical ideas, representing these emotion areas, react in real-time while the reader's speech rate and dynamics influence the final result. This automated data sonification process also integrates an interactive notation system which translates the emotion circumplex into traditional music notation for live performance applications.

Soundwriter was also developed to promote children's sound cognition skill of inferring emotions by offering an emotive experience through the interweaving of three dimensions of stimuli: 1) semantic cues, 2) prosodic expressions, and 3) musical expressions. Long-term research is needed for further investigation and development.

7. REFERENCES

[1] M. M. Bradley and P. J. Lang, "Affective Norms for English Words (ANEW): Instruction Manual and Affective Ratings," 1999.

[2] C. Alm, R. S.-N. E. C. on Speech, and undefined 2005, "Perceptions of emotions in expressive storytelling," *isca-speech.org*.

[3] A. Kirke and E. Miranda, "Aiding Soundtrack Composer Creativity through Automated Film Script-profiled Algorithmic Composition," 2017.

[4] D. Vidas, G. A. Dingle, and N. L. Nelson, "Children's recognition of emotion in music and speech," *Music Sci.*, vol. 1, p. 205920431876265, Jan. 2018.

[5] J. Freeman, "Extreme Sight-Reading, Mediated Expression, and Audience Participation: Real-Time Music Notation in Live Performance," 2008.

[6] J. R.-J. of personality and social psychology and undefined 1980, "A circumplex model of affect.," *psycnet.apa.org*.

[7] S. Rubin, M. A.-P. of the 27th annual A. symposium, and undefined 2014, "Generating emotionally relevant musical scores for audio stories," *dl.acm.org*.

[8] O. Kwon, K. Chan, J. Hao, T. L.-E. E. Conference, and undefined 2003, "Emotion recognition by speech signals," *isca-speech.org*.

[9] M. S.-S. E. C. on Speech and undefined 2001, "Emotional speech synthesis: A review," *isca-speech.org*.

[10] T. Bänziger, K. S.-S. communication, and undefined 2005, "The role of intonation in emotional expressions," *Elsevier*.

[11] T. Vogt, E. André, and J. Wagner, "Automatic Recognition of Emotions from Speech: A Review of the Literature and Recommendations for Practical Realisation," in *Affect and Emotion in Human-Computer Interaction*, Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 75–91.

[12] S. D. Bella, I. Peretz, L. Rousseau, N. G.-Cognition, and undefined 2001, "A developmental study of the affective value of tempo and mode in music," *Elsevier*.

[13] J. Bruce Morton and S. E. Trehub, "Children's judgements of emotion in song," *Psychol. Music*, vol. 35, no. 4, pp. 629–639, Oct. 2007.

[14] S. Rubin, F. Berthouzoz, G. Mysore, … W. L.-P. of the 25th, and undefined 2012, "UnderScore: musical underlays for audio stories," *dl.acm.org*.