Legere: A Visualizer for Spoken Audio

Alex Lamar
Virginia Tech
Blacksburg, VA, USA
alawi@vt.edu

Timmy Meyer
Virginia Tech
Blacksburg, VA, USA
tmeyer15@vt.edu

Loran Steinberger
Virginia Tech
Blacksburg, VA
loran3e16@gmail.com

Steve Harrison
Virginia Tech
Blacksburg, VA, USA
srh@cs.vt.edu

ABSTRACT
Legere is a work of critical technology-art that examines the intersection between novels and visual media as two different forms of entertainment. It is set to mimic television -- the program, displayed on an old television set, has a set number of channels that the user can flip through with a remote. Each channel concurrently plays a long-running audiobook, and using speech-recognition, the program flashes the book's text at the user in sync with the narration. The exhibit is meant to mock a living room atmosphere by adding a couch, coffee table, and other peripherals like a rug, to the project space.

Author Keywords
HCI demo; speech recognition; audiovisual demo.

ACM Classification Keywords
H.5.m. Information interfaces and presentation (e.g., HCI): Miscellaneous.

INTRODUCTION
Legere is a piece of technology-art that examines the intersection between novels and visual media as two different forms of entertainment. In today's world, the two are quite separated from one another – on one hand, books and novels as a medium are almost exclusively consumed by one individual in a private context (that is, the interaction is private between a single person and the book, others do not and cannot participate because the medium does not lend itself well to group consumption). On the other hand, visual media such as television or movies lend themselves much more towards group consumption – whether it's in a movie theater or a living room, it is easy for a group of people to sit around and watch, much more so than it would be for a group to huddle around and read *To Kill a Mockingbird*.

The purpose of Legere is to break down the barrier-to-entry inherent with books as an entertainment medium – to make books easy to pick up and enjoy in any context. In doing so, we can show a piece that critically examines how the result fits in with modern entertainment mediums as we know them.

5th Decennial Aarhus Conference on Critical Alternatives
August 17 – 21, 2015, Aarhus Denmark

DOI: http://dx.doi.org/10.7146/aahcc.v1i1.21470

SYSTEM
At its heart, Legere is a computer visualization program that takes in an arbitrary audio input along with a transcription, and, after some processing, outputs a time-stamped reading of the transcription in sync with the original track. The result is a program that adds a visual medium to the spoken audio in the form of text flashing on the screen as the audio progresses.

This creates an experience that is actually quite different from simply listening to an audio track. The text on the screen, although only a seemingly slight addition to the audio, creates an important effect where the audience's attention is forced onto a single point in the room. Because of this, it is much harder to get distracted by anything other than the program, as is often the case when listening to pure audio. The visual component encourages the audience to sit down and watch, which is key to creating any engaging medium.

Legere also has the capability to switch between audio tracks with a handheld remote, giving it the loose appearance of a television set that can switch between T.V. channels. It is in this context that the program is used as an exhibit.

Audio Books
The format of the program was designed with audiobooks in mind. The program requires two items to function: a source of audio, and a transcription of that audio. Any unabridged audiobook is guaranteed to have those two things, one being the .mp3 or .wav track, and the other being the original text of the book. The length of the input does not matter, so the program can handle long-running audiobooks in excess of 10 hours.

EXHIBIT
As an exhibit, *Legere* does two things: it creates an environment intimately familiar to the audience (we set up a mock-living room), while displaying an instance of the *Legere* program that teeters between the line of familiarity and strangeness (it is set to play classic novels that most people have at least some familiarity with, but the format, being visual, is inherently unfamiliar to the audience). The program is meant to be engaging first and foremost, but because of its slight unfamiliarity, it also causes the audience to think critically about what exactly they are looking at.

The purpose of the exhibit is to examine exactly how people view differences between something like a television program, and a book or a novel. What makes the two different? It raises questions as to where exactly novels have gone as a once-predominant medium of entertainment, and why exactly they have been replaced, by and large, by movies and television.

But, beyond all of this, the exhibit is *engaging*, and because of that it can very likely act as a mechanism that rekindles an audience member's interest in books and novels, making him ask himself exactly *why* reading has been placed on the backburner in favor of other mediums of entertainment.

TECHNOLOGY

The technology behind *Legere* is quite novel, and is worth dedicating a short description to. The program uses speech recognition technology (the Sphinx4 library, a speech recognition library developed at Carnegie Melon University) to parse audio input into timestamped text. Speech recognition as we know it, however, is *very* far from perfect, and a naïve attempt at parsing audio input using just the recognition library will yield around ~30% recognition accuracy, and will take in excess of a day for a single, hour-long passage, which is more or less unusable for our purposes.

However, by progressively chunking the input audio, and with some clever pre-processing (by guiding small portions of audio with text from the transcription), we can achieve nearer ~80% recognition accuracy along with a linear parsing time (linear, but still faster than real-time). After that, running the transcribed audio through a post-processor that matches it exactly with the transcribed text then gives us perfect to near-perfect accuracy when reading back the text to the user.

Speech recognition is a finicky process, and the field as a whole is still very far from perfect (and it most likely never will be), but it is still capable of being very, very accurate when it is used in novel ways (in this case by the use of an audio transcription).