Tools for Expressive Text-To-Speech Markup

Erik Blankinship
M.I.T. Media Laboratory
20 Ames St.
E15-311
Cambridge, MA 02139-4307, USA
Tel: 1-617-253-8142
E-mail: erikb@media.mit.edu

Richard Beckwith
Intel Corporation
JF3-377
2111 N.E. 25th Avenue
Hillsboro, OR 97124-5961, USA
Tel: 1-503-264-4760
E-mail: richard.beckwith@intel.com

ABSTRACT
This paper describes handicapped accessible text-to-speech markup software developed for poetry and performance. Most text-to-speech software allows the user to select a voice, but provides no control over performance parameters such as rate, volume, and pitch. For users with vocal disabilities, the default "computer voice" is often dreaded since it provides no personalization. Evolving standards exist for text-to-speech markup (Sable, Java Speech Markup Language, Spoken Text Markup Language), but few tools exist for non-experts to modify documents using these prosody options [1, 5]. Furthermore, we could find fewer tools allowing for straightforward live performance using a synthesized voice [3]. Thus we created an easy to learn text-to-speech markup tool that requires little training to use.

KEYWORDS: Assistive technology, text-to-speech, universal design.

INTRODUCTION
Art Honeyman is a well-known and published poet in the Portland, Oregon area. Art is also severely handicapped by cerebral palsy, limiting his vocal capabilities (speaking a word can take upwards of a minute) and his physical dexterity to just his right foot. With the assistance of the Portland State University's Assistive Technology Center, Art had tried using existing text-to-speech software (DECtalk in an embedded device) to create marked-up readings of his poetry. It was a laborious process given Art's handicap. He had to type XML-like brackets with his spastic hands to specify pitch changes and then close them, all the while avoiding confusing syntax errors that rendered his poetry unspeakable.

We met Art in January of 2001 and saw an opportunity to build a tool for his special needs. We proposed building a "poetry machine" he could use to easily markup the performance parameters of his poetry. Furthermore, we proposed that our tool be easy enough for Art to use in live performances.

FIELD RESEARCH
We knew that existing text-to-speech systems were inadequate but we needed to discover in which ways these systems were inadequate. Furthermore, we needed to understand what would make for an adequate system. We interviewed poets in the Portland area about their preparations for performance, their review process, and what they listen for in their performances. We also attended a large number of open microphone poetry performances, watched poetry slam competitions on tape, and read poets' writings on the poetry process.

Nearly all of the poets we interviewed spoke about the 'poetry voice', defined as a rising intonation at the end of every line often adopted by beginning poets. None of these poets would use it at this point in their career but all recognized it as a common style. Other poets spoke about the importance of pauses for audience feedback, either dramatic or to wait for the audience to stop laughing after a joke (or to rush ahead if no one laughed). The bottom line was that volume, pitch, pausing, and rate would all need to be user adjustable. While by no means an exhaustive survey of performance theory, our efforts did glean us insight into where we could begin our development efforts.

GRAPHICAL MARKUP OF TTS
Adding countless markup brackets to modify prosody of a text document is an overwhelming task (Figure 1). An apt analogy is trying to create a complex web page without GUI tools like Macromedia's *Dreamweaver* or Microsoft's *FrontPage*; a daunting task for the uninitiated.

```
<PITCH CONTOUR="0.0 104.0;0.029
114.0;0.058 124.0;0.088 129.0;0.382
147.0;0.411 150.0;0.441 150.0;0.470
156.0;0.5 156.0;0.711 161.0;0.941
161.0;0.970 161.0;1 161.0"
```

```
RANGE="90%">Gliding</PITCH>
```

Figure 1: An example of a marked up text-to-speech pitch contour. Relative time and a pitch value are indicated in pairs.

We developed an application, called *Poet Shop*, to let a user graphically modify volume and pitch contours. In this case, two horizontal lines running behind a word represent volume and pitch assignments. Moving a line up causes an increase in volume or a rising in pitch. Moving a line down has the opposite effect. Volume and pitch of a word or line can also be changed with a pop-up control panel, as can their rate (Figure 2).

Figure 2: Differently colored pitch and volume contour lines can be modified with the mouse within each word. A higher line indicates a greater value (i.e., raising the pitch line results with the word being spoken at a higher pitch), while a lower line indicates a lower prosodic value. Volume and pitch arrow buttons raise and lower their corresponding contour lines, while the rate arrow buttons increase or decrease a word's spoken rate (graphically represented by elongated and shrunken representations of the word). "Insert" and "delete" buttons allow for words and lines to be added and removed from the document.

Prosodic contours are often shared across lines of poetry, for notorious example, take the aforementioned 'poetry voice'. Taking our cue from aural cascading style sheets [4], we crafted a few macros that could be applied to a text document to change its performance characteristics.

PERFORMANCE TIME MARKUP OF TTS

Art is adept at using a trackball to control his PC, but is especially impressive when steering his wheelchair by joystick with his right foot. We decided to take advantage of Art's pedal dexterity in our software design. We added joystick support to our application, letting a user modify prosody with the joystick as the poem is read aloud by the text-to-speech engine. In this way, the joystick also allows the user to "perform" poetry. As the text document is read aloud by the text-to-speech engine, the performer can modify speech parameters with the joystick (ie., move the joystick to the right to make the poem louder, or moving the joystick down to lower pitch for dramatic effect). Holding the joystick button down pauses the text-to-speech engine. The user can specify the mappings of the volume, pitch, and rate contours to the X and Y axis of the joystick.

DIRECTIONS

Our work does not create a more natural sounding synthetic voice, nor is it "smart" in knowing where to place appropriate stress [2]. Rather, we claim it is nearly impossible for a machine to get poetry performance "right", as it is a highly personal endeavor. While cadence, rhyming schemes, alliteration, and other attributes of oral performance could all be programmatically emphasized we assert there will always be a desire for human personalization – especially for those with no voice of their own. The deployment of *Poet Shop* to the Portland State University's Assistive Technology Center is a step in figuring out what elements are most important for personalization and how to make their modification the most accessible.

ACKNOWLEDGEMENTS

Poet Shop adheres to the OGISable 1.3 text-to-speech markup standard. The authors would like to acknowledge the contributions of the Oregon Graduate Institute for the CSLU Toolkit and the University of Edinburgh for the Festival text-to-speech engine, both of which are used in our software's implementation.

REFERENCES

1. Bell Labs (1997). *Bell Labs Text-to-Speech Synthesis*. Available: http://www.bell-labs.com/project/tts/voices-java.html.

2. Cahn, Janet (1998). *A Computational Memory and Processing Model for Prosody*. Doctoral Dissertation, Massachusetts Institute of Technology.

3. Fels, Sidney & Hinton, Geoffrey. (1995). GloveTalkII: An Adaptive Gesture-to-Formant Interface. In *Proceedings of ACM CHI'95 Conference on Human Factors in Computing Systems*, vol. 1, (pp.456-463).

4. Lilley, Chris (1997). *Aural Cascading Style Sheets (ACSS)*. Available: http://www.w3.org/Style/CSS/Speech/NOTE-ACSS.

5. Wouters, J., Rundle, B., & Macon, M. (1999). Authoring tools for Speech Synthesis using the Sable Markup Standard. In *Proceedings of Eurospeech*, vol. 2, (pp. 963-966).