AN INVESTIGATION INTO CROSSMODAL AND MULTIMODAL HUMAN COMPUTER INTERACTION: HANDWRITING, MUSIC GENERATION AND BODY GESTURE

TANG WAI WA

MPhil

The Hong Kong Polytechnic University

2015

THE HONG KONG POLYTECHNIC UNIVERSITY

DEPARTMENT OF COMPUTING

An Investigation into Crossmodal and Multimodal Human Computer Interaction: Handwriting, Music Generation and Body Gesture

Tang Wai Wa

A thesis submitted in partial fulfilment of
the requirements for the degree of
Master of Philosophy

January 2015

CERTIFICATE OF ORIGINALITY

I hereby declare that this thesis is my own work and that, to the best of my knowledge and belief, it reproduces no material previously published or written, nor material that has been accepted for the award of any other degree or diploma, except where due acknowledgement has been made in the text.

(Signed)

TANG Wai Wa (Name of Student)

ABSTRACT

This thesis presents an investigation into cross-modal and multimodal human computer interaction. Human interaction is by nature multimodal and crossmodal. Human beings communicate through multiple channels of speech, gestures, facial expressions, body language, etc. In addition, cross-modal transformations, where ideas and feelings are expressed in multiple modalities, are fairly common in art. However, these modes of expression and communication have so far been mostly ignored in human computer interaction.

We focus on the transformation of input modalities to alternate output modalities. The primary focus is on the transformation from writing to music, specifically Chinese calligraphy to Chinese music. We investigate both real-time as well as offline modes of transformation, and experiment with different levels of human control versus statistically driven modeling. We present evaluations that show that the music that is generated is both pleasing to the ear as well as correlates with the Chinese style of music.

As an additional investigation, we also consider two other cross-modal transformations: the mapping of body motions and gestures to music manipulation operations, and the mapping of handwriting to graphics. Both transformations are motivated and demonstrated with real-life applications; preliminary evaluations show that they can encourage interaction, stimulate creativity as well as serve as a motivation to practice the input modality.

PUBLICATIONS ARISING FROM THE THESIS

1. Lo, K. W.; Lau, C. K.; Huang, M. X.; Tang, W. W.; Ngai, G. & Chan, S. C. Mobile DJ: a Tangible, Mobile Platform for Active and Collaborative Music Listening. In Proc. of International Conference on New Interfaces for Musical Expression (NIME), 2013, 217-222.

2. Tang, W. W. W.; Chan, S.; Ngai, G. & Leong, H. V. Computer Assisted Melo-rhythmic Generation of Traditional Chinese Music from Ink Brush Calligraphy. In Proc. of International Conference on New Interfaces for Musical Expression (NIME), 2013, 84-89.

3. Tang, W. W. W.; Leong, H. V.; Ngai, G. & Chan, S. C. F. Detecting Handwriting Errors with Visual Feedback in Early Childhood for Chinese Characters. In Proc. of the 2014 Conference on Interaction Design and Children (IDC), 2013, 273-276.

ACKNOWLEDGEMENTS

There are many people I would like to thank for helping and guiding me throughout this process.

First and foremost, I would like to thank my supervisors, Dr. Stephen Chan and Dr. Grace Ngai. There has been tough time and they provided valuable advice and support on the thesis and my personal issues. It has been my great pleasure to work under their supervision, I have learnt a lot.

I would also like to thank Dr. Leong Hong-va and Dr. Alvin Chan, for their kind and valuable comment and advice on my study. Also a big thank you for colleagues and friends in the eToy Teaching Lab, it has been fun and productive working with them. I know I still own all of you a big meal.

I would like to express my special thank to Mr. Sandy for his generous support of my study. Without him I would not have the chance to step into the world of academia and discover a world full of wonders. I hope I will contribute to the society just as the way he do.

Last but not least, I would like to thank my family and friends, for their unconditional love and support in all the moment of my life.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

1. INTRODUCTION

Everyday human communication is multi-dimensional. Language, gestures, facial expressions and body language all play a part in communicating information from person to person. Human-Computer Interaction (HCI), however, has mostly been uni-dimensional so far; humans usually communicate with the computer via the keyboard or the mouse, and only one point of input is used at any one time. In recent years, multimodal HCI has gained much interest, as it is becoming increasingly evident that a system that allows interaction beyond monitor, keyboard and mouse, such as touch screen, voice recognition and various sensory inputs like motion and location cab provide for a richer user experience. However, even though the range of input media has grown, the output to the user is still very limited in common usage, primarily consisting of text and graphics, with the occasional notification sound and vibration. Multimodal human-computer interaction has therefore mostly focused on the study of input modalities and the fusion of these inputs. Study on the output aspect, which is usually text-based with graphical elements, usually enters around the addition of modalities, such as auditory and tactile modes, as redundancy channels.

We hypothesise that the transformation of these input modalities into output of other forms may result in a synergistic effect that seems more natural to humans. In particular, we focus on the aesthetic and artistic domains. There are traditions dating back to ancient times, that different art forms are closely related and can be transformed into each other. Therefore, in this study we seek to explore whether this aforementioned transformation can be exploited for HCI.

One of the strongest relationships between art forms is believed to be Chinese calligraphy and music. (Qian and Fang 2007) Some of these beliefs can be evidenced in popular media. In movies, whenever there is a scene involving Chinese calligraphy or Chinese painting, traditional Chinese music is inevitably playing in the background. The same is also true for dancing and music. This suggests that all of these art forms stimulate and deploy multiple channels of personal expression.

For some users, using textual output as the only output may not be the most efficient. For example, when children are learning to write, presenting their mistakes in the form of graphical representation may make more sense to them, especially since Chinese characters originate from drawings of real objects. Other modalities may also be useful for tasks other than providing information, e.g. art, creativity inspiration.

1.2. Objectives

Modality transformation has not been widely studied, and this thesis seeks to fill in some of the gaps that are lacking in the current literature. The aim is to provide insight and case study to help understanding the methodology and difficulty in implementing such transformation. We investigated different aspects in the transformation of input to output in multimodal human-computer interaction, and developed practical mappings based on the concept of metaphoric congruence. The mappings have been implemented and developed as working prototypes. Evaluation methods for each mapping have been proposed and carried out to facilitate further design and research in the field.

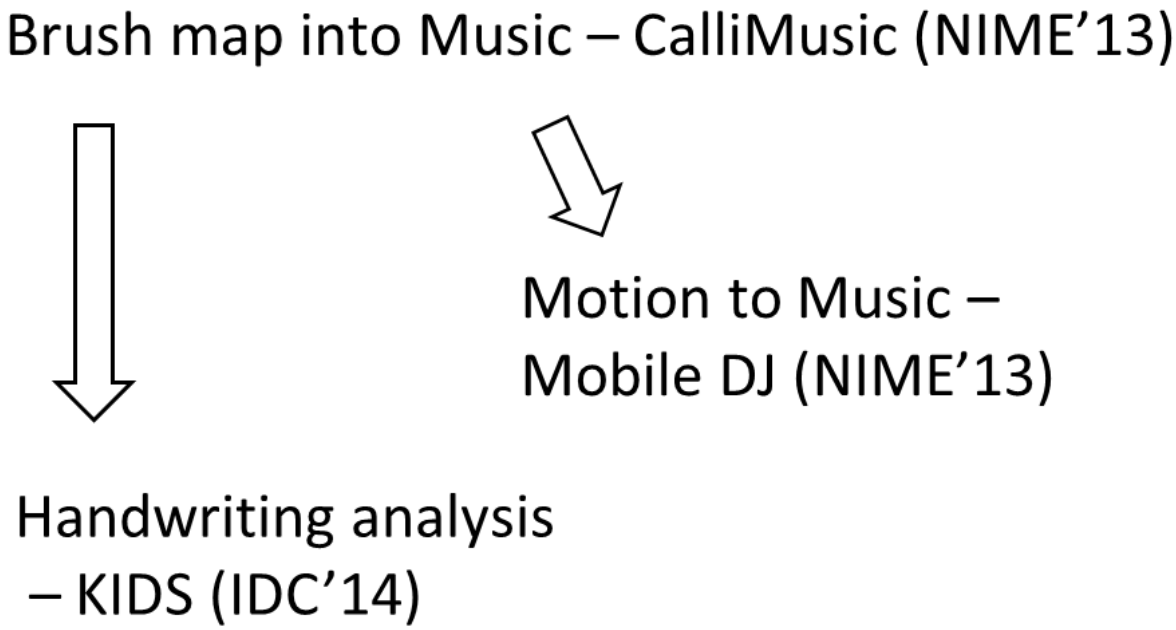


Figure 1.1. Publication diagram.

1.3. Organization of the Thesis

The thesis is organized as follows. Chapter 2 describes the current state-of-the-art work related to this study. The three related topics explored in this thesis are presented in Chapter 3 to Chapter 5. Chapter 3 describes a study of the transformation of writing to music, which was the first area we investigated in transforming multimodal human-computer inputs to alternate feedback. It describes a project called CalliMusic that convert brush painting into Chinese classical music, which was accepted into NIME conference in 2013. Having the hand-in experience of music generation, in

Chapter 4 we further explore the relationship of human body motion and music, together with social elements. This project, MobileDJ, was also accepted in the same conference. Chapter 5 describes an application that utilizes our handwriting analysis knowledge and modality transformation — a tool for training handwriting in children, that visualizes the correctness of their writings as an alternate feedback. It has been presented in the IDC conference in 2014. Chapter 6 provides a summary for the whole thesis.

2. BACKGROUND AND LITERATURE REVIEW

2.1. Human-Computer Interaction

Human-computer interaction is critical to the success of consumer electronics, mobile devices and home entertainment. It involves the study, design, implementation and evaluation of the interface between user and computer, minimizing the barrier between a human's mental model and a computer's structural operation procedure. With the advance in the processing power of modern computer, the goal of computing have shifted from people learning to coupe with computer to computer being designed to help people to deal with everyday problem. The change from command line to graphical user interface (GUI), from mouse to multi-touch interface, from keyboard to voice input, each of the advance provides a more natural interface for human to work with the computer to increase efficiency. The main focus of HCI research are designing new methodologies, prototyping and evaluating new devices, proposing new interaction model and theories.

2.2. Multimodal Human-Computer Interaction

A multimodal system can be defined as one that supports communication through multiple modalities or different types of communication channels. Modes of communication in multimodal human-computer interaction typically are visual (text, graphics) and audio (speech, sound), with non-verbal information (touch, hand gesture, head orientation, face recognition, gaze, etc.) complementing for a more flexible, more efficient or less ambiguous interaction. Daily human-to-human interaction also usually involves more than one mode of communication, such as facial expressions, or emotions expressed in the change of tone and/or hand gestures. Therefore, multimodal human-computer interaction can be referred to as an "interaction with the virtual and physical environment through natural modes of communication". (Bourguet 2003) Since the introduction of "Put that there" demonstration that integrates hand gesture and speech, the field of multimodal human-computer interaction has expanded rapidly.

Along with the study in alternate input and output modality, one focus od multimodal human-computer interaction is the combination of input modalities to increase usability, and the other is the mapping of information to various output modalities to achieve synergy and redundancy effect.

To design such robust and well-integrated systems, many disciplines have to be involved, for example computer science, software engineering, social science, psychology, human factors, ergonomics, design and linguistics. These requirements pose a challenge to multimodal system development. In light of that, we have previously developed a general multimodal framework for easier prototyping with different devices, i.e. i*Chameleon (Lo, Tang et al. 2012), that we will use in most of the projects in this paper.

2.3. Modality Transformation

There has been much previous work on combining human motion or activity with sound to create a new art form. Some involve intentional motions: for instance, in the form of tangible interaction (Sergi, Günter et al. 2007) 2D drawing (HyperScore) or 3D spatial interaction (Sonic wire sculptors); other approaches use statistical models to generate music corresponding to user interactions or signals (Zeljko 2005, Ian, Dan et al. 2008, Eric, Dan et al. 2009).

There have been some efforts on generating music by combining tangible and digital media. For example, there have been some efforts on generating music in response to brushstrokes. DrawSound (Kazuhiro 2008) explicitly maps the position and pressure of brushstrokes on a conductive surface to sound frequencies. The Hé (Kang and Chien 2010) and MelodicBrush (Michael Xuelin, Will et al. 2012) systems generate music corresponding to Chinese calligraphic strokes. For motion-based music generation, there has been study on using motion data of sensors on skeleton joint positions of a performer to facilitate live sound transformation. (Fuhrmann, Kretz et al. 2013)

2.4. Tangible Interface

Tangible interfaces are an inspiring research topic for many disciplines, especially in human computer interaction, as they can effectively facilitate intuitive and direct manipulation. CoolMag (Cheng, Li et al. 2011) applied a series of magnetic sensors to detect the touch of a user's fingers, which allowed users to adopt objects in daily life as carriers of musical instruments. Ubiquitous Drums (Boris and Mark 2010) embedded multiple force sensors on clothes, which enabled the user to generate drumbeats using percussive gestures. These applications focused either on the design of a particular instrument or on a single user's experience without the involvement of social collaboration. Freqtric Drums (Tetsuaki, Taketoshi et al. 2007) used electric current to turn the audience into drums. It required all the users to hold the device, which implies its necessity for face-to-face interaction. Similarly, Jam-O-Drum (Tina and Tim 2000) was an interactive musical system, which allowed multiple users to play and collaborate simultaneously. However, it was based on sharing a single surface, and not intended to facilitate remote social interaction.

2.5. Music

2.5.1. Music Generation

Computer-generated music is not new. A wide range of techniques have been used already, with varying degrees of success. The involvement of computer in music composition started as early as 1963, where Hiller and Robert Baker developed Musicomp, the first computer-assisted environment for music composition (Nierhaus 2009). In 1970, Max Mathews and Richard F. Moore developed the GROOVE (Generated Real-time Output Operations on Voltage-controlled Equipment) system, which is the first music synthesis system that can be controlled and perform in real-time by interacting with human (Mathews and Moore 1970).

2.5.2. Markov Model

Probability calculus and statistics have been used in the field of automatic music generation early on (Jones 1981). Markov model (Camurri 1995) and the n-gram technique borrowed from natural language processing (Ponsford, Wiggins et al. 1999) were applied to generate music statistical model with success.

Jazz chord sequences generation (Steedman 1984), along with identification and generation of hierarchical structure in chorale melodies composed by J.S. Bach (Craig and Ian 1997) have proofed the success of this approach. Novel jazz solos can be generated by automated learning of stochastic context-free grammar. Another powerful method for algorithmic music composition and analysis is generative grammar (Ponsford, Wiggins et al. 1999).

2.5.3. Chaos System

Recently, Lindenmayer systems or L-systems, named after the botanist Artistid Lindenmayer developed a formal language to represent the growth of plants (Lindenmayer 1968), have been used as the basis for many music composition applications due to its randomness and at the same time its similarity. Many form of pictorial transformation has been studied to map generated image to music. Many of the techniques can be used to transform the pictorial form of writing into music.

2.5.4. Ink Brush Stroke to Music Transformation

Generating music from Chinese ink brush strokes (or writing in general), on the other hand, is quite new, and there have been few related work. DrawSound (Kazuhiro 2008) generates sound frequencies with the writing of ordinary paper and conductive brush on top of a multi-touch input surface. However, it uses a special brush which is not designed for traditional Chinese calligraphy.

The Hé system (Kang and Chien 2010) takes a closer idea that it generates musical notes with ink brush painting. However, it is designed to accept arbitrary brush movements and does not take the stroke type or the character into account.

2.5.5 Body Movement to Music

On the other hand, body motion and music are tightly tied together. Dancing and DJ scratching are good examples of people performing both art form at the same time. Therefore it is possible to transform the input of one form into another, most likely in real time. Active music listening (Jack and Philippe 2009) enables users to interactively mold and modify the music they are listening to in real time. Such interactions allow the users' active participation to influence the music listening experience. The expectation is that this interaction with the music would lead to stronger engagement and musical embodiment on the part of the user. Mobile and wearable devices have become almost pervasive in recent years. These devices offer the possibility of social and collaborative interaction in different locations and situations. Riding on the success of mature sensory systems and well developed hardware technologies, mobile devices nowadays possess processing power that is capable of handling multimodal and multimedia interactions. This integration of mobile devices offers the possibility of an embodied dimension and richer social network interaction. Therefore, it makes it possible for the user to transform the art of body motion to music.

2.5.6. Active Music

There has also been much work in active music listening, deploying various techniques such as such as computer vision and wearable sensors (Demey, Leman et al. 2008, Godbout and Boyd 2010). Examples of interactive sonification systems include gesture-based music synthesis as well as motion-based synchronization (Wanderley and Depalle 2004). Active music has also been used for dance music generation (Camurri 1995) and sport training (Effenberg 2005, Godbout and Boyd 2010, Eriksson, Halvorsen et al. 2011).

Research into social interaction through active music listening is still in its infancy. Investigations originating from the low level (e.g. motion rhythm) single user-centric active listening has recently focused on the listening interaction between multiple users and the corresponding high level expressive and emotional processes. Stockholm and Pasquier (2009) applied reinforcement learning to enhance the users' mood and mixed audio to increase collaboration and empathy among the users. Varni (2009) noticed a positive correlation between phase synchronization and social interaction empathy and dominance. They also developed a multi-channel musical system to explore the effectiveness of user centric media on mobile phones in affecting the social interaction and synchronization (Varni, Mancini et al. 2011). Three different interactive sonification methods, generating music according to the users' motion coordination metric, were introduced and compared (Varni, Dubus et al. 2012). By linking the musical attributes with the human motion, researchers observed a correlated effect of the audio stream on users' synchronization of rhythmic motions, entertainments, empathy, and collaborative social behaviors.

2.5.7 Social Interaction

Interactional synchrony covers the elements of temporal coordination, interpersonal communication (Burgoon, Stern et al. 1995) and emotional contagion (Hatfield, Cacioppo et al. 1994) in social interaction. It is believed to have a strong connection with rapport for nonverbal human interaction, which is crucial for trust, closeness, and acknowledgement (Scheflen 1964). Interactional synchrony, which originated in psychology, has been applied in human robot interaction (Andry, Gaussier et al. 2001, Marek, Selma et al. 2007, Meisner, Sabanovic et al. 2009). Motion rhythms were used as an important way to enhance the interactional synchrony during interaction. Michalowski et al. (2007) developed a rhythm-driven robotic system, which provided the user with visual feedback of synchrony during the interaction, demonstrating the enhancement arising from interactional synchrony. The study of Shadowplay, a robot human interface trained with real human-human interaction data, showed the effectiveness of interactional synchrony, imitation and correlated sequences of behaviors during interaction (Meisner, Sabanovic et al. 2009). Other applications also indicated the value of interactional synchrony. Gait alignment has been detected and augmented to facilitate interactional synchrony during mobile phone conversation (Robert 1992). Balaam (2011) introduced an ambient display to enhance the interactional synchrony, which was proved to have an influence on the participants' nonverbal behaviors.

We believe by investigating the designs of social interaction factors, we can apply the positive effects on social interaction from tangible interaction, active music listening, and interactional synchrony, to establish a collaborative social music network that would encourage social bonding and engagement.

2.6. Chinese Culture

2.6.1. Chinese Characters

Oriental languages, and Chinese in particular, possess a complex structure. While Latin languages are composed of a linear sequence of basic alphabetical components, written Chinese exhibits a two-dimensional pictographic composition nature, in which the basic ingredients, strokes, can be spatially arranged in many different ways.

2.6.2. Childhood Education in Chinese Writing

Early childhood education is an important sector of education. In Hong Kong, an increasing amount of public educational resource has been allocated towards early childhood education. The availability of ubiquitous computing devices has also led to the development of increasing numbers of technology-enabled education methodologies and platforms, with the main focus being on languages (English and Chinese) and Arithmetic.

To maintain literacy, children learn to write at an early age. In Hong Kong, for example, formal reading and writing starts before elementary school, with most kindergarten children starting to learn penmanship as early as age 3 or 4. By the time they are in elementary school, children regularly spend as much as 50% of their time in handwriting tasks (Tseng and Chow 2000), and a failure to achieve handwriting competency has been shown to have long-term effects on self-esteem (Feder and Majnemer 2007), not to mention the impact on elementary study.

Figure 2.1. Sample of practicing Chinese writing from elementary school child.

The most common learning approach is learning by practicing: teachers ask students to write the characters on a gridded sheet of paper, such that each character falls neatly within one gridded square, as shown in **Figure 2.1**. They are “graded” according to the correctness of the written character, whether they exceed the boundary of the grid, whether they are too big or too small, or presence of any obvious irregularity. Grading by hand is tedious and oftentimes subjective and inconsistent. The feedback is also delayed, which means that children often practice incorrect methods of writing for extended periods of time before they are identified and corrected. Finally, the feedback inflicts negative feeling upon the children, especially when the evaluation is not positive.

There is certain software for children to practice Chinese character writing, in following an animated stroke sequence in a manner similar to practicing using the Latin copybook, and scoring in whether the stroke order is followed correctly.

2.6.3. Traditional Chinese Music

Most traditional Chinese music uses a pentatonic system, i.e. a scale with five notes in an octave. Traditionally these five notes are named as 宫 (Gong), 商 (Shang), 角 (Jue), 徵 (Zhi) and 羽 (Yu), which correspond to the notes Do, Re, Mi, Sol and La respectively in Western music. Traditional Chinese music does not have harmony, so there are no simultaneous notes being played at any given time. In other words, all notes are played one by one in series.

3. MAKING MUSIC FROM WRITING — CALLIMUSIC

Our first investigation into cross-modal and multimodal interaction involves the mapping between music and writing. CalliMusic captures user input in the form of brush motion in traditional Chinese calligraphy writing, without affecting the way the calligraphy is written - where a person uses a real brush with real ink to paint on a piece of paper. As the user writes, the system simultaneously analyzes the writing content and mechanics and generates Chinese music where the pitch and rhythm are determined by the written characters and writing intervals.

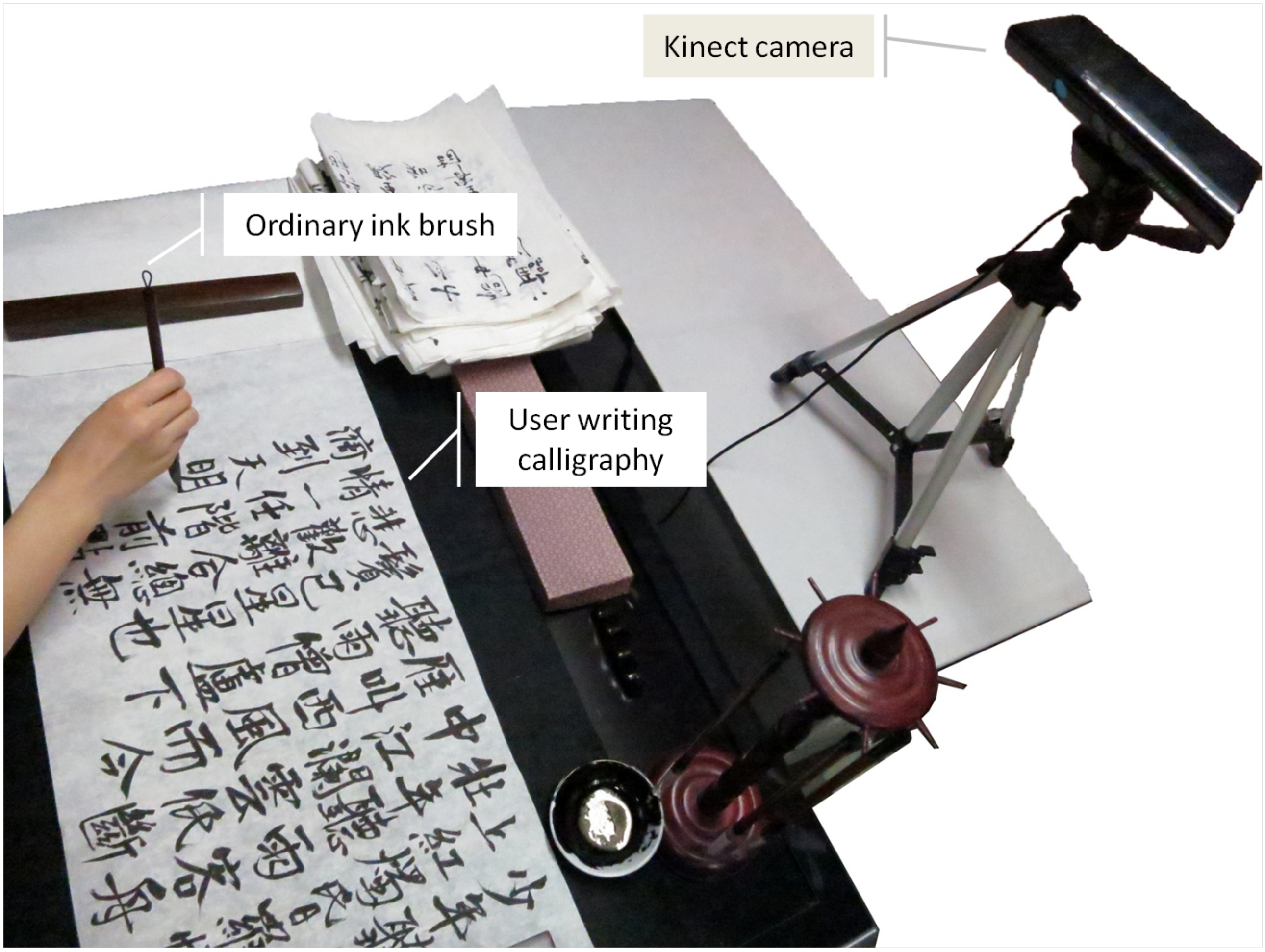


Figure 3.1. Writing platform: user writing on the paper, and a Kinect camera in front of the desk capturing the motion.

3.1. Architecture and System Design

The CalliMusic system utilizes the Microsoft Kinect depth camera to turn a non-sensing surface into a touchpad. The system consists of vision, graphics and music generation components. Users draw strokes on a screen surface with a normal calligraphic brush, and manipulate the drawing area with their free hand (see **Figure 3.1**). The depth camera captures the writing state of the user, including the position of the brush and the free hand. The brush effect is recreated in real time. At the same time, the melody is generated according to stroke parameters.

3.1.1. Vision-Based Modeling of Writing Mechanics

In the CalliMusic system, the fingertips modeled as rigid objects and the brush bristles as deformable objects. Background extraction is performed on the temporal filtered depth map obtained from the camera. The locations of the fingertips are then estimated by projecting the centroids of the convex points from the hand silhouette into 3D space. In the brush estimation, a point-based detector is used to approximate the brush handle in the depth map. Probabilistic Hough transformation is applied to form the brush contour, and Kalman filter to track the brush motion between frames.

3.1.2. Graphical Recreation

The user motions detected by the vision estimation component are processed into graphical feedback for the users. The writing pressure is modeled as a function of the distance from the brush-end to the writing surface and the brush orientation, and the strokes segmentation is determined by the brush bristles leaving the writing surface. To generate a stroke, the detected touch points are first smoothed using Gaussian filter, and then connected by polygons along the stroke trajectory. The size and transparency of the polygons are adjusted with changes in the writing pressure (as presented in **Figure 3.2**).

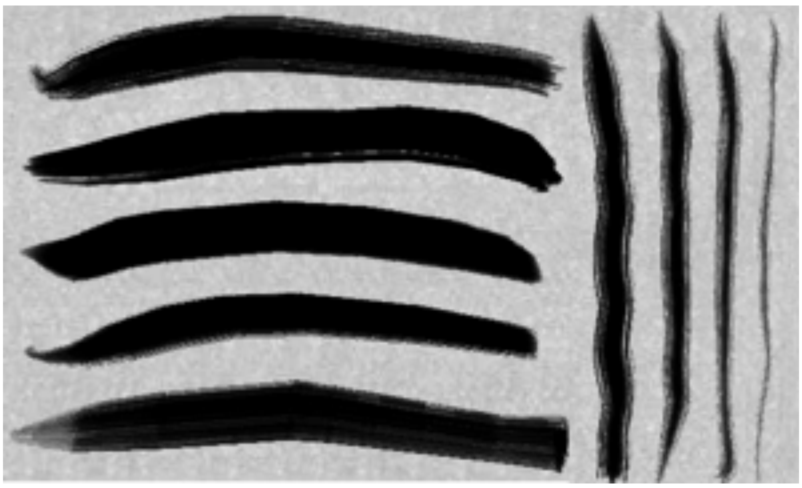


Figure 3.2. The brush stroke effects with different bristle textures and writing pressure levels.

Our graphical recreation system is not intended to completely replicate physical ink brush writing on paper. Instead, we aim to provide visual feedback to the calligrapher with digital signals. This allows us to create visual effects and feedback that would be impossible under the physical paper medium, such as special smoothing effects, or user interaction effects.

3.2. Mapping Writing to Music

The three main components of a piece of music are pitch, rhythm and harmony. Traditional Chinese music did not have a strong concept of chord and not much in the form of harmony. Usually, one musical instrument or many musical instruments of the same type will take the lead and dominate a musical piece. Therefore, to create music that sounds like traditional Chinese music, our system considers only the pitch and rhythm components. To gain a deeper understanding of the effect of each individual part, these two parts will be processed independently in the beginning of this project and integrated together subsequently. Since both calligraphy painting and music writing involve some additional effects to better express detail emotion, we tried to add in another layer besides these two components.

All the sheet scores and generated music are accessible by the following link: http://goo.gl/8qGrke

3.2.1. Traditional Chinese Music

For our work, we investigated 120 pieces of traditional Chinese music that were composed from 1027BC (Zhou Dynasty) to 1911AD (Qing Dynasty), ranging from court music to military music to folk music. We found that the rhythm of traditional Chinese music is very simple: tied notes or complex beat, such as triplets, were very rare in the corpus. We therefore postulate that simple beat representation is sufficient for notation. Also, most of the pieces are in basic time signature of 4/4 or 2/4, with a few exceptions of other complex time signatures such as 10/4. Since there is no change of time signature in the middle of any piece in our corpus, most pieces can be rearranged to 4/4 for the whole song.

3.2.2. Music Generation

In this project, we investigate several different means of mapping Chinese characters to music both in real-time (online processing) as well as off-line.

3.2.2.1. Analyzing and Transforming Writing Strokes

To transform between music and writing, we need a basic unit on both sides for processing. On the music side, a note seems to be the most natural unit, while Chinese characters are composed of strokes, which also seem to be a natural basic unit. We therefore investigate the possibility of map from writing strokes to musical notes.

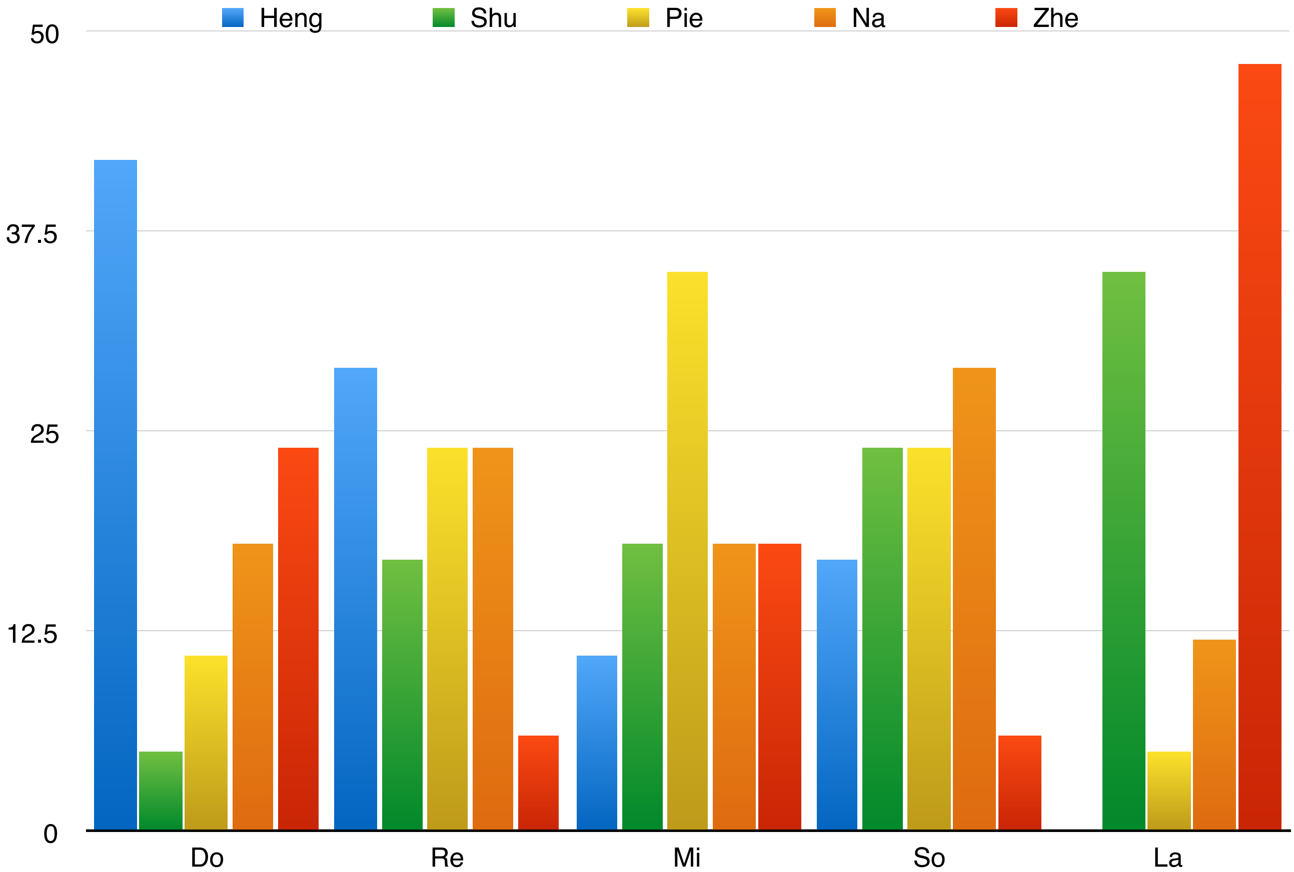


Figure 3.3. Result of experiment mapping calligraphic stroke to musical notes. There is a significant correlation between the stroke and the musical note that is mapped to it ($F_{4, 80}=3.65$, $p<0.01$), especially for the lowest, highest, and middle pitched notes.

Writing strokes can be mapped to music notes on two dimensions. Firstly, the temporal order in which the strokes are written within a character can be matched to the note sequence in music. Secondly, we postulate that there may be a natural mapping from the pictorial *type* of the stroke to the pitch of the musical note. We therefore investigate the likelihood of a metaphoric transference from Chinese calligraphic strokes to musical notes. The 39 types of strokes from Chinese calligraphy writing were categorized into 5 major classes as seen in one of the most commonly used Chinese input methods for computerization, 五筆 (Wubi, literally meaning five strokes). Since Chinese music uses a pentatonic scale (a scale with five notes per octave), we investigate a one-to-one mapping between stroke type and note pitch.

We asked 17 participants aged 20-28 to listen to generated notes, and to look at generated calligraphy strokes. They were then asked to identify the calligraphy stroke that they felt most corresponded to each note. All of the participants were able to read and write Chinese, and 10 of them had some degree of musical training and were thus familiar with notes and tones. The order of presentation was randomized and the experiments repeated 10 times to eliminate order effects. The results (see **Figure 3.3**) suggests that there is a consistent and natural (although not universal) correlation between stroke type and pitch according to human perception.

3.2.2.1. Real-time Music Generation

For real-time generation, an according note will be played almost immediately when the user writes something, and supplementary notes will be added in real-time between stroke, which will lead to a better sense of direct manipulation and user experience. The supplementary notes are generated using a best-fit optimization step. The mapping between stroke type and its corresponding note is defined in **Table 3.1**.

Stroke Type	Heng	Shu	Pie	Na	Zhe
Stroke					
Music Note	Do	Re	Mi	Sol	La

Table 3.1. Stroke types and the corresponding music notes.

3.2.2.2. Off-line Music Generation

Off-line generation from writing to music analyses the user's writing as a whole piece. Since more information is available, especially about writing time sequence and which character that is being written, this method can produce a better mapping and pleasing sound. For the purpose of visual comparison, generated melody and rhythm will be converted to GNU LilyPond file format (Lilypond), along with other information such as the characters and stroke types that contribute to a particular segment of music. From the file, a music score sheet can be automatically engraved that can be studied and be played by musicians using Chinese musical instruments. The powerful customization of the score sheet format allows us to study various mapping easily (**Figure 3.4**).

More importantly, a MIDI file can also be generated from that LilyPond file, which can be further synthesized using various synthesizers to produce realistic traditional Chinese music.

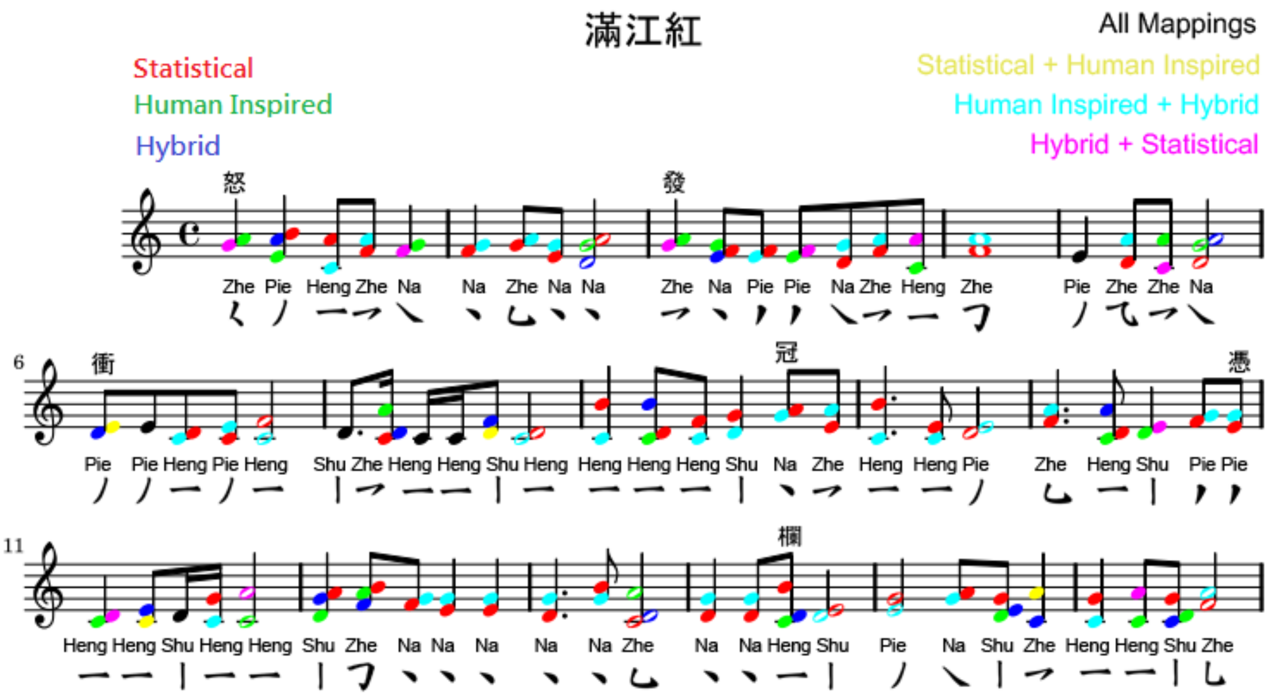


Figure 3.4. Score sheet generated by LilyPond for the study of differences between mappings. Different colors indicate different mappings used to generate that note.

3.3. Generating the Pitch

In this project, we investigate four mappings for the generation of pitch. To afford human control, we use the human-inspired mapping from **Table 3.1** as our basis. (Michael, Will et al. 2012) A further study uses the pictorial mapping from written image to produce sound. Then we test a statistical mapping for a more naturally pleasing sound. Finally, a hybrid mapping is developed by complementing human control with statistical input. The result is quite promising.

3.3.1. Human-Inspired Mapping

When a user writes a character with a sequence of strokes, a sequence of notes will be generated accordingly. For example, **Figure 3.5** shows the music sequence that is generated when the character 木 (wood) is written.

Figure 3.5. Musical sequence generated in response to writing the character 木 (wood).

The advantage of this human-inspired mapping is that it takes into consideration the correspondence between the perception of a calligraphic stroke and the pitch of a note. For example, some strokes give the perception of stability, and these are perceived to correspond with stable note pitches. However, there is a major shortcoming to this mapping. Although the user can directly control the generation of pitches, the generated music may not sound appealing if the music is generated strictly from a piece of written prose, or even poetry, since the writer cannot completely control the writing at the level of individual strokes - the sequence of strokes is dictated by the characters chosen. Even if the writer is allowed to write strokes without regard to whether the strokes form actual characters, probably only very adapt musicians can generate pleasing music in this manner. One more shortcoming is that many common Chinese words tended to repeat strokes, at least for a good portion of their stroke sequence. The corollary, repeating the same note for a good portion of the musical sequence, is not as common. Another, relatively minor, shortcoming is that the whole song will stay in one octave if no further parameter is involved. To address this issue, we extend the mapping in allowing the musical sequence to traverse octave boundaries if it would minimize the number of notes traversed between the current note and the following note. Constraints are applied to avoid notes deviating too far away from the octave of middle C that will not please the human ear.

3.3.2. Whole Character Mapping

In addition to the transformation from raw strokes to musical notes, we also investigate a *whole-character transformation*, which maps a written calligraphy image to a series of code that control the generation of music notes. This method is inspired by cellular automata music generation from previous work. (Bilotta and Pantano 2002)

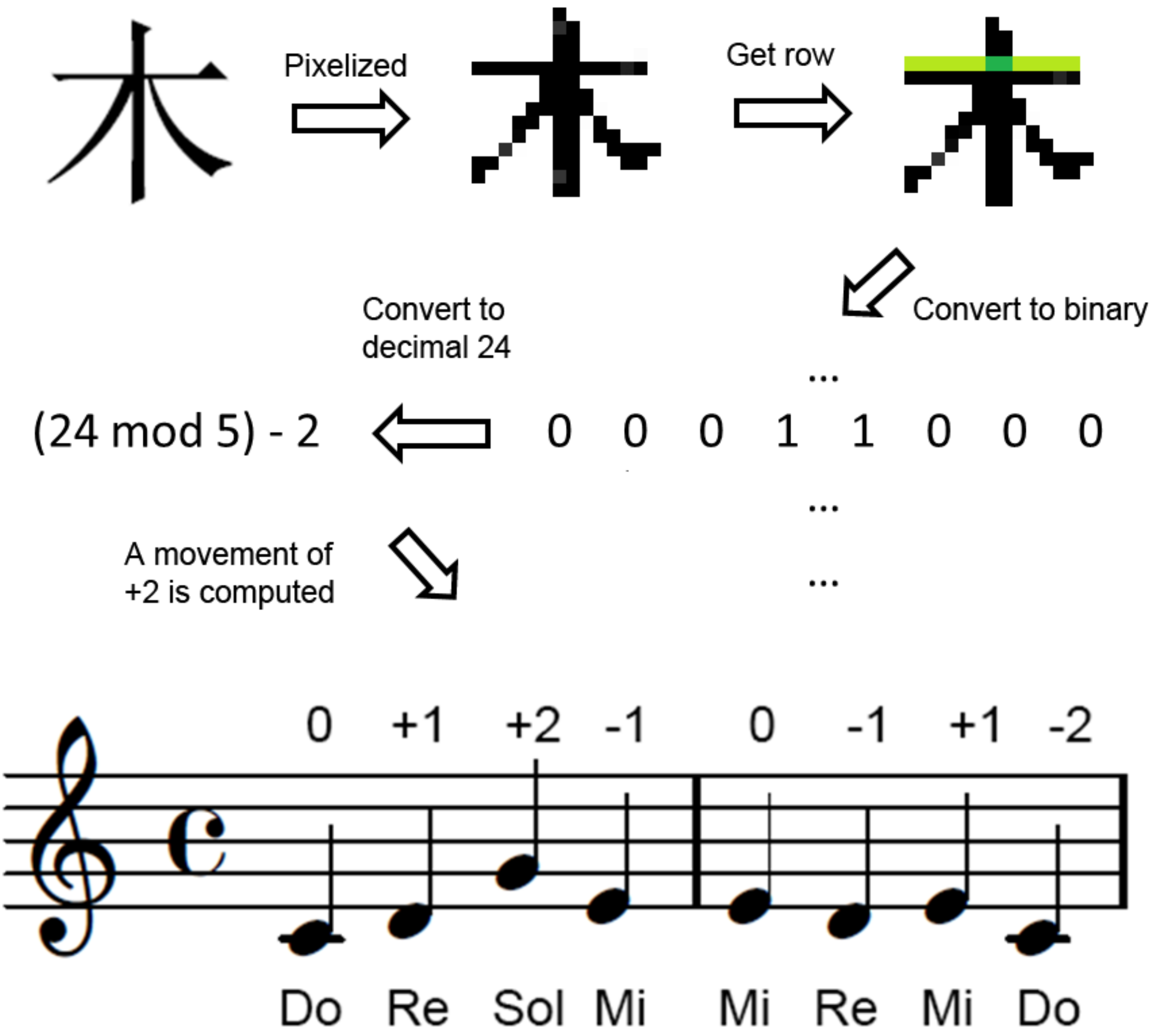


Figure 3.6. Pictorial mapping process showing pixels in a row of writing image being transformed to notes.

Figure 3.6 shows the process. We first pixelate the character as a black-and-white image. Each row of the image therefore consists of a sequence of 0's and 1's, such as 001100, that can be viewed as a binary number. Unlike the approach used in many cellular automata music generation where the numbers have certain self-similarity (Miranda 2001), in this case direct mapping of such numbers to a certain pitch will create randomized noise. Therefore we modify the approach and use these numbers to control the pitch *movement* of the *following* music note. We perform two modulo operations on each row number by 5 (1 octave), and 2 (half octave). This maps the movement of the following note to three possibilities: moving half an octave downwards, staying on the same pitch or moving up half an octave. The rational behind this is that in Traditional Chinese music notes rarely “jump” too far. To avoid having too low or too high pitch that is unpleasant, the sign of the number, i.e. the note movement direction, will be changed if the following pitch goes beyond constraints. Starting from middle C, we will generate flows of music pitch that are controlled by the resolution of the written calligraphy image. Row by row, the pixels in the written image can be transformed to a sequence of notes as shown in **Figure 3.6**.

3.3.3. Statistical Mapping

To ensure that our resulting music is consistent with the Chinese style, we develop a statistical model trained on a corpus of traditional Chinese music to recreate a musical piece that carries the same probabilistic characteristics of traditional Chinese music. Our corpus consists of 120 pieces of well-known or typical traditional Chinese music that were composed between 1027BC and 1911AD that are preprocessed by transposing the scale to C Major, fixing special markings on the notes and normalizing the time signature. Our statistical model uses the conventional trigram model. For each stroke c, a note with pitch p is picked to maximize:

$$P(p_c \mid p_{c-1} p_{c-2})$$

The choice of trigrams rather than longer sequences was made to capture the characteristics of the music without memorizing actual portions. Smoothing by backing off to an interpolated bigram-unigram combination is used when the sequence of the previous two notes cannot be found as the first two notes of a trigram in the corpus. (Ponsford, Wiggins et al. 1999) The issue of this mapping is that the generated music bears no obviously discernible relationship to the writing style or written strokes, save for the number of notes generated. It is simply automatic music generation based on a statistical model. By itself, it cannot achieve our objectives. **Figure 3.7** is a sample score generated by statistical pitch mapping with rhythm extracted from a classical song.

Figure 3.7. A sample score generated by statistical pitch mapping with rhythm extracted from a classical song.

3.3.4. Hybrid Mapping

Direct human control affords human control but does not guarantee pleasing music. Statistical mapping generates pleasing music but does not allow human control. Hence neither human-inspired mapping or statistical mapping alone fully provides what we need. But combined they can. Hence, a back-tracking algorithm has been developed to maximize the likeliness of n-gram probability while still respecting the written stroke types. The algorithm goes through each possible path of note sequences presented in the corpus, remembering the most likely path as a back-track path for each run. The result is that the user can, to some degree, control the generation of pitch sequences, while at the same time the computer helps her to generate pleasing sequences of pitches. In statistical mapping, the generation of pitch is a probability of previous pitches. In human-inspired mapping, the relationship between stroke and pitch is also a probability. So combining these probabilities together we get the following probability to be used to generate pitch. For each stroke c of type t, a note with pitch p is generated to maximize:

$$Max\left(P\left(p_c \mid t_c\right), P\left(p_c \mid p_{c-1} p_{c-2}\right)\right)$$

where $P\left(p_c \mid t_c\right)$ is the most likely pitch that is correlated with the stroke type according to human perception, and $P\left(p_c \mid p_{c-1} p_{c-2}\right)$ gives the most likely pitch according to our trigram model. Our model thus takes into account both human perception as well as musical theory, as evidenced by the patterns that are commonly found in real musical pieces.

Figure 3.8. Score generated by hybrid melodic mapping. The text (怒) on top is the character being written while the texts on bottom corresponds to the stroke types of that character. Note that different pitches may be generated even with same stroke type Na, but the same pitch for the first and third Zhe.

3.4. Generating the Rhythm

Our work investigates three different means of rhythmic mapping. First we used a real-time mapping in which the user controls the rhythm of the generated music in real-time, as opposed to other mappings which are not real-time. Then a statistical mapping was used to generate rhythm. For human-inspired mapping, we tried to capture the writing mechanics presented in calligraphy, along with the syntax of the content being written.

3.4.1. Real-Time Mapping

Using the same technique as statistical mapping for pitch but applying it to rhythm, we built a model of traditional Chinese music corpus with beats. Trigram and smoothing techniques are used to model the sequence of simple beats. However, it also introduces some complexities due to its real-time nature. When the user is writing slowly, notes are “due” before the next stroke is written. In this situation, the system generates the pitch for those notes according to the pitch model, first using trigrams, and then backing off to bigrams and unigrams as necessary. These system-generated notes continue as needed until the next stroke is written, whereby a user-generated note is again created using the stroke-note mapping. The next system-generated notes will use the pitch of the user-generated note in their trigrams and bigrams, thus incorporating the user-generated notes into the melody. When the user is writing quickly, the strokes are generated faster than the rhythm model generates notes. For example, suppose that the rhythm model has prescribed a bar of 4 quarters (or crochets, single-beat notes), but the second stroke is written while the first note is still being played. In situations such as these, the model seeks for an alternate rhythm that would best fit the situation (for example, a bar with a triplet in the first beat), and replaces the prescribed bar with the new one. Our music generation component combines cross-modal mapping and computational modeling to allow users to create music in response to their calligraphic writing, which is also musically sound and supported by statistical models.

3.4.2. Statistical Mapping

In statistical mapping, although we use the same technique as in real-time mapping, many of its real-time issues does not exist anymore, but it introduces new challenges of its own. One issue found in the statistical generation of rhythm is that sometimes the generated beats do not fit in the length of one bar, i.e. one of the beats will be placed across two bars. To address the issue, in the case of picking up a beat that will exceed the length of the current bar, shorter beats that will not exceed the bar will be considered instead as long as such beats exist.

Another issue raised by this mapping is similar to that of statistical pitch mapping where, in the statistical model, there are no corresponding beats for the sequence of strokes in question. However, in this case the writing time intervals provide a basic rhythm to be used as a skeleton. This allows the “writing rhythm” to be used instead of the statistical model.

3.4.3. Human-Inspired Mapping

The writing of Chinese calligraphy has its own rhythm but it is not desirable to convert the writing rhythm to the musical rhythm mechanically, because the writing rhythm very often cannot be mapped exactly to simple beats that conform to music theory. Therefore, we need a mapping to transform the writing rhythm to a more musical rhythm while still maintaining its writing characteristic. For example, as shown in Table 1, instead of transforming the Chinese character 木 to a rhythm of unevenly divided beats (93 : 181 : 114 : 208), the transformed rhythm should be in the form of short, long, short, long (♪, ♩, ♪, ♩) to reflect the writing rhythm and remain appealing to human ear by conforming to music theory. Our transformation process works on multiple levels of stroke, character and sentence.

Stroke Type	 Heng	 Shu	 Pie	 Na
Writing time intervals	93ms	181ms	114ms	208ms
Rhythm	♪	♩	♪	♩

Table 3.2. Stroke writing time intervals and corresponding rhythm of Chinese character 木 (wood).

3.4.3.1. Definition

When a user writes character c with n_c number of strokes, writing time gaps will be formed between strokes.

$w(c,s)$ = Writing time of stroke s in millisecond

$g(c,s)$ = Gap time between stroke s and $s+1$ in millisecond, let $g(c,0)$ = $g(c,1)$ and $g(c,n_c)$ = $g(c,n_{c-1})$ to handle leading and tailing strokes which have only one gap time adjacent to them

t_c = Total writing time of character $c = \sum_{i=1}^{n_c} w_{c,1} + \sum_{j=1}^{n_{c-1}} g_{c,j}$ in millisecond

e_c = Total writing time of character c excluding gap time = $\sum_{i=1}^{n_c} w_{c,1}$,1 in millisecond

b_c = Proposed number of bars = $\lceil t_c \times BPM \rceil$ rounded up to the nearest integer or half, where we fixed ρ = 4 for 4/4 time signature and BPM = 120 for easier handling

The shortest note length we used in this project is sixteenth note, so min_note_length = 1 ÷ 16 = 6.25% time of a bar.

3.4.3.2. Stroke-Level Optimization

For stroke s in a character c:

q = Proposed quantized note duration in where $\text{min\_note\_length} \le q \le \text{max\_note\_length}$, for example 25% is quarter note ♩, 12.5% is eighth note ♪, etc.

The Quantization Cost $Z_{c,s}(q)$ reflecting how well q is fitting to stroke s would be:

$$Z_{c,s}(q) = E_{c,s}(q) + P_{c,s}(q)$$

The Error Score $E_{c,s}(q)$ is defined as the deviation in lengths from ideal note length:

$$E_{c,s}(q) = \left| \frac{W_{c,s}}{e_c} \times b_c - q \right|$$

The Penalty Score $P_{c,s}(q)$ is defined as the total penalty in different aspects of music theory:

$$P_{c,s}(q) = I_{c,s}(q) + R_{c,s}(q)$$

The Irregularity Score $I_{c,s}(q)$ is defined to avoid unusual note lengths, where whole or half notes are encouraged:

$$I_{c,s}(q) = \begin{cases} 0, & q \text{ can be represented by } 1 \text{ or } 1\frac{1}{2} \text{ note} \\ \lambda \times b_c, & \text{otherwise} \end{cases}$$

The Confine Score $R_{c,s}(q)$ is defined to keep q close to the stroke writing time:

$$R_{c,s}(q) = \begin{cases} 0, & q \text{ can be represented by } 1 \text{ or } 1\frac{1}{2} \text{ note} \\ \mu \times b_c, & \text{otherwise} \end{cases}$$

where $\mu = 0.1$

3.4.3.3. Character-Level Optimization

For a character c and a quantization vector $\vec{v} = \begin{bmatrix} q_1 & q_2 & \dots & q_s \end{bmatrix}$ with s note durations, we define the Character Quantization Cost:

$$A_c(\vec{v}) = \left(\sum_{i=1}^{s} Z_{c,i}(q_i)\right) + \sum_{j=1}^{s} \left|b_c - q_j\right| \times b_c$$

For best fit notes with total note length l up to stroke s, Character Length Quantization Score $L_c(l,s)$ is defined as:

$$L_c(l,s) = \min\left(A_c(\vec{v}) \forall q_0, q_1, \dots, q_i \text{ where } q_0 + q_1 + \dots + q_i = l\right)$$

For the best possible fit note lengths $\vec{p}$, it will be the vector $\vec{v}$ with the minimum score of

$L_c(b_c, n_c)$. Since the process to find the best note lengths requires a lot of resources and not much difference were experienced during practical usage between the optimized and the sub-optimized, a heuristic algorithm (**Algorithm 3.1**) will be applied instead to find the sub-optimized note lengths.

Heuristic Fit Algorithm:

Input: Current character c, Number of stroke s,
Total note length l, Stroke writing time $w_{c,s}$,
Total writing time e_c

1. Set note lengths $\vec{p}$ to $\begin{bmatrix} q_1 & q_2 & \dots & q_s \end{bmatrix}$ where $q_i = \left\lceil \frac{w_{c,s}}{e_i} \right\rceil$
2. Set *difference* to $l - \sum_{i=0}^{s} q_i$
3. While *difference* is greater than 0
4. Repeat for $i = 1$ to s
5. Set p_i to $q_i + min\_note\_length$
6. Set $score_i$ to $Z_{c,i}(p_i)$
7. Find the smallest $score$ and set j to be its index
8. Set q_j in $\vec{p}$ to $q_j + min\_note\_length$
9. $\vec{p}$ is output as the sub-optimized note lengths

Algorithm 3.1. Heuristic Fit Algorithm for finding corresponding note lengths

Character 青 (green)	Stroke s								
	0	一 1	一 2	丨 3	一 4	丿 5	㇆ 6	一 7	一 8
$w_{1,s}$	/	93	94	203	219	125	578	109	93
$g_{1,s}$	172	172	156	125	172	172	188	141	141
$w_{1,s}/t_1$	/	3.52%	3.56%	7.69%	8.30%	4.73%	21.89%	4.13%	3.52%
$w_{1,s}/e_1$	/	6.14%	6.21%	13.41%	14.46%	8.26%	38.18%	7.20%	6.14%
$g_{c,s-1}+w_{c,s}+g_{c,s}/t_1$	/	16.55%	15.98%	18.33%	19.55%	17.77%	35.53%	16.59%	14.20%

Table 3.3. Example score table for the character 青 (green).

3.4.3.4. Example

Referring to **Table 3.3**, suppose we want to calculate the score for character written as $c=1$, $n_1=8$, $t_1=2460$, $e_1=1514$, $b_1=1.5$.

$$E_{1,1}(q)=\left|\frac{w_{1,1}}{e_1}\times b_1-q\right|=\left|\frac{93}{2640}\times 1.5-q\right|$$

For $q=6.25\%=0.0625$, $E_{1,1}(6.25\%)=|0.0921-0.0625|=0.0296$

Since q can be represented as a sixteenth note, $I_{1,1}(6.25\%)=0$. Also $\frac{w_{1,1}}{t_1}\times b_1=0.0528$ and $\frac{g_{1,0}+w_{1,1}+g_{1,1}}{t_1}\times b_1=0.2483$, and $0.0528<0.0625<0.2483$,

$$\therefore R_{c,s}(6.25\%)=0$$

$$Z_{1,1}(6.25\%)=E_{1,1}(6.25\%)+I_{1,1}(6.25\%)+R_{1,1}(6.25\%)=0.0296+0+0=0.0296$$

For $q=31.25\%=0.3125$,

Since q can be represented at best as one quarter note plus one sixteenth,

$$\therefore I_{1,1}(31.25\%) = |0.0921 = 0.3125| = 0.2204$$

Also $\frac{W_{1,1}}{t_1} \times b_1 = 0.0528$, $\frac{g_{1,0} + w_{1,1} + g_{1,1}}{t_1} \times b_1 = 0.2483$ and $0.0528 < 0.3125 > 0.2483$,

$$\therefore R_{c,s}(31.25\%) = 0.1 \times 1.5 = 0.15$$

$$Z_{1,1}(31.25\%) = E_{1,1}(31.25\%) + I_{1,1}(31.25\%) + R_{1,1}(31.25\%) = 0.2204 + 0.15 + 0.15 = 0.5204$$

3.4.3.5. Sentence-Level Optimization

Apart from character-level optimization, our system takes into consideration the syntactic function of the characters. Since Chinese is written without spaces between characters, word segmentation is carried out to group characters into words. Using an automatic Chinese Part-Of-Speech (POS) parser and tagger, each word is then grouped and tagged with its POS G_c , e.g. ADJ, VERB or NOUN.

Sentence Confine Penalty $A_c(q)$ is defined as:

$$A_c(q) = \begin{cases} \omega_c, & \text{Beginning of a word and } \left[(G_{c-1} = \text{ADJ and } G_c = \text{NOUN}) \text{ or } (G_{c-1} = \text{ADJ and } G_c = \text{VERB}) \right] \\ \omega_f, & \text{Beginning of a word and } (G_{c-1} = \text{VERB and } G_c = \text{NOUN}) \\ \omega_b, & \text{Beginning of a word but not in previous cases} \\ 0, & \text{Not the beginning of a word} \\ \omega_n & \textit{otherwise} \end{cases}$$

where we set the parameters $\omega_c = 0.1$, $\omega_f = 0.15$, $\omega_b = 0.2$, $\omega_n = 1$.

So instead of making the notes fit into the proposed note duration strictly for each character, we consider all possible lengths and try to concatenate these characters with the least overall Sentence Confine Penalty. A similar heuristic will be applied to get the sentence-level sub-optimized best possible fit note lengths.

3.4.4. More Music Parameters

In Chinese ink brush calligraphy, the calligrapher aims to express him/herself through the shape of the strokes, the pressure applied to the brush, and the speed of the strokes. Hence, we applied these writing mechanics to our early prototype.

These mechanics are mapped to parameters from music that share similar metaphors. Writing pressure is mapped to sound volume, which is intuitive as playing a louder note requires plucking the string with greater pressure. The speed of the stroke is mapped to the rhythm of the music.

3.4.5. User Interactions

For the real-time online generation system, a number of special features were also added to the system according to user feedback, which adds to the experience of generating music through writing:

3.4.5.1. Viewport Dragging and Scaling

The writing surface of the CalliMusic system is presented as a viewport on a piece of paper of infinite size. Using their free hand, users may position the viewport on any part of the paper. They may also scale the viewport to get a bigger or more detailed view of their work.

3.4.5.2. Visual Music Generation Feedback

To give the user a better sense of his composition, the musical score is generated across the bottom of the viewport. Notes that correspond directly to calligraphic strokes are shown in red, notes generated by the bigram model as harmony are shown in black.

3.4.5.3. Self-Collaboration and Multi-Part Composition

In addition to real-time composition, CalliMusic also enables self-collaboration and multi-part composition through record, reset and undo functions. Users may replay a portion of their artwork, while simultaneously writing new strokes, thus generating a composite multi-part melody. The ink color, bristle texture and paper texture may also be changed, and changes in the ink color and bristle textures are mapped to changes in the musical instrument, which adds to the richness and diversity of the artwork.

3.5. Performance Evaluation

We tested most of the combinations of each melodic mapping with each rhythmic mapping, including original pitch and rhythm sequence, to determine the combinations that give the best results. In order to not require the evaluator to listen to too many songs, pictorial mapping was been omitted due to its lack of user manipulation and unsatisfactory initial result. To study how a melodic mapping performs independently from the rhythmic mapping, and vice versa, a fixed rhythm or melody part extracted from an original music piece is used. **Table 3.4** is a summary of satisfaction levels of most of the mapping combinations reported by 4 subjects, all of whom have at least a basic level of musical training. Each subject gave a satisfaction level ranging from very bad, quite bad, bad, okay, good, quite good and very good, which was mapped to a score of -3, -2, -1, 0, 1, 2 and 3. The overall satisfaction level was then determined by averaging the score and converted back to satisfaction level.

Not all mappings generated music pleasing to subjects. Statistical mappings provided a reasonably good result. The proposed hybrid pitch mapping and human inspired rhythm mapping received the best level of satisfaction, as expected.

Satisfaction Level		**Pitch Mapping**			
		Original	**Statistical**	**Human Inspired**	**Hybrid**
Rhythm Mapping	**Original**	*(N/A)*	Okay	Okay	Good
	Statistical	Okay	Good	Okay	Good
	Human Inspired	Quite Good	Good	Good	Very Good

Table 3.4. A summary of satisfaction level of most of the mapping combinations.

The projections made when we designed CalliMusic have also been validated in the user test. We focus on the hybrid pitch mapping and human-inspired rhythm mappings since these give the user control over the music generation. (1) The same user writing the same sentence in the same style generated essentially the same piece of music, especially for expert calligraphers whose writing are more stable than those of novices. (2) The same person who writes the same sentence in different style — in this case we tested 楷書 (KaiShu, regular) and 行書 (XingShu, semi-cursive) styles — generated essentially the same set of pitches but different sets of rhythms. Only in some rare cases among expert users where their XingShu style simplified and changed the sequence of stroke, the set of generated pitches were different. (3) Different people writing the same sentence in the same style generated essentially the same set of pitches but different sets of rhythm. We observed that novice users tend to think more before actually writing on the paper, so the music rhythms they generated are slower and more regular than those by expert users.

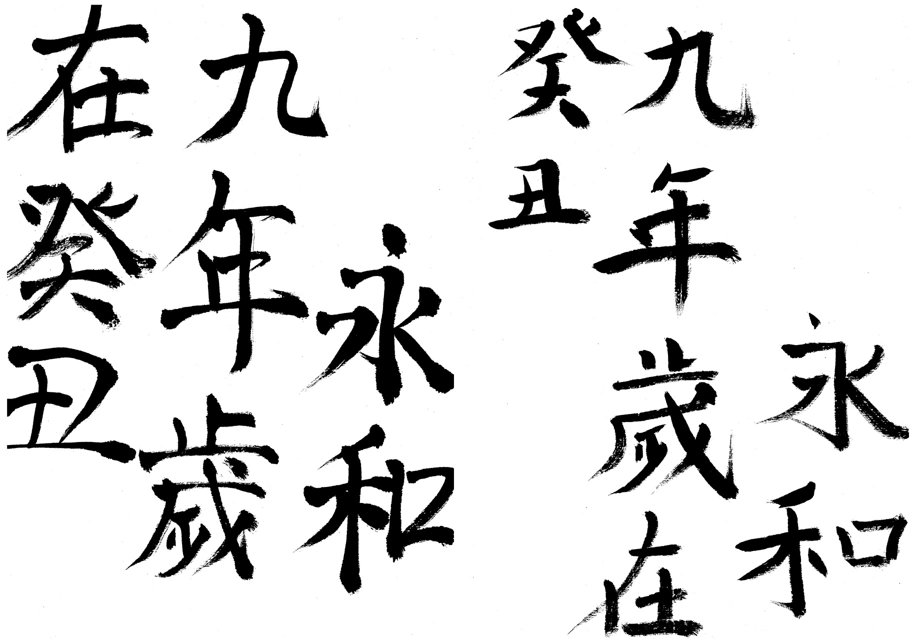


Figure 3.9. Two pieces of writing (portion only) by two subjects in the same writing style. On the left, (1) a more accomplished calligrapher; no the right, (2) a novice. (not really an expert)

Figure 3.9 shows two pieces of writing by two experiment subjects, using the same writing style. The subjects were asked to write a well-known classical calligraphy work 蘭亭集序 (Lantingji Xu) by 王羲之 (Wang Xizhi), written in semi-cursive style and composed in year 353. They both copied from the same work placed in front of their desk, so the stroke size and order were all the same. As a result, the sets of pitch generated were the same. However, subject 1, the more accomplished calligrapher, wrote his work faster and with more complex rhythm throughout the writing of each stroke, showing more confidence. More nuanced details can be found in the starting and ending of every stroke. On the other hand, subject 2, the novice, wrote it slower and with more constant speed for each stroke, pausing and thinking for a longer time between strokes. Therefore the rhythms of their generated music were different in that the music of the first subject is more lively and pleasing to ear, while the music of the second one is more formal and boring. (**Figure 3.10**)

Figure 3.10. Two according ices of music (portion only) generated by these subjects using hybrid rhythmic mapping and human-inspired melodic mapping. Note that the work of subject 1 shows more rhythmic features than that of subject 2.

3.6. Conclusion and Future Work

We have presented a cross-modal mapping between Chinese calligraphy and Chinese music. This allows even users who do not know music to generate reasonably pleasing music through different proposed mappings, as long as she can write Chinese. Just like many other sophisticated art forms, it takes a lot to become an expert Chinese calligrapher. But it is relatively easy to learn to use the Chinese ink brush to write basic Chinese calligraphy. Hence our methods make Chinese music accessible to a wide range of users, from novices to experts.

3.7. Further Investigation

This project focuses on the physical and syntactical aspects of Chinese calligraphy and music. Taking the semantics into account can result in better mappings in terms of meaning and emotional expression.

In addition, some of the parameters used in the mappings are set through heuristic processes.

More complete experiments are planned in order to improve the mapping quality through parameter tuning.

Finally, the techniques that we have developed are, of course, applicable to other forms of writing, and other types of music. Applications in performance arts, therapeutic and rehabilitative domains are also being explored.

4. MAKING MUSIC FROM MOTION - MOBILE DJ

Mobile DJ is a system that we created to investigate tangible and music transformation, which consists of a tangible musical control interface that is connected to a mobile device for signal processing and social interaction. In Mobile DJ (Figure 4.1), musical control can be divided into two functions: musical effect control and audio time-based control. Players can actively enrich the active playing track by inserting chords, changing instruments or rhythm while scratching is implemented to offer time-based control. The following section discusses the system architecture and interaction design.

Figure 4.1. A Mobile DJ user with connected *TouchPad* and iPhone app.

4.1. Architecture and System Design

The overview of the system architecture of Mobile DJ is shown in **Figure 4.2**. The system consists of three components: Tangible Musical Control Interface, Active Music Listening Application and Application Server.

The Tangible Musical Control Interface is responsible for capturing the user's interactions. Together with the mobile device, it forms a self-contained digital musical mixing platform. Active music listening on a single-user basis is enabled when a user connects the Musical Control Interface to device and registers it with the application.

The Application Server provides support for multi-user active music listening which is implemented with i*Chameleon web services middleware (Lo, Tang et al. 2012). It provides a comprehensive protocol to model and integrate the Musical Control Interface together. Portable music player, such as an Apple iPhone, that is running an Active Music Listening Application can be connected by standardized web services.

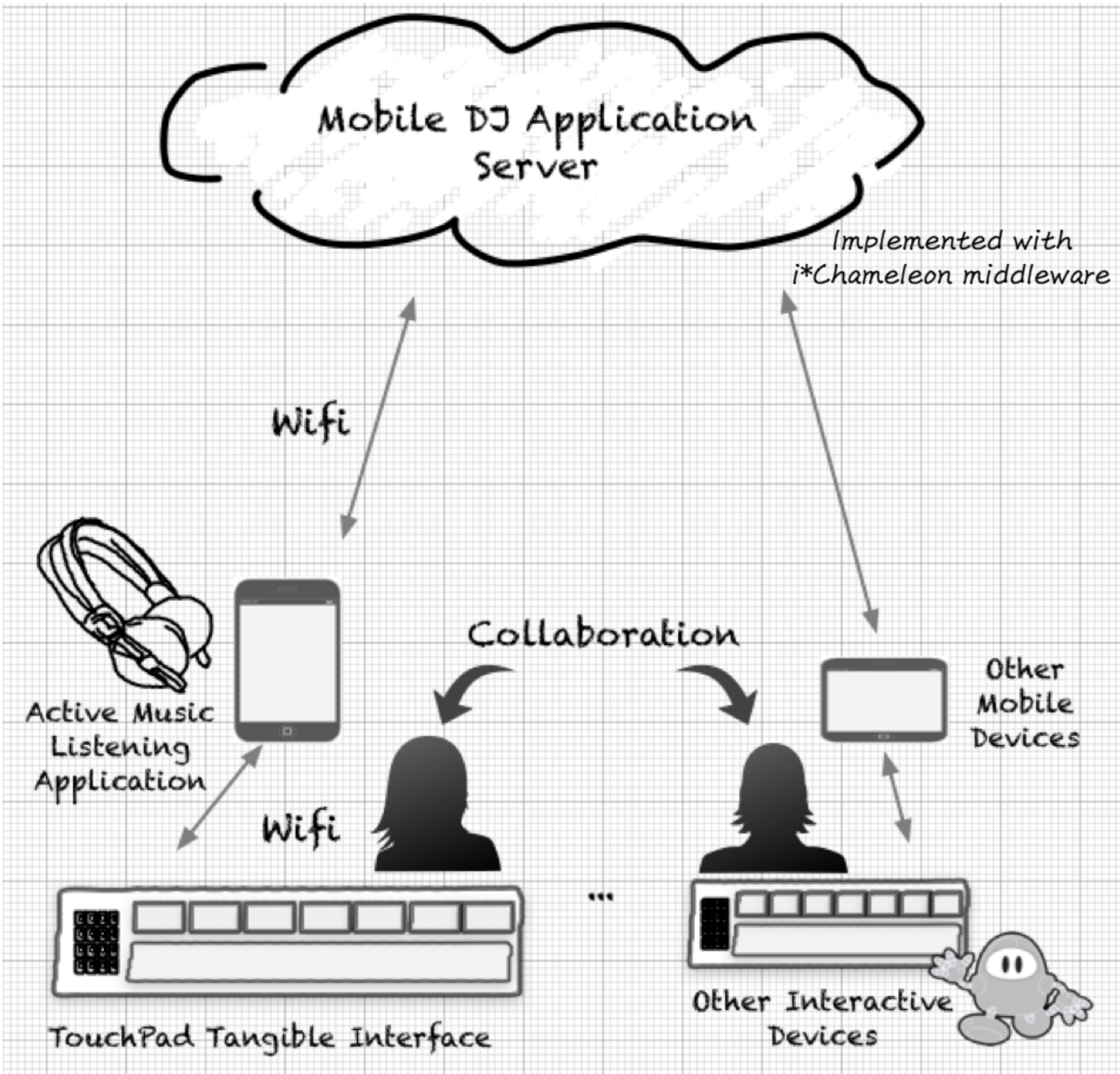


Figure 4.2. Mobile DJ architecture.

The prototype Active Music Listening Application was implemented to run on iOS in Objective-C, and the BASS audio library (Bass Audio Library) was used to implement special sound effects such as scratching.

4.2. Tangible Musical Control Interface

The Tangible Musical Control Interface is intended to be portable, and, to a certain degree, wearable. We performed some preliminary user tests to gauge the intuitiveness of different form factors for the interface. The final prototyped design was a lightweight armband equipped with sensors, which would be manipulated by a gloved hand.

There has been some work in predicting social acceptance of a new interface through analysis of the user interaction (Assaf, Emmanuel Munguia et al. 2005, Sami, Jonna et al. 2007). Reeves (Demey, Leman et al. 2008) suggested using manipulations and effects to evaluate the design of public interfaces. This also fits in with our intuitions on the best way to facilitate the immersive experience of the user. Based on this conclusion, we designed our Tangible Musical Control Interface to have high degrees of manipulation and effect. In other words, we designed our interaction to require large scale body motions, and to produce rich feedback and distinct effects.

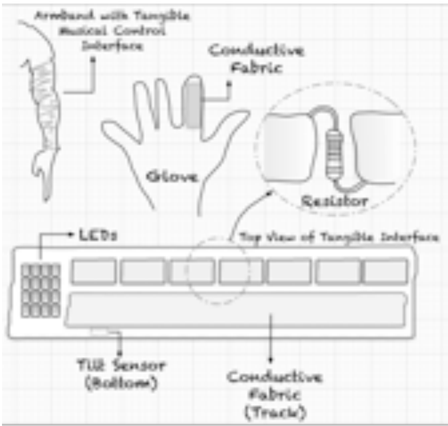


Figure 4.3. Hardware implementation of *TouchPad*

Figure 4.3 shows the hardware implementation of the interface. We required the interface to be intuitive and to support different modes of interaction with the music. From a wearability standpoint, the form factor of the interface as an armband implied that users should be able to interact with the music through movements of the other hand on the arm, or by swinging the arm. In addition, we wished to facilitate immersion into the musical listening experience by encouraging exaggerated interaction motions from the user.

4.3. Interaction Design

4.3.1. Scratching

Moving a vinyl record back and forth to distort the melody and produce distinctive sounds is considered to be an art form in and of itself in hip-hop music. Traditionally, scratching requires a DJ mixer, vinyl record, and a turntable. The user physically interacts with the record by moving it forward and backward and butting the sound on and off with the crossfader (Hansen and Bresin 2003).

The Musical Control Interface for Mobile DJ supports scratching through a slidebar potentiometer incorporated into the surface of the armband. The slidebar consists of two tracks. The lower track is in one piece, while the upper track consists of segments connected with constant-valued resistors. When a user touches both the upper and lower tracks at the same time, the resistance between the tracks changes in proportion to the number of resistors that have been bypassed, thus allowing us to determine the upper track segment that is currently being touched. For aesthetics and usability, the tracks in the potentiometer are constructed from conductive fabric, which makes them congruent to textiles and garments.

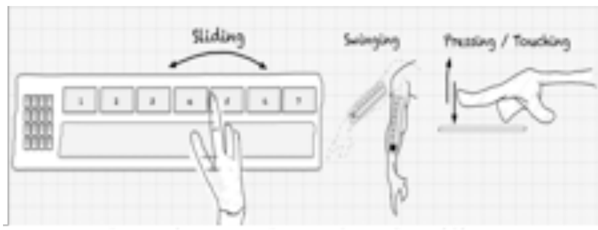


Figure 4.4. Interaction design of Mobile DJ

The connection between the tracks created some challenges. We designed a conductive glove to enable the potentiometer to better and more stably detect changes in the resistance. At the same time, wearing the glove on one hand enhances the user's immersive experience, similar to how stage costumes enhance the ability of actors to immerse into their character.

The user triggers a scratching interaction by sliding his/her hand over the tracks. The amplitude (number of sections that the sliding motion moves over), velocity and direction (up or down the arm) of the motion are mapped to angular displacement of the vinyl record, speed of the scratching motion and the direction of the motion. Our preliminary tests showed that an angular displacement of twenty degrees per track segment generated the best effect.

4.3.2. Swinging

Swinging, swaying and bouncing along to the beat is a common response to music, and certainly very commonly seen in active music listening experiences. The armband form factor of the tangible interface also suggested interaction through swinging or waving of the arm.

This mode of interaction is supported via the incorporation of a tilt sensor in the armband, which detects movements of the user's body. A low sensitivity filter is used to distinguish conscious movements, such as swings, waves and bounces, from normal movement.

In preliminary user evaluations, we observed that swings and waves of the arm were often used to communicate intentions in multi-user, remote collaboration scenarios. These intentions were often "handshaking" interactions, such as signaling the beginning or the end of a turn, or inviting the other user to begin his/her turn. These swings and waves are detected and communicated to the other users via the visual feedback mechanism to be described later.

4.3.3. Pressing

In addition to sliding, the slidebar potentiometer on the tangible interface also supports a pressing interaction. In Mobile DJ, the pressing motion maps to the enrichment of the musical track by interspersing the melody with chords and different instruments, or changing the rhythm and the harmony.

The interface supports both a short "tap" as well as a long "press". The "buttons" are the upper track segments. A tap on a "button" adds a chord to the music; the system will automatically generate an appropriate chord based on the key and the instrument of the currently playing song. Since we have seven "buttons", this allows the user to generate up to seven different chords. Long presses trigger a change in rhythm or harmony of the music in a similar fashion.

4.3.4. Visual Feedback

In general, in active listening, the user's interactions will receive audio feedback through the rearranged or modified music track. However, since one of the objectives of Mobile DJ is to support remote multi-user interaction, it was deemed necessary to incorporate visual feedback to enhance the interaction between users.

The visual feedback is supported by a matrix of multicolored LEDs that is incorporated into the tangible interface above the slidebar potentiometer. The LEDs flash in different colors to signal interactions from the collaborating partner, which provides a channel that allows a certain degree of communication and signaling, but without interfering with the experience or imposing upon the center of attention of the user. In the absence of signals from the other party, or in single-user mode, the LEDs serve as an additional form of feedback to enhance user immersion by flashing along according to the rhythm of the music.

4.4. Collaborative and Social Interaction

Gesture	Duration	Section	Function
Sliding	-	1 - 7	Scratching
Pressing	< 1 sec	1 - 7	Chords
	> 1 sec	1 - 2	Changing background music
	> 1 sec	3 - 4	Changing chord style
	> 1 sec	7 - 8	Changing instruments
Swing	< 1 sec	-	Indication of end of performance and it will reflect on other connected players

Table 4.1. Gestures and corresponding functions.

To support the social interaction between users, Wearable DJ includes a location-based, user discovery function, which allows users to discover other “unconnected” players nearby. For example, player A can see that player B, who is 500m away from him, is still unconnected, and invite him to join the music network. Once two players are connected, they share their chords and music. Both of them can manipulate the same background sound track by adding their own records. Additionally, each player can wield different effects, or can alternate effects, which provide for better effect on collaboration and socialization.

4.5. Preliminary User Feedback

To explore the usability of Mobile DJ and investigate the ways in which the different system design factors affect the interaction style and collaborative effect, we conducted a task- driven experiment.

9 participants (6 male, 3 female) between the ages of 18-30 were recruited for this experiment. None of them had worked with the device before. 4 of them had prior musical training. The participants were first given three minutes to explore the device by themselves, after which, a short tutorial was given to explain proper usage of the interface. They were then asked to spend around 5 minutes testing the device. Finally, they were given an iPhone running an application that could also simulate the same interactions (scratching, pressing and swinging), and asked to repeat their interactions with the phone. The experiments were all video-recorded for analysis. Post-experiment, the subjects were asked to compare the two systems, and post-experiment interviews were conducted for more in-depth investigations.

Our first observation relates to the initial posture chosen by the experiment subjects when they were given the devices, as this is likely to affect the manner in which they are used. For example, if the users preferred to sit while interacting with the device, the scale of motion would likely be limited. Interestingly, almost all of our users chose to interact with Mobile DJ in a standing position – which we noticed was not the case when they were presented with the mobile phone.

Almost all of the users were also able to discover the sliding and pressing interactions without any help. Surprisingly, however, none of them discovered the swinging/waving interaction during their explorations.

To evaluate the ability of the system to encourage movement during active listening, we used the skeleton model from the Kinect motion-capture camera to record the body motion in terms of human translation and joint variation. The motion was then quantified using the concept of motion energy, which is calculated as a weighted sum of movement levels of the body's joints. We tracked the subjects' joints' locations throughout the experiment. At intervals of 100 milliseconds, we calculate the motion energy as $\sum w_i \cdot \Delta d_i$, where w_i and Δd_i respectively denote the weight and the spatial translations of the i^{th} joint. The weight of the torso and wrists is set to be 1.5 times that of other joints', which allows the model to focus more on full- body movement and hand movement.

Figure 4.5 shows the motion energy of the 9 subjects during the experiment that tested Mobile DJ and its iPhone version. The central mark denotes the median energy level, while the edges of the box indicate the 25th and 75th percentiles. It is clear that the subjects' movements were generally

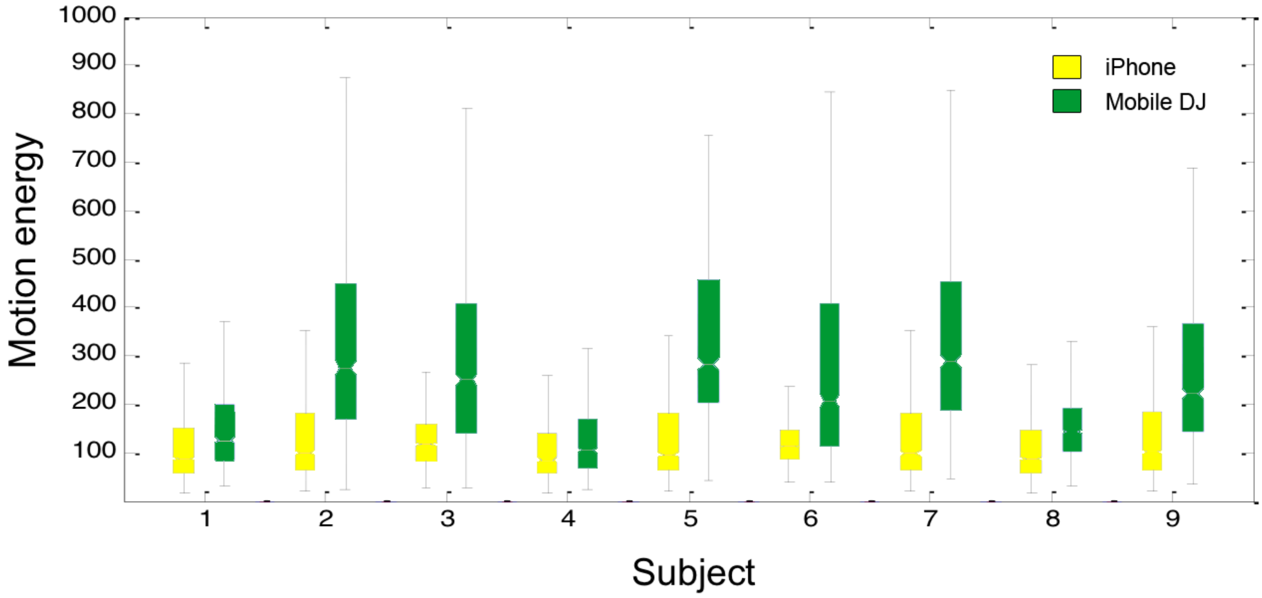


Figure 4.5. Motion energy comparison between Mobile DJ (right bar) and iPhone counterpart (left bar) of 9 subjects

larger when using the Mobile DJ device than when they were using the iPhone version. This is consistent with our observations during the experiment. We can therefore conclude that the Mobile DJ interface succeeds in motivating a larger body movement.

Features	iPhone	Mobile DJ
Functionality	77%	11%
Collaborative Engagement	0%	77%
Portability	44%	44%
Visibility of Actions	22%	66%

Table 4.2 Comparison of user feedback between Mobile DJ and iPhone.

Table 4.2 shows the experiment subjects' evaluations of the Mobile DJ device and its iPhone counterpart. The users were asked to compare Mobile DJ with the iPhone on four dimensions: functionality, collaborative engagement, portability and visibility of actions. For each dimension, they were asked to pick the device that they felt was better. They were also given the option of "neither" if they were not able to decide.

From the table, it can be seen that although the users judged that the iPhone's functionality was better than that of Mobile DJ, Mobile DJ was judged to be better at providing a collaborative experience. From our interviews with the users, they felt that a mobile application would draw the focus of players to the interaction between the device and the user, rather than to other players. When pressed for their reasons, some subjects suggested that this may be due to the fact that Mobile DJ has better error tolerance and offers eye-free control, which removes the need for the users to focus on the device.

Our subjects also felt that the portability of Mobile DJ approached that of its iPhone counterpart. There were suggestions that even though the current prototype is large, as it is designed as a wearable interface, it could be potentially embedded into clothes, instead of an external armband. In terms of visibility of actions, nearly all subjects felt that the interface of Mobile DJ was intuitive and “fun to control”. They also felt that the movement required from the tangible system is closer to the human movement than that required by the iPhone application. Confirming our hypothesis that the tangible interface makes for a better music listening experience, most users reported that the new interface helped them to engage better into the active listening experience and raised their interest in musical manipulation. Some users also remarked on the need for real-time response with as little delay as possible. Since the larger scale of motion resulted in heightened emotion, they suggested that any delay in the response feedback would result in heightened frustration from the user!

4.6. Evaluation

Our previous experiment was designed to test the affordances of the Mobile DJ tangible control interface by analyzing the way in which novice users interacted with the system. However, as Mobile DJ is very much a musical instrument (albeit an unconventional one), and collaborative musical manipulation or performance requires performers who have established rapport with each other, we felt that a complete set of experiments should involve users who were trained in music, familiar with the system, and were also familiar with each other. We therefore asked the subjects from the previous experiment who were more quickly comfortable with the Mobile DJ interface and also had musical training. Our objective is to investigate the way in which our system design supports user interaction, especially in the aspects of collaboration and contextual awareness. Contextual awareness includes the self-awareness of operation, the mutual awareness of the counterpart and the awareness of the performance.

Our experiment subjects were introduced to the Mobile DJ system during a ninety-minute session. During the first half- hour, they were asked to explore the interactions and come up with a collaborative, face-to-face performance. Despite them both having musical training, their improvised performance was acceptable neither to themselves nor to the observers.

The second half-hour involved a musical imitation game, in which the experiment subjects were each given a Mobile DJ system, and asked to take turns imitating and elaborating upon the special effects introduced into the melody by the previous person. Their initial response was tentative, but they warmed up quickly and were able to perform an enjoyable show.

During the last half hour, the experiment subjects were placed in separate rooms and repeat the previous improvisation and competition tasks. The experiment was recorded for analysis, and the subjects were interviewed after the experiment. From the analysis and the interview, we obtained the following observations and conclusions.

4.6.1. System affordance

It was reported that the subjects were distracted by trying to remember the different functionalities and operations of the Mobile DJ interface during the first half-hour. Their experience improved dramatically in the second half-hour, after they had gotten used to the interface and were able to focus entirely upon the other person and the music. This was an indicator to us of the interface's intuitiveness and affordances in the collaborative experience.

4.6.2. Interaction rules and collaboration

The failure of our experiment subjects to produce an acceptable collaboration in the first half-hour demonstrated that even though both users were musically trained, it was not enough to produce a satisfying collaboration on a new device. This was not unexpected, as it would seem to be the case that it would be difficult for two individuals to be able to improvise a performance within such a short time with a new instrument, and without prior collaborative practice. However, in the second half-hour, the subjects were able to successfully use the system in the game that we set for them, which used a turn-based musical imitation approach, and they reported that they enjoyed the game as well. This indicates the utility of having appropriate interaction rules to guide active listening and collaboration, especially for novice users.

4.6.3. Contextual awareness and collaboration

When the subjects were placed into different physical locations in the last half-hour, we observed that the subjects were able to start their turn almost immediately after the previous turn had ended, similar to the situation when they were situated in the same location. The subjects quickly learned to make use of the ability of the interface to sense signaling gestures such as swinging to indicate “I’m done, here’s your turn” to enhance each other’s awareness. This suggests that contextual awareness plays an important role for remote musical collaboration, an absence of which highly impacts the overall performance effect. In addition, we also observed much motion synchrony between the experiment subjects. Both subjects tended to move their bodies in response to the beat. Although the way in which they moved their body was different, their movements were synchronized with each other and with the music. This suggests that the subjects were able to immerse into the musical experience.

Figure 4.6 Face-to-face collaborative performance

4.7. Discussion

Design principles provide developers with concepts from which to design new types of interactive devices. Based on Mobile DJ and its experimental results, we concluded the design problems and suggest recommendation for two aspects: tangible musical interfaces and active collaboration interfaces.

First, wearable devices promote portability. For tangible musical interfaces, sensors could be embedded in clothes to detect the body motion from which different musical effects could be produced. From our experiment, larger scale of motion directly relates to the enjoyment of the experience. This suggests that the design of the interaction should involve as much of the body as possible for these types of interfaces.

We also observed that larger scale of motion increases the frequency of interaction from the user. Therefore, instant feedback and exaggerated effect are suggested as essential factors for the success of tangible musical interfaces. Even a short delay time will affect the usability and acceptance of the device.

The location of the device obviously has a significant impact on the active listening experience. The positioning, surprisingly, of sensors, has not been studied extensively. Gemperle (Gemperle, Kasabach et al. 1998) states that the criteria for location vary with the accessibility of target user and needs of functionality. However, it also summarized and concluded that eight parts of human body are the most unobtrusive locations for wearable objects: the collar area, rear of upper arm, forearm, ribcage, waist, thigh, shin and top of the foot. The most suitable locations for wearable devices in tangible active music listening, however, needs further investigation.

For the aspect of collaboration, reflecting the user status of other players is an important element. When different users interact over the Internet and are required to collaboratively contribute to the same task, it is obviously crucial for them to know and understand the status of the others in order to provide instant feedback. This feedback, however, does not need to be in the center of the user's attention.

4.8. Conclusion

This chapter presents a mobile system, Mobile DJ, which uses active music listening and tangible interface to provide users with a social and collaborative music listening experience. Users can apply different styles of musical interaction, such as swinging or sliding, to control the audio track. In addition, it is possible to collaboratively manipulate the music even when the users are in different physical locations. Users can remix, scratch the same sound track or intersperse the melody with specific chords and different instruments. Our experiments showed that the prototype was well received by the users, who enjoyed the tangible interface and the collaborative effort.

The tangible interface of our system is successful at motivating larger body movement when compared to touchscreen mobile devices. In addition, it is possible for experts to use touch exclusively to interact with the system. We also noticed that interaction synchrony can be achieved, even when the users are in different physical locations. Mobile DJ therefore provides a tangible and wearable interface for playing active music, and also makes it possible for remote collaborative active music listening.

4.9 Further investigation

Gaining insight from this project, we have seen how cross-modal interaction and alternate outputs can facilitate creative and collaborative effects from users. In the following chapter, we will adapt what we learnt to another project that explores more about transformation of correctness of writing into graphics.

5. WRITING ANALYSIS AND FEEDBACK - KID

5.1. Introduction

Early childhood education is an important sector of education. In Hong Kong, an increasing amount of public educational resource has been allocated towards early childhood education. The availability of ubiquitous computing devices has also led to the development of increasing numbers of technology-enabled education methodologies and platforms, with the main focus being on languages (English and Chinese) and Arithmetic.

The Oriental languages, and Chinese in particular, have a complex structure. While Latin languages are composed of a linear sequence of basic alphabetical components, written Chinese exhibits a two-dimensional pictographic composition nature, in which the basic ingredients, strokes, can be spatially arranged in many different ways. To maintain literacy, children learn to write at an early age. In Hong Kong, for example, formal reading and writing starts before elementary school, with most kindergarten children starting to learn penmanship as early as age 3 or 4. By the time they are in elementary school, children regularly spend as much as 50% of their time in handwriting tasks (Tseng and Chow 2000), and a failure to achieve handwriting competency has been shown to have long-term effects on self-esteem (Feder and Majnemer 2007), not to mention the impact on elementary study.

The most common learning approach is learning by practicing: teachers ask students to write the characters on a gridded sheet of paper, such that each character falls neatly within one gridded square, as shown in **Figure 5.1**. They are "graded" according to the correctness of the written character, whether they exceed the boundary of the grid, whether they are too big or too small, or any obvious irregularity. Grading by hand is tedious and oftentimes subjective and inconsistent. The feedback is also delayed, which means that children often practice incorrect methods of writing for extended periods of time before they are identified and corrected. Finally, the feedback is sensationally negative to the children, especially when the evaluation is not positive.

This part presents a potential solution to this problem through an app called KID (Kindergarten Interactive Demo-app). KID is built to run on an Android tablet and will detect the child's written strokes and determine whether the character is correct. Errors in incorrect characters are

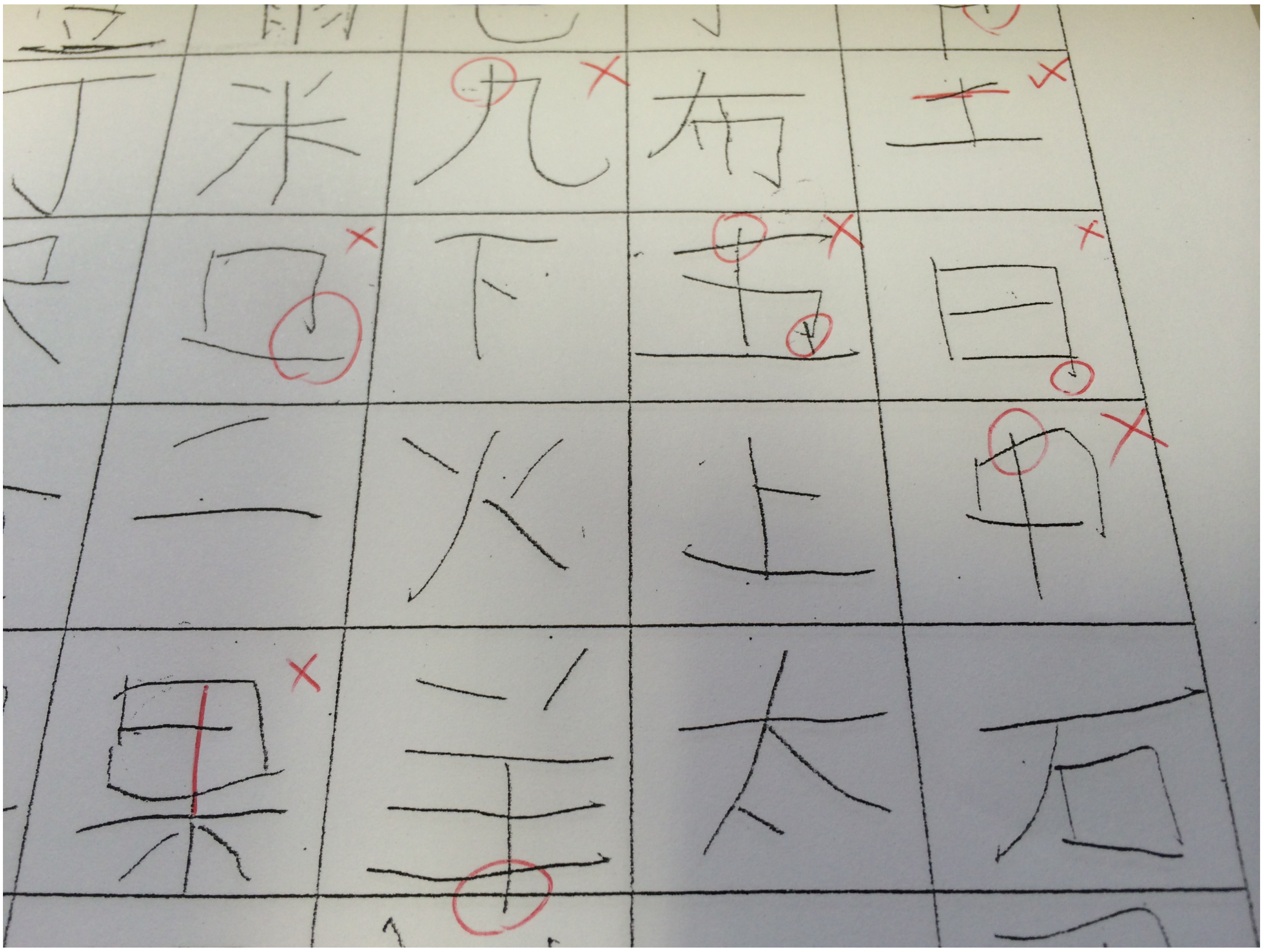


Figure 5.1. Sample of graded penmanship from elementary school child, with teacher-annotated error.

highlighted to provide immediate feedback. To motivate the children and to add an element of fun, correctly written characters are "rewarded" with their corresponding pictorial form displayed. Incorrect characters produce a pictorial with missing features or misalignment. This reinforces their learning with immediate, interesting and rewarding feedback. **Figure 5.3** shows an example for the Chinese character 牛 (cow). The character on the left is correct, producing a graphic of a normal-looking cow. The character on the center is badly written, so the computer circles the error (rightmost character) and produces a crooked-looking cow.

In this project, we sought to achieve these contributions: (1) KID, which facilitates children to learn and practice Chinese writing; (2) algorithms for efficiently evaluating the correctness of written character; (3) positive reinforcement feedback to motivate learning; (4) decent user interface design based upon focus group interviews; and (5) preliminary field studies with positive results.

5.2. Methodology

In KID, we adopt a dictionary with 60 common Chinese characters as recommended by the kindergartens operated under the Hong Kong Christian Service. The KID prototypes used a subset of 10 common characters, with various pictorial formats, including “deformed” graphics that correspond to incorrectly written characters. These characters have been selected because they are not overly simple (the character “human” 人 with just two strokes would not be selected) or too complex, and because they are nouns that lead to easy exposition in pictorial form. In future work, we will populate KID with the remaining 50 characters and extend it to include more characters as appropriate. The list of characters selected is shown in **Table 5.1**.

Character	豆	口	牛	羊	果
Meaning	bean	mouth	cow	sheep	fruit
Character	草	土	火	雨	門
Meaning	grass	earth	fire	rain	door

Table 5.1. Characters with pictorial forms in KID.

5.3. Architecture and System Design

Figure 5.2 shows an overview of the structure of our KID system. The raw input captured is a sequence of points produced by the child using a pointing device (a stylus), each point in the form of a tuple <x, y, pressure, time>, indicating the position and the sensed stylus pressure on the tablet. When the stylus is lifted above the tablet, the absence of captured points for the period indicates the pen-lift time, which provides additional information towards the writing style analysis. The set of consecutive points will be merged into a line. Successive connected lines are structured into a stroke. The strokes are then composed into a character. This is depicted in the written character composition on the left, indicating the part-of relationship between entity classes.

In order to perform the mapping more efficiently and accurately, we need to handle imprecision in each of the steps, as shown in the character evaluation module on the right. The first step is to recognize a trace from the set of points captured by the touch screen device. We adopt a moving window to smoothen the trace until there is a sharp turn (Michael Xuelin, Will et al. 2012). The collection of successive lines oriented in a similar manner is considered as a simple stroke. Successive simple strokes are then considered as a whole for consolidation into a compound stroke. This information is then used to determine the stroke level mistakes committed.

Figure 5.3. Visual feedback from KID. A correctly written "cow" character (left) generates a proper-looking cow (second to left). An improperly written character (center) is annotated with the error (second right) and generates a crooked-looking cow (far right).

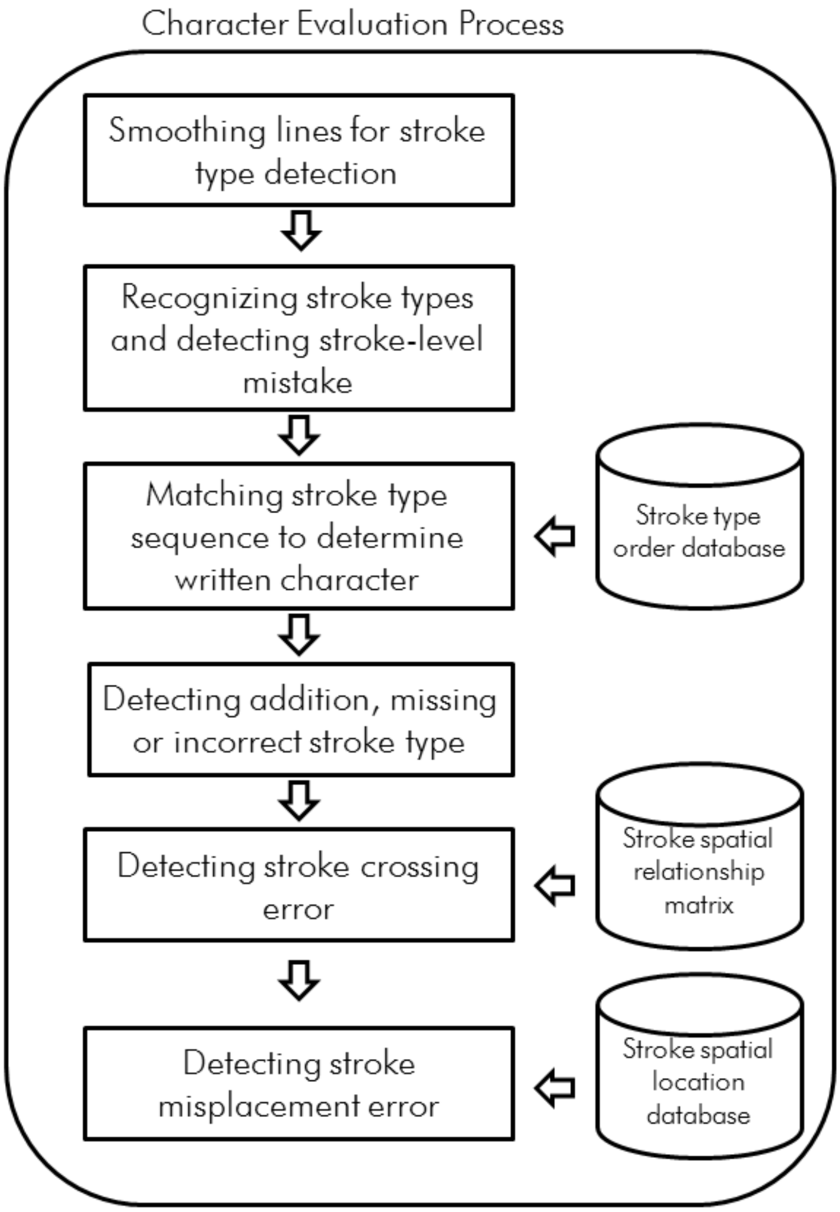


Figure 5.2. System architecture of KID

We use the character 牛 (cow) as depicted in **Figure 5.3** as an example. The first stroke is written as a slightly concave curve going to the left. The variation in orientation indicates the stability of the stroke, quality or even correctness. The second stroke (the top horizontal stroke) should touch the first stroke, but it does not. The third stroke (the second horizontal stroke) should be straight, but it is written as a line inclined at 20 degrees, indicating instability in orientation and a badly written character.

The collection of strokes is then processed against a dictionary of characters for the best matching one. A first filtering step is to use the concatenated stroke order (a sequence) as an index into the dictionary for one or more exactly matched characters. To cater for characters with incorrect, missing or superfluous strokes, we perform character stroke matching for characters whose stroke count falls within a pre-defined tolerance level, e.g., 2 strokes. Using a dynamic programming approach based on the edit-distance measure, we compare the sequence of input strokes and the strokes of a particular character. The set of characters producing the smallest edit distance are considered the candidate set. Discrepancy revealed by the edit-distance matrix provides us the feedback information on the mistake(s) committed by the child, e.g., an insertion error may indicate the presence of extra strokes. The information is then displayed visually as feedback.

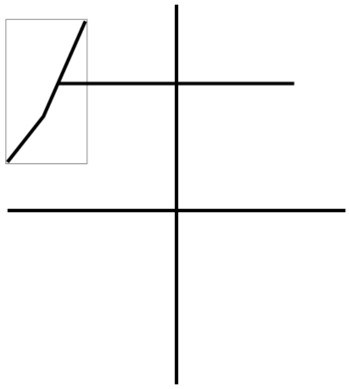
The next step is to resolve the possibility of multiple matching cases, such as two characters that are represented by the same stroke sequence. We take into account the spatial relationship between the strokes of a character to indicate the structure of the character. We build up a stroke spatial relationship matrix for each character. Each element S_{ij} indicates the spatial relationship between stroke i and stroke j, whether they intersect or they touch. The spatial relationship matrix for characters 牛 (cow) and 午 (noon) is shown in **Figure 5.4**, where a blank entry means the strokes are unrelated, t means touching and c means crossing. This matrix is symmetric and relatively sparse.

牛	/	-	-	I
/		t		
-	t			c
-				c
I		c	c	

午	/	-	-	I
/		t		
-	t			t
-				c
I		t	c	

Figure 5.4. The matrices determining spatial relationships of strokes. On the left is 牛 (cow); on the right is 午 (noon). The only difference between the two characters is their stroke spatial relationship.

This matrix allows us to disambiguate seemingly isomorphic characters. A stroke that fails to intersect another stroke when there should be an intersection is easily detected by this spatial relationship matrix. A poorly written character like the one in **Figure 5.3** could be evaluated, where the entry S_{12} in the written character became a blank. The spatial relationship matrix provides a simplified spatial analysis on all the strokes with respect to their proper positions in an actual character. It is highly resilient to character shape distortion error, a very common problem for kids.



Stroke	Bounding Box
/	[0.04, 0.06] x [0.30, 0.45]
-	[0.20, 0.24] x [0.85, 0.26]
-	[0.00, 0.59] x [1.00, 0.61]
I	[0.49, 0.00] x [0.51, 1.00]

Figure 5.5. Using bounding boxes for spatial analysis.

In order to more precisely evaluate how well the written character conforms to the "template" character, we measure the actual spatial location of the strokes for the written character and the candidate character. For each character in our dictionary, we associate a bounding box for each of the strokes, normalized with respect to the extent of the character in $[0,1]^2 \times [0,1]^2$. **Figure 5.5** shows an example for the character 牛 (cow). In a well-written character, the location of the input stroke and the candidate stroke should be close enough. We proceed by computing the bounding box for the input character and the locations of the strokes are then normalized with respect to this character bounding box. We measure the distance between the centers of the bounding boxes of the strokes. If the individual distance and the total distance for all the strokes are below pre-established thresholds, we assume that the strokes of the two characters match in location. This allows KID to detect mistakes resulting from misplacement of strokes that do not touch or cross other strokes, and also allows the extent of each stroke to be compared for the writing quality.

5.4. User Interface

KID has two modes of display. The first is the traditional teaching mode, as would normally be found in similar contemporary software, but with enriched features. The second mode is the pictorial feedback mode to provide more visual feedback and motivate kids for better learning. This is depicted in **Figure 5.6**.

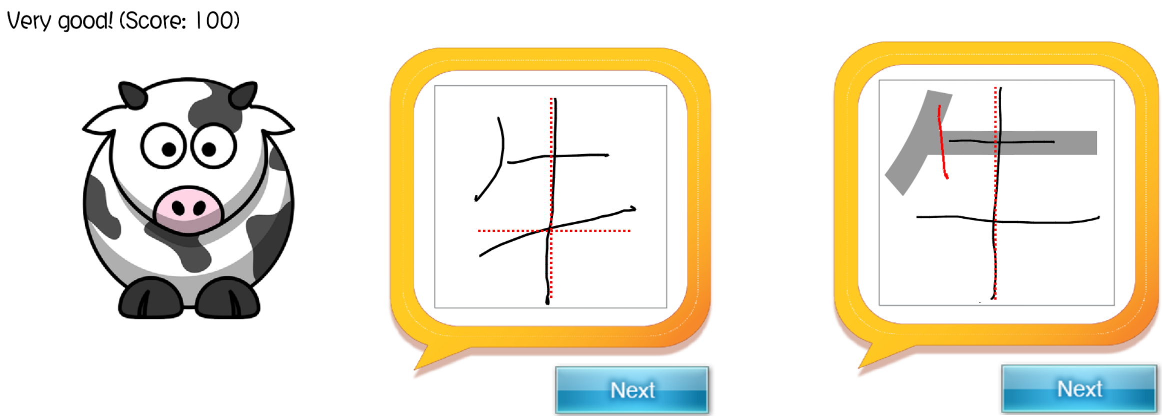


Figure 5.6. Screenshots from KID in operation.

In the traditional teaching mode, after a kid writes the character, KID analyzes the quality of each stroke. When strokes of inferior quality are detected, e.g. lack of "straightness", or uneven pressure along a stroke, a reference "desirable" stroke will be displayed over the problematic stroke (**Figure 5.6 center and right**), so that the kid knows that there is room for improvement on his/her stroke. If the stroke type is incorrect, as detected by the dynamic programming approach, the offending stroke is marked in red (**Figure 5.6 right**). Similarly, misaligned strokes with spatial error, as detected by the spatial relationship matrix, will be marked with a correct reference line. For example, **Figure 5.3 second right** shows a red circle where there is a too-large gap between the first two strokes of the character 牛 (cow). In either case, a correct character will be displayed for comparison. If the strokes are in wrong order, not only the character will be displayed, but also the stroke order along side the strokes of the character (**Figure 5.3 right**).

The pictorial feedback mode applies concepts of metaphoric congruence transfer (Maurer, Pathman et al. 2006, Michael Xuelin, Will et al. 2012) and uses a spatial metaphor between the written character and a pictorial image linked to the character to provide visual feedback on mistakes. When the character is written correctly, a "cute" pictorial representation of the character is displayed with an affirmative sound. However, when a missing stroke is detected, the same image is displayed but with some features missing, e.g., missing eyes or missing feet, depending on where the missing stroke is, together with a denial sound. Extra strokes lead to the same image with displaced features and denial sound, e.g., a cow with a horn. Incorrect stroke order or type will cause the image to become twisted, like the one in **Figure 5.3 right**. When an error is reported, the correct character will be displayed, and an animation displays the proper writing stroke order, stroke by stroke.

5.5. Deployment

We implement the KID prototype on an Android tablet (a Samsung Note 10.1 in our case) with a digitizer surface. To ensure natural writing movement, the touch feature for the tablet is disabled, so that only stylus strokes will be recognized. The digitizer surface can detect the pressure applied to the stylus as well as when the stylus is hovering (i.e. not touching) the writing surface, which allows us to monitor writing defects such as hesitation and uneven pressure.

5.6. Evaluation

We performed a preliminary evaluation on KID to gauge its usability and acceptance among children and their parents. We seek further understanding of the intended user's point-of-view when designing our system features and user interface.

5.6.1. First Prototype (KID-1)

We built the first prototype based on our past experience on developing software with input from frontline teachers.

With the first KID prototype in **Figure 5.6**, we invited two kids, aged 6 and 7, for a pilot exploration. They were asked to write the 10 Chinese characters in each of the two modes. Their accompanying parents were asked to observe their performance, and more importantly, their reaction towards the prototype. They would also ask their own kids which mode they preferred, whether they enjoyed the system, and the part of the system that they did not like. We conducted focus group interviews with the parents to uncover the difficulties that their kids encountered, based on their observation of their kids' performance, as well as talking to their kids. We also listened carefully to their idea of potential cause of the difficulties, and possible suggestions.

From the focus group interview, as well as our own observation of the kids, both kids like the electronic platform, as it is more interesting than writing on paper. Kids did not know where to write on the relatively big screen. Parents commented that they may write really big or really small characters, adding on the “unsure” writing position. As a result, the revised prototype includes an outline bounding box within which the children are supposed to write their character, similar to how they would usually write a character within a gridded box on paper.

Kids also got frustrated if no correct writing form could be displayed when they believed they had written correctly and did not know what / how to correct. There is a discrepancy between what they thought was correct and what was really correct. As such, we have tuned the tolerance level on the stroke type classification for better alignment of user expectation in our modification.

Another issue brought up by the children was that of partial credit. Children generally felt that they should be rewarded if they got their work mostly right, e.g., the character is correct except that the stroke order is not quite correct, or just one stroke is not quite straight but a little misaligned. In other words, an imperfect score of 9 (out of 10) points would be expected! However, the first prototype kept on displaying distorted images with a denial sound. This was perceived as negative feedback and made them unhappy.

We also recognized that the binary pictorial feedback was not sufficient to differentiate between almost-properly-written versus poorly-written characters. There is a degree of transition, when a kid gradually improves his/her writing. If a kid keeps on improving, continually receiving the abnormal image would be quite demoralizing.

To remedy this, we give a grand score to a whole exercise of written character practice on the 10 selected characters. If the overall score is higher than a threshold, a “mostly correct” reward is given. The reward can be of multiple levels, depending on the degree of attainment. In the future, we also plan to reward better writing with nicer visualization, such as adding animations when the level of attainment is high. This is where the bounding box algorithm would be more useful.

5.6.2. Second Prototype (KID-2)

We then conducted a second evaluation with the same kids for the improved prototype. They seemed more engaged and enjoyed writing more. They liked the demonstrative animation of how to write the character strokes correctly. With the bounding box introduced, they now often wrote the upper left part of the character larger than the other parts, due to a lack of sense of space. The

Figure 5.7. A child playing with KID.

result is characters that are out-of-bounds or disproportionate. Parents suggested that a 3x3 grid could be laid in the background to help, and more feedback would be helpful for poorly written characters until improvement is achieved. A guidance or reward system is suggested, so as to avoid kids exploring the written mistakes and playing around with the malformed feedback images.

5.6.3. Evaluation Summary

In summary, our initial evaluation brings us important feedback to improve on the prototype features as well as the user interface design. We believe that the first round of interview has enabled us to build a useful and user-friendly KID to teach kids the challenging task of learning Chinese characters, as evidenced by results in the second evaluation.

5.7. Conclusion and Future Work

Our KID-1 prototype is able to help to motivate children to learn to write Chinese character by providing real-time interaction and in the form of visual and audio feedback. We conducted a focus group interview and evaluation by subjects in unearthing issues and possible solution in implementing an improved KID-2, which is greeted with positive feedback.

In future, we would like to extend KID-2 further with animations, a larger set of vocabulary, and user design feedback for usability. More dimensions in providing visual feedback will be explored, for better accuracy and better motivation for children. We would also like to extend our analysis towards the writing characteristic from the collected data to build a model for children learning to write Chinese characters. This intelligent tutoring approach enables us to detect abnormalities in writing and automatically devise corresponding training modules for the kids. Statistical results and evaluation would also be provided to inform their parents of potential learning problems.

Upon building a more comprehensive KID, we would like to conduct large-scale experiments on kindergarten children. Eventually, we would like to explore its usability on children with dyslexia, making use of the writing model we built and additional data collected specifically with these children.

6. CONCLUSION AND FUTURE WORK

Three different investigations on the transformation of multimodal human-computer interaction have been conducted, i.e. writing to music, motion to music and writing to alternate visual feedback. With different use cases in mind, methods of transformation has been proposed, implemented and evaluated. Observation suggested that these novel interfaces provides a more intuitive interface for user in their specific context. Implementation and evaluation problems were pointed out to help future researcher on similar multimodal HCI system. Our work demonstrated the feasibility of such mappings and provides directions for further investigations.

Other than the modalities investigated in this project, there are of course other modalities that are commonly used by humans that should be investigated, for instance, head gesture and movement, gaze tracking and facial expression. For the propose of a more intuitive interface in other contexts, these modalities should be further investigated for their specific roles in transformation and combination. One of the main challenges in this project was that the creative/generative nature of the problems makes it difficult to measure the quality of the transformation or even the output modality itself, especially for the music output. Further research on evaluating creative modal transformation should be carried out to address this issue.

BIBLIOGRAPHY

Andry, P., P. Gaussier, S. Moga, J. P. Banquet and J. Nadel (2001). "Learning and communication via imitation: an autonomous robot perspective." Systems, Man and Cybernetics, Part A: Systems and Humans, IEEE Transactions on 31(5): 431-442.

Assaf, F., T. Emmanuel Munguia, S. Sajid, M. Pattie and S. Chris (2005). ReachMedia: On-the-move interaction with everyday objects. Proceedings of the Ninth IEEE International Symposium on Wearable Computers, IEEE Computer Society.

Bass Audio Library. from http://www.un4seen.com/.

Bilotta, E. and P. Pantano (2002). "Synthetic Harmonies: An Approach to Musical Semiosis by Means of Cellular Automata." Leonardo Music Journal 35(2): 153-159.

Boris, S. and D. G. Mark (2010). Ubiquitous drums: a tangible, wearable musical interface. CHI '10 Extended Abstracts on Human Factors in Computing Systems, Atlanta, Georgia, USA, ACM.

Bourguet, M.-I. (2003). Designing and Prototyping Multimodal Commands. INTERACT'03.

Burgoon, J. K., L. A. Stern and L. Dillman (1995). Interpersonal adaptation : dyadic interaction patterns. Cambridge, New York, Cambridge University Press.

Camurri, A. (1995). Interactive dance/music systems. International Computer Music Conference (ICMC).

Cheng, Z., S. Li, W. Danli, T. Feng and W. Hongan (2011). CoolMag: a tangible interaction tool to customize instruments for children in music education. Proceedings of the 13th international conference on Ubiquitous computing, Beijing, China, ACM.

Craig, G. N.-M. and H. W. Ian (1997). "Identifying hierarchical structure in sequences: a linear-time algorithm." J. Artif. Int. Res. 7(1): 67-82.

Demey, M., M. Leman, F. Bossuyt and J. Vanfleteren (2008). The Musical Synchrotron: using wireless motion sensors to study how social interaction affects synchronization with musical tempo. New Interfaces for Musical Expression (NIME), Genova, Italy.

Effenberg, A. O. (2005). "Movement sonification: Effects on perception and action." MultiMedia, IEEE 12(2): 53-59.

Eric, N., M. Dan and B. Sumit (2009). Data-driven exploration of musical chord sequences. Proceedings of the 14th international conference on Intelligent user interfaces, Sanibel Island, Florida, USA, ACM.

Eriksson, M., K. A. Halvorsen and L. Gullstrand (2011). "Immediate effect of visual and auditory feedback to control the running mechanics of well-trained athletes." Journal of Sports Science 29(3): 253-262.

Feder, K. P. and A. Majnemer (2007). "Handwriting development, competency, and intervention." Developmental Medicine & Child Neurology 49(4): 312-317.

Fuhrmann, A., J. Kretz and P. Burwik (2013). Multi Sensor Tracking for Live Sound Transformation. Proc. of the International Conference on New Interfaces for Musical Expression.

Gemperle, F., C. Kasabach, J. Stivoric, M. Bauer and R. Martin (1998). Design for wearability. Second International Symposium on Wearable Computers.

Godbout, A. and J. E. Boyd (2010). "Corrective sonic feedback for speed skating: a case study." Proceedings of the 16th international conference on auditory display: 23-30.

Hansen, K. F. and R. Bresin (2003). DJ scratching performance techniques: Analysis and synthesis. Stockholm Music Acoustics Conference.

Hatfield, E., J. T. Cacioppo and R. L. Rapson (1994). Emotional Contagion, Cambridge University Press.

HyperScore. from http://www.hyperscore.com/.

Ian, S., M. Dan and B. Sumit (2008). MySong: automatic accompaniment generation for vocal melodies. Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, Florence, Italy, ACM.

Jack, S. and P. Philippe (2009). Reinforcement Learning of Listener Response for Mood Classification of Audio. Proceedings of the 2009 International Conference on Computational Science and Engineering - Volume 04, IEEE Computer Society.

Jones, K. (1981). "Compositional Applications Of Stochastic Processes." Computer Music Journal 5(2): 45-61.

Kang, L. and H.-Y. Chien (2010). Hé: Calligraphy as a Musical Interface. Proceedings of the International Conference on New Interfaces for Musical Expression (NIME).

Kazuhiro, J. (2008). DrawSound: a drawing instrument for sound performance. Proceedings of the 2nd international conference on Tangible and embedded interaction, Bonn, Germany, ACM.

Lilypond. from http://www.lilypond.org.

Lindenmayer, A. (1968). "Mathematical models for cellular interactions in development I. Filaments with one-sided inputs." Journal of Theoretical Biology 18(3): 280-299.

Lo, K. W. K., W. W. Tang, H. V. Leong, A. Chan, S. Chan and G. Ngai (2012). i*Chameleon: A unified web service framework for integrating multimodal interaction devices. Pervasive Computing and Communications Workshops (PERCOM Workshops), 2012 IEEE International Conference on.

Madeline, B., F. Geraldine, G. Judith and H. Eric (2011). Enhancing interactional synchrony with an ambient display. Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. Vancouver, BC, Canada, ACM.

Marek, P. M., S. Selma and K. Hideki (2007). A dancing robot for rhythmic social interaction. Proceedings of the ACM/IEEE international conference on Human-robot interaction. Arlington, Virginia, USA, ACM.

Marek, P. M., S. Selma and K. Hideki (2007). A dancing robot for rhythmic social interaction. Proceedings of the ACM/IEEE international conference on Human-robot interaction, Arlington, Virginia, USA, ACM.

Mathews, M. V. and F. R. Moore (1970). "GROOVE—a program to compose, store, and edit functions of time." Commun. ACM 13(12): 715-721.

Maurer, D., T. Pathman and C. J. Mondloch (2006). "The shape of boubas: sound-shape correspondences in toddlers and adults." Developmental Science 9(3): 316-322.

Meisner, E., S. Sabanovic, V. Isler, L. R. Caporael and J. Trinkle (2009). ShadowPlay: A generative model for nonverbal human-robot interaction. Human-Robot Interaction (HRI), 2009 4th ACM/IEEE International Conference on.

Michael, X. H., T. Will, W. K. L. Kenneth, C. K. Lau, N. Grace and C. Stephen (2012). MelodicBrush: a cross-modal link between ancient and digital art forms. CHI '12 Extended Abstracts on Human Factors in Computing Systems. Austin, Texas, USA, ACM.

Michael Xuelin, H., W. W. T. Will, W. K. L. Kenneth, C. K. Lau, N. Grace and C. Stephen (2012). MelodicBrush: a novel system for cross-modal digital art creation linking calligraphy and music. Proceedings of the Designing Interactive Systems Conference, Newcastle Upon Tyne, United Kingdom, ACM.

Miranda, E. (2001). Evolving cellular automata music: From sound synthesis to composition. Workshop on artificial life models for musical applications (ALMMA).

Nierhaus, G. (2009). Algorithmic Composition: paradigms of Automated Music Generation, Springer.

Ponsford, D., G. Wiggins and C. Mellish (1999). "Statistical learning of harmonic movement." Journal of New Music Research 28: 150-177.

Qian, Z. and D. Fang (2007). "Towards Chinese Calligraphy." Macalester International 18.

Robert, R. (1992). Interactive music systems: machine listening and composing, MIT Press.

Sami, R., H. Jonna, kkil, K. Saana, C. Ashley and L. Jukka (2007). Tap input as an embedded interaction method for mobile devices. Proceedings of the 1st international conference on Tangible and embedded interaction, Baton Rouge, Louisiana, ACM.

Scheflen, A. E. (1964). "The Significance of Posture in Communication Systems." Psychiatry 27: 316-331.

Sergi, J., G. Günter, A. Marcos and K. Martin (2007). The reacTable: exploring the synergy between live music performance and tabletop tangible interfaces. Proceedings of the 1st international conference on Tangible and embedded interaction, Baton Rouge, Louisiana, ACM.

Sonic wire sculptors. from http://sws.cc/index.html.

Steedman, M. J. (1984). "A Generative Grammar for Jazz Chord Sequences." Music Perception: An Interdisciplinary Journal 2(1): 52-77.

Tetsuaki, B., U. Taketoshi and T. Kiyoshi (2007). Freqtric drums: a musical instrument that uses skin contact as an interface. Proceedings of the 7th international conference on New interfaces for musical expression, New York, New York, ACM.

Tina, B. and P. Tim (2000). The Jam-O-Drum interactive music system: a study in interaction design. Proceedings of the 3rd conference on Designing interactive systems: processes, practices, methods, and techniques, New York City, New York, USA, ACM.

Tseng, M. H. and S. M. Chow (2000). "Perceptual-motor function of school-age children with slow handwriting speed." American Journal of Occupational Therapy (AJOT) 54(1): 83-88.

Varni, G., A. Camurri, P. Coletta and G. Volpe (2009). Toward a Real-Time Automated Measure of

Empathy and Dominance. International Conference on Computational Science and Engineering.

Varni, G., G. Dubus, S. Oksanen, G. Volpe, M. Fabiani, R. Bresin, J. Kleimola, V. Välimäki and A. Camurri (2012). "Interactive sonification of synchronisation of motoric behaviour in social active listening to music with mobile devices." Journal on Multimodal User Interfaces 5(3-4): 157-173.

Varni, G., M. Mancini, G. Volpe and A. Camurri (2011). "A System for Mobile Active Music Listening Based on Social Interaction and Embodiment." Mobile Networks and Applications 16(3): 375-384.

Wanderley, M. M. and P. Depalle (2004). "Gestural control of sound synthesis." Proceedings of the IEEE 92(4): 632-644.

Zeljko, O. (2005). A flexible system for creating music while interacting with the computer. Proceedings of the 13th annual ACM international conference on Multimedia, Hilton, Singapore, ACM.