Taking Sides: Dynamic Text and Hip-Hop Performance

Jason Lewis
Obx Labs, Concordia University
1455 de Maisonneuve Blvd. W.
Montreal, QC H3G 1M8
1 514 848 2424 x4813
jason.lewis@concordia.ca

Yannick Assogba
Obx Labs, Concordia University
1455 de Maisonneuve Blvd. W.
Montreal, QC H3G 1M8

yannick.assogba@yahoo.com

ABSTRACT

In this paper we describe *Taking Sides*, a performance using a real-time speech visualization software system called TextEngine. *Taking Sides* is a collaboration between our research studio and Montreal hip-hop artist Dwayne Hanley. Our primary goal was to create a strong conceptual link between the text visualization, the content of the artist's lyrics, and his performance style. Additionally we wanted to test the flexibility of TextEngine in developing customized performance applications. Pursuing these goals led us through a three month development effort that cycled tightly between design, performance and programmatic iterations.

Categories and Subject Descriptors

J.5 [**Computer Applications**]: Arts and Humanities – *Performing arts.*

General Terms

Performance, Design, Human Factors

Keywords

Real-Time Media, Speech Visualisation, Hip-Hop, Rap

1. INTRODUCTION

The integration of real-time computational media with the performing arts is challenging for a multitude of reasons. The aesthetic and conceptual concerns that are paramount in those performing arts involving the body can make the introduction of a computer system feel intrusive. However, as advances in computer technology allow for a more seamless integration with human action, augmenting the performing arts with computational media is increasingly feasible and desirable.

Taking Sides explores several possibilities for interfacing computational media with performing arts practice, in this case rap. Rap provides a perfect vehicle for this exploration because its highly lyrical and poetic nature complements our interest in working with text both as language and image. *Taking Sides* is built upon TextEngine [8], which provides support for the real time acquisition, recognition, analysis and display of spoken word performance. TextEngine in turn is built upon NextText [7] which provides facilities for creating and rendering text with dynamic visual behaviours. This project attempts, through the use of dynamic text, to create a complementary visual form that captures conceptual and semantic aspects of the rap battle, a genre of live hip-hop performance.

2. RELATED WORK

This project is informed by previous work at our studio, particularly *Intralocutor* [6] which attempts to visualize the dynamics of everyday conversation. While *Intralocutor* is geared more towards spontaneous un-choreographed participation from members of the general public, *Taking Sides* has been developed for a particular performer.

J. Andrews work *Nio* [2] raises similar conceptual issues, particularly those of written representation of acoustic events. In *Nio*, the artist created letterforms and animations to form visual poetry representative of abstract sound poems that he had recorded earlier. The user was then able to trigger sounds and their corresponding animations to create new compositions. However, the content of this work focused on the sonic, rather than semantic, content of speech and its relation to image. Additionally, although it is interactive and the compositions generated by users are unique, the sounds, images and animations have all been pre-produced.

Re:mark by G. Levin and Z. Lieberman [5] is an installation that employed phoneme recognition and other signal processing techniques to capture aspects of participants' speech which were used to create visualizations comprising text and abstract shapes. Like *Nio,* it focuses on the sonic qualities of speech rather than the semantic. Their later work, *Messa di Voce* [4], is also abstract in its sonic and visual content but relates to our work in the sense that is a collaboration between Levin, Liebermann and established vocal artists Jaap Blonk and Joan La Barbara.

Another related work is *Generative Audiovisual Systems* by J. Kreft [3] which, like *Taking Sides*, is targeted towards generating visuals for live hip-hop performance. However, Kreft's work is strongly graphic- and image-based, whereas ours is more textual and focused on language.

MM'06, October 23–27, 2006, Santa Barbara, California, USA.
Copyright 2006 ACM 1-59593-447-2/06/0010...$5.00.

3. DESCRIPTION

3.1 Technical Description

Taking Sides consists of human and computational elements. The human element consists of a live performer enacting a rap battle. A rap battle is a form of competitive rap in which two contestants compete against each other to establish their superior lyrical prowess. This is usually in the form of the rappers taking turns to insult or disparage each other in rhyme. However, in our performance there is only one performer, who creates and enacts the two 'characters' that take part in the battle. This is expressed by the performer in a change of voice and differences in the content of the verses for the two characters that reflect their respective demeanour. There are two verses per character as well as a chorus shared by both.

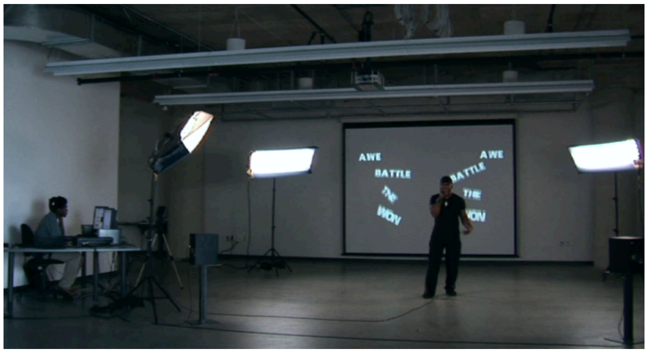

Figure 1: Wide shot of system being used for performance

The computational part of the system can be subdivided into capture and visualization components. The capture component consists of a video capture system and speech recognition system. A custom visualization was built using NextText to animate the text and project the final output; loudspeakers are used to diffuse the sound. (Figure 2)

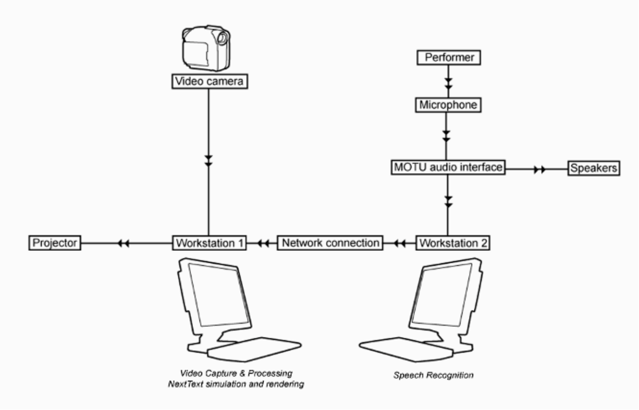

Figure 2: System diagram.

3.1.1 Capture

The video capture system utilizes a consumer-grade digital camcorder which feeds into a custom video signal processing module within TextEngine to extract the performer's body. The camera is oriented in such a manner as to capture the performer's profile. In order to extract the performer's body from the video stream, we use an implementation of the object segmentation algorithm described by A. Amer in [1].

For speech recognition we use the commercially available Dragon Naturally Speaking 8 Software Development Kit from Nuance. This software allows us to capture the utterances of the performer in real time, with a reasonable (approx. 75% after training) accuracy rate.

3.1.2 NextText text visualization library

The NextText library is a Java based library developed by our lab for the creation and visualization of text with dynamic behaviours. Behaviours are a set of rules that determine what a portion of text should be doing. A core library of behaviours is provided with the library, and these are added to and combined in various ways by each application that uses the library. NextText stores text in a hierarchy that allows us to manipulate text at various levels, from the individual glyphs all the way up to sentences and paragraphs. Additionally facilities are provided for the inclusion of rendering processes not related to text, such as the rendering of a performer's silhouette.

3.1.3 Visualization

The figure of the performer extracted from the video stream is turned into a silhouette for projection, and appears on different sides of the screen depending on which character is currently delivering a verse. As the performer's figure is captured in profile, the projected silhouette appears to be facing the side of the screen opposite to where it is standing i.e. towards its opponent. Because the screen starts out black, the performer's silhouette is initially invisible, and becomes more visible over the course of the performance as the background becomes filled in with text.

The words generated by the performer's speech emerges from the silhouette's head and moves to the opposite side of the screen, where they stack up in a loose column, partially overlapping words already present in the column. As words come to rest they briefly brighten and then fade slightly to become translucent. The result is that each word is emphasized for a brief moment after which, due to the overlap, it shifts into near illegibility. When a column is filled, a new column is formed next to it that grows in the opposite direction. This continues for the remainder of the verse, with the result that the background in the opposing character's space fills up with text (Figure 3).

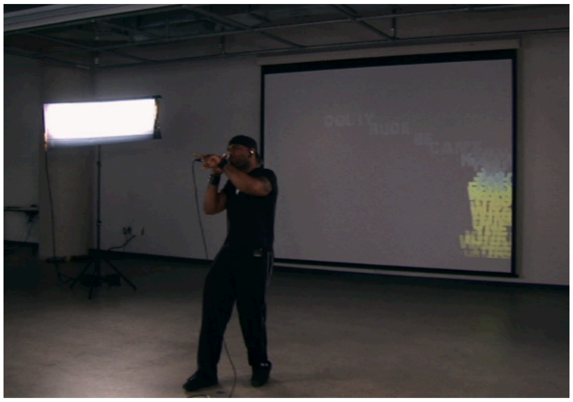

Figure 3: First character delivering a verse with silhouette invisible.

When the verse changes, the silhouette disappears from one side and appears on the other. The opposing character performs his verse in response and the text of his verse is differentiated from that of the previous character by colour. It is also in this moment that the silhouette is first fully revealed, as it now appears on top of the text delivered previously by the other character (Figure 4).

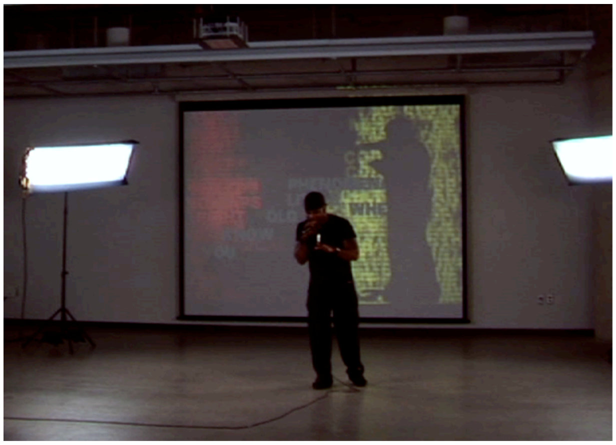

Figure 4: Second character delivering a verse.

During the chorus a different behaviour is applied to the text and to the silhouette of the performer. The silhouette is duplicated and displayed as an outline on both sides of the screen; the text of the chorus is also duplicated and displayed emerging from the heads of both silhouettes. The chorus text hovers in front of its respective silhouette for a few seconds and then tumbles, slightly off balance, until it falls off the screen (Figure 5).

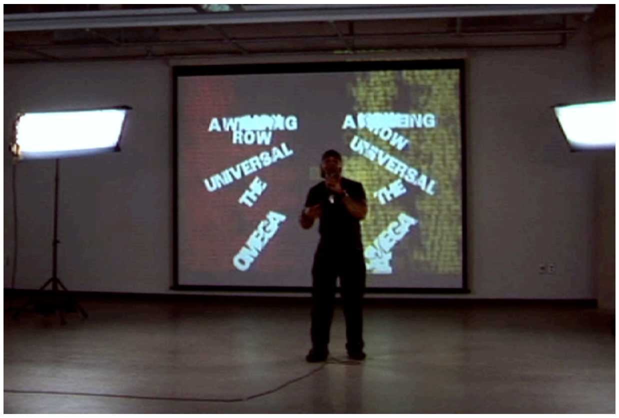

Figure 5: Chorus behaviour.

3.2 Conceptual Description

In this section we shall describe the conceptual and artistic motivations for the choices made in the design of the visualization.

The motion of the text from one side to the other reflects the back-and-forth exchange of the rap battle. The text builds up on either side as it attempts to assert its dominance over the performance and the opposing character. This slow growth of the text masses towards the center echoes the increase in tension and energy of the live performance as the battle goes on (Figure 6).

The masses of text also serve a secondary purpose as they are used to reveal the characters silhouettes; in effect the silhouette is only viewed through the words of its adversary, similar to how a human battler is perceived by the audience through the words said about him by his opponent. The behaviour used during the chorus was constructed to exhibit the fact that the lyrics of the chorus are the only parts of the song that are self-referential. As such the words do not travel towards the opposing character but instead fall at the feet of the character that uttered them, growing slowly as they fall to reflect the bravado of the statements made.

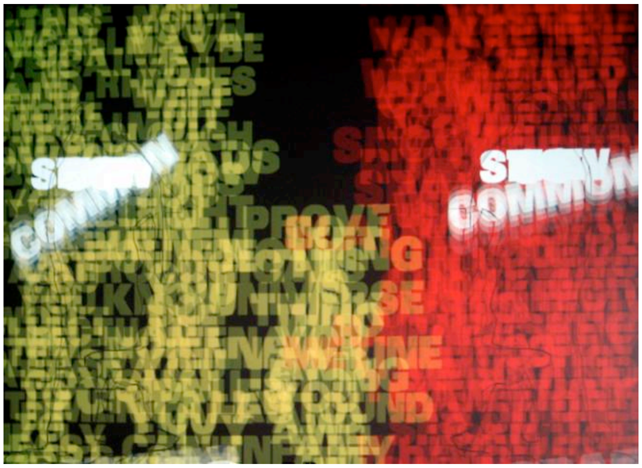

Figure 6: Screen-capture sequence showing resulting build up of text.

We also experimented with the boundary between legibility and illegibility of the text. In creating the visuals we did not wish to create a direct inscription of the performance that could exist separate from the performance itself. Rather, we wished to create a visual representation of the performance that shared in its dynamic and temporal aspects, including the difficulty in making out the lyrics of a live rap performance (particularly if one has not heard the song before). To this end text moves from the legible, allowing the audience to read in written form what they might have missed in spoken form, toward the illegible, where the words form a texture and begins to act more as an image than as language.

4. DISCUSSION AND FUTURE WORK

We feel that *Taking Sides* was successful from a number of perspectives. Most important was the positive feedback we received from the artist that we developed the piece with. He feels the visuals integrate well with the nature of the performance and bring something new to the genre. Success in this regard can be partially attributed to the participatory design approach that was used in developing the work. This approach also allowed us to deal with certain challenges posed by the artist, and the limits of the technology, gracefully.

For example, as we experimented with the speech recognition software, we noted that it had difficulty keeping up with Dwayne's rap style because of the lack of pauses in his speech and the speed at which he could rhyme. Thus in writing the final song used for the performance, the artist knew what tempo he could perform with successfully and could also adjust the length of his bars to control the size of the audio chunks that would be processed by the speech recognition software. This adjustment in speed also helped give the speech and the visualized text better temporal coherency.

We also had to adapt the visualization system to deal with the large number of words generated by Dwayne's performance. The NextText library is a vector based library that had never previously needed to deal with such a large number of glyphs simultaneously as was presented by the songs lyrics. Our solution in this case was to transform glyphs that no longer needed to be animated into bitmap form and composite them onto a single image, which was then rendered behind the animated glyphs and silhouette. The seamless transformation allowed NextText to maintain its performance and support the large number of words that needed to be displayed.

TextEngine's foundation on the NextText text visualization library, and its integration of speech and video recognition modules, enabled us to prototype quite rapidly in response to the artist's aesthetic desires. The ease with which we could do this allowed us to spend most of our time focused on design rather than technical issues.

The artist's satisfaction with the outcome has encouraged us to engage in another series of collaborations with outside artists. We are currently working with another rapper, a typographer/poet and a performance poet to create two additional performances. We are particularly keen on getting the performances out of the studio by configuring the system in such a way that we can rapidly deploy it in venues such as small clubs where rap and spoken word performances often occur.

5. ACKNOWLEDGMENTS

We would like to thank Dwayne Hanley, Alex Taler, Maroussia Levesque and Lysanne Bellemare for assistance in developing this project. This work was supported by Hexagram and the Fonds québécois de la recherche sur la société et la culture.

6. REFERENCES

[1] Amer, A. Object and Event Extraction for Video Processing and representation in On-Line Video Applications, http://users.encs.concordia.ca/~amer/paper/phd/

[2] Andrews, J. Nio, http://www.vispo.com/nio/

[3] Kreft, J. Generative Audiovisual Systems, http://www.deinlieblingsgestalter.de/

[4] Levin, G. and Lieberman, Z. Messa di Voce, http://www.tmema.org/messa/

[5] Levin, G. and Lieberman, Z. Re:mark, http://www.flong.com/remark/index.html#remark

[6] Lewis, J. Intralocutor, http://obx.hybrid.concordia.ca/research/semantics/intralocutor/research_intralocuteur.htm

[7] Lewis, J. NextText, http://obx.hybrid.concordia.ca/research/nexttext/research_nextext.htm

[8] Lewis, J. TextEngine, http://obx.hybrid.concordia.ca/research/nexttext/textengine/research_textengine.htm