Automatic Note Taker for the Impaired (ANTI)

Roland Anderson, Patrick Galloway, Casey Miville

Dept. of Electrical Engineering and Computer Science, University of Central Florida, Orlando, Florida, 32816-2450

***Abstract* — The Automatic Note Taker for the Impaired (ANTI) is small robotic system that tracks a professor or presenter using image recognition. The system uses a high fidelity directional microphone to record audio and a static high-resolution camera to record video with post-processed subtitles for redundancy purposes. A connected Windows device has a speech to text software that will make notes for the user. Therefore, the user will not only have the notes from the presentation or class but also audio and video with subtitles for a more detailed account of the event.**

***Index Terms* — Computer vision, face detection, servomotors, embedded software, microphones, webcams.**

I. Introduction

The ANTI device was designed to function in a class sized environment with various external inputs. The device tracks the presenter and is able to pan 180 degrees and tilt within a 90 degree range to cover the presenter's movement. The device has a footprint smaller than 9 inches in diameter. The size is to account for an average sized desk in a classroom that the device will have to be resting on alongside a laptop computer. The device is designed to be able to track movement from over 10 feet and be able to direct the microphone to the target for greater audio pickup. PC software includes a graphic user interface (GUI) with easy to use options for disabled students and needs to be installed on a portable Windows device with a microphone jack. ANTI device is powered through external power supply.

II. Motivation

A need was seen for this low cost portable solution to help students with disabilities. The current paid student note taker system is inherently flawed, because the disabled student is reliant on the note taking skills of another student, whom might make mistakes or not take as detailed notes as possible to help the disabled student better understand the lecture. Some student note takers may ignore or miss something the professor or lecturer stated because they may think it not important or common knowledge. The simple basic fact is when someone is fully relying on notes from someone else's perception or understanding of a subject matter the account of that presentation will always be skewed. This system would fix that flaw by having a detailed comprehensive account of the lecture or event that is thorough and unbiased.

Every semester most of the UCF student body receive emails offering to pay students $150 per class to be note takers for the impaired. This system would save the school a lot of money by offering this low cost, portable, all-in-one solution to multiple students at time, instead of each disabled student waiting on a note taker from his or her class to email the notes.

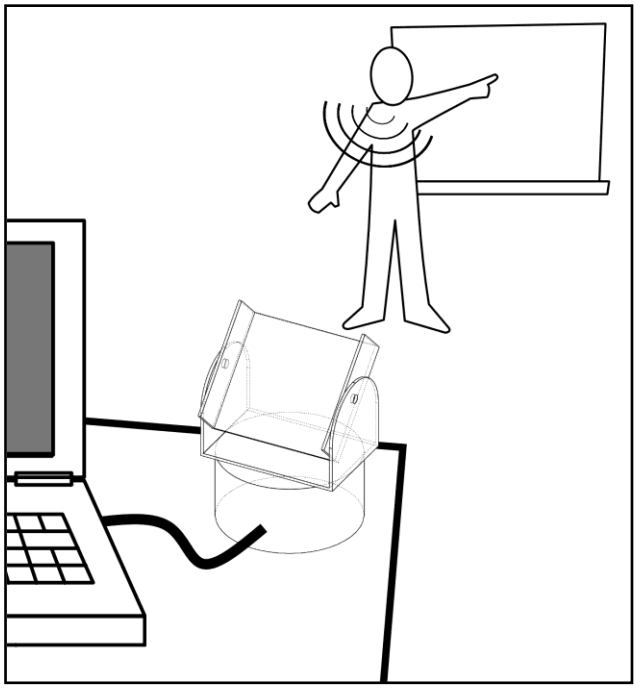


Fig. 1. Initial concept: Laptop PC and ANTI recorder in a classroom setting.

III. Objectives

This device will be modified in stages to add functionality to the finished product. Each phase represents a large leap in software or hardware design and integration. Phase 1 represents the base functionality that this design strives to accomplish while the latter phases are optional additions and bring forth new challenges to be overcome.

A. Phase I

Phase I consists of accomplishing the goal of taking notes in a written form for the user. The main functionality of the project is split between two major devices; The laptop PC and the ANTI recorder are connected and allow the PC to acquire input from the microphone. The ANTI recorder is an independent system that is capable of identifying a presenter, the target, and position the microphone to face the target. The ANTI recorder accomplishes the tracking via an on-board camera and pan/tilt servo base. The microphone is mounted to the pan/tilt head, passes it's output through the ANTI recorder, and connects directly to the laptop PC. The PC is responsible for running a "speech-to-text" software with the microphone as it's input. Phase I requires the software to output a text file of what the target has spoken that can be read at a later time.

B. Phase II

Phase II consisted of adding a static, second camera mounted to the laptop PC. The static camera records the entire scene in front of the user for greater context. Also, in Phase II, the software will attach the sound input to the video file and then uses the "speech-to-text" software to append subtitles to the saved video output. The output is represented in Figure 2 below.

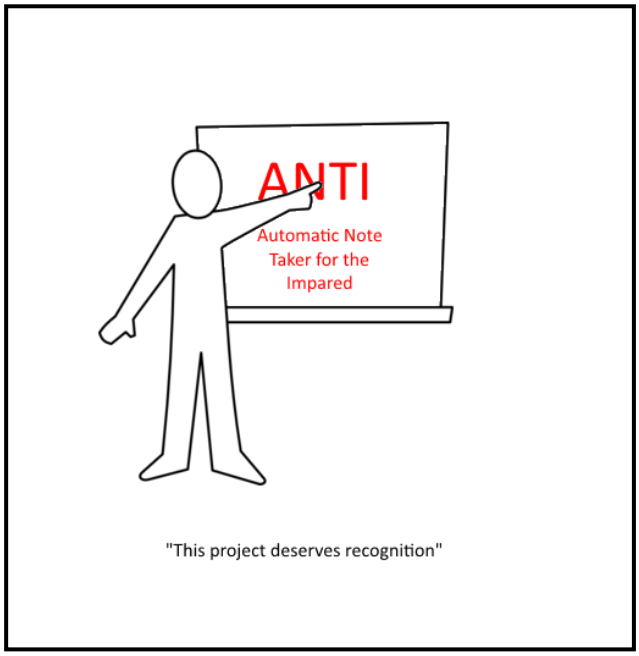


Fig. 2. Initial concept for Phase II output.

IV. Hardware and Specifications

All hardware described in the following subsections is independent of the Windows device. No parts were fabricated or specially ordered for this application. The static webcam mounted to the portable Windows device is interchangeable with any Windows compatible webcam.

A. BeagleBone Black

The BeagleBone Black is a development board by Texas Instruments featuring an AM3359 ARM Cortex A8 microprocessor with a 1 GHz clock rate and 32 bit registers, and has features to increase throughput during video and image processing computations. The BeagleBone Black is running Debian, a distribution of Linux, providing easy integration of many important libraries and communication protocols. This processor and board will perform all of the functions of the ANTI device.

The main function this processor will be expected to perform is video processing using the OpenCV computer vision libraries. To this end, the MPU includes a NEON SIMD (single instruction multiple data) coprocessor. The NEON has 32 registers, each 64 bits wide, which can be used to perform simultaneous computations which do not rely on each other's results. These types of computations are typical in image and video processing, and will allow for much faster processing of the video data to provide more frequent tracking instructions to the motion control class.

The BeagleBone Black has a number of available I/O peripherals included, including USB, SPI, UART, I2C, and 2 PRUs which allow very customizable input and output methods. In this application we are using the USB port to communicate with the webcam and 2 GPIO pins to control the 2 servo motors via pulse width modulation.

A custom pin header (also known as a cape) is attached to the BeagleBone Black with soldered connections for the servo motor signal wires. A ground wire is also connected between signal ground and power ground of the servo motors. Figure 3, below shows the BeagleBone Black with expansion header "cape".

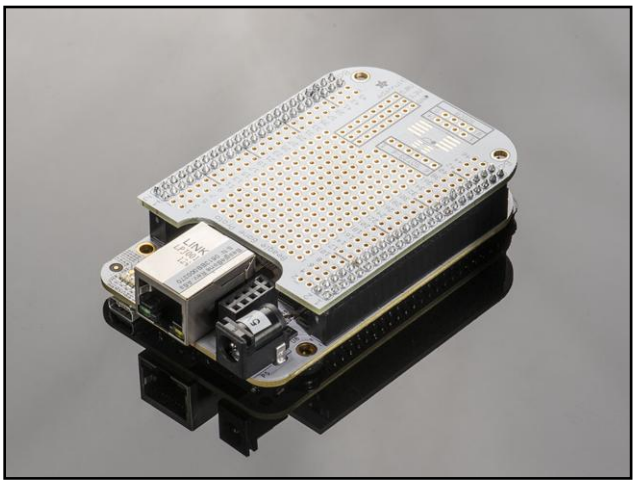


Fig. 3. BeagleBone Black with cape.

Power is provided to the BeagleBone Black through an external wall adapter purchased for this application. The board and custom expansion cape is housed inside an acrylic case to protect the wiring. Holes are cut into the acrylic to provide access with the various ports on the BeagleBone Black.

B. Microphone

Shotgun microphones are long cylindrical devices that have the ability to pick up audio in a tight conical shape at a close range, while filtering out the rest. These types of microphones are used in "boom microphones" that are held just above the performer out of a camera's view box to isolate the audio of the performer in a video. The shotgun microphone is also used in handheld camera applications to record audio in the direction that the lens is facing. These microphones contain two major components that allow it to pick up sound in a single direction, the interference tube and the hypercardioid microphone. These components give the shotgun microphone the ability to capture sound directionally as seen in Figure 4 below.

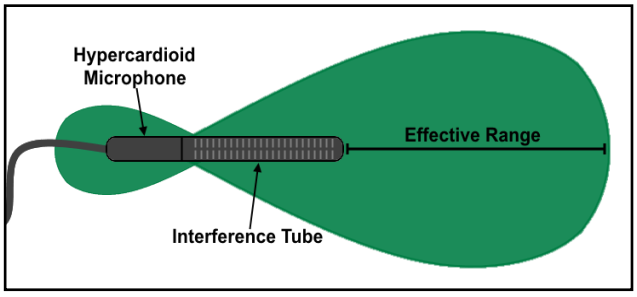


Fig. 4. Shotgun microphone area of effect.

Shotgun microphones have the ability to cancel noise from surrounding inputs, but have a limited range of operation. In the case of the design of this project, if the presenter is a large distance away from the microphone, his voice will still sound faint and distant. Shotgun microphones are quite sensitive, so if the microphone experiences movement or force, the audio quality will not be as clear. These factors entail that the user of the device be close, but not too close, to the presenter when operating.

The ANTI device includes an Audio-Technica ATR6550, and has its specifications listed in Table I below. This device was chosen based upon the description from the manufacturer. In its 'Tele' range setting, this cardioid condenser is engineered to pick up dialogue and sound effects at a distance, while bypassing ambient noise such as traffic, air-handling systems, room reverberation and mechanically coupled vibrations[1].

The microphone is mounted to the head of the pan/tilt device described in subsection D. For easy removal and replacement, the microphone can be detached from the mount.

TABLE I
AUDIO-TECHNICA ATR6550 SPECIFICATIONS

Element	Mono Condenser
Polar Pattern (Modes)	Normal: Cardioid Tele: Supercardioid
Frequency Response	70Hz - 18kHz
Open Circuit Sensitivity	Normal: -56 dB Tele: -45 dB
Impedance	Normal: 1,000 ohms Tele: 2,200 ohms
Cables	1m cable with 3.5mm TRC plug
Weight	4 oz

The output cable of the microphone is a male 3.5mm TRS connection, where TRS stands for "tip, ring and sleeve," which describes the shape of the male audio connector. This cable extends and connects directly into the microphone port on the portable Windows device.

C. Webcams

This project utilizes 2 separate webcam devices. The first is mounted to the head of the pan/tilt arm on the ANTI device while the other is mounted to the portable Windows device. Both devices communicate over a USB connection with their host devices.

The Logitech B910 HD webcam is mounted to the tilt head of the ANTI device alongside the microphone and another is mounted to the mobile Windows device. Webcam specifications are listed below in Table II.

TABLE II
LOGITECH B910 HD SPECIFICATIONS

Field of view	78-degree wide-angle
Color Depth	24-bit true color
Frame Rate	30 frames per second @ 720p and VGA mode
Photo Capture Resolution	5 million pixels
Microphone	Built-in dual microphones
Output Interface	Hi-speed USB 2.0

D. Pan/Tilt Base

The ANTI device is comprised of 2 major moving components; Pan and tilt systems were acquired to allow the device to direct the microphone towards the presenter. Both movements are powers by standard size Hitec HS-322HD servo motors. These motors allow 180° movements by both pan and tilt components. Figure 5 below shows the pan/tilt pre-rendering where object B is the pan base and object A is the tilt head.

The pan base houses one Hitec HS-322HD servo motor and is placed on the table. Atop the pan base, mounted on a bearding shaft, sits the tilt head, which also houses the second servo motor. The tilt head tray holds the Logitech B910 HD webcam and the Audio-Technica ATR6550 microphone. All attached cables are fed through the base pan and to their components.

Fig. 5. Pan/Tilt base.

E. Servo Motors

Two servo motors are required for movement of the Pan/Tilt base as described in subsection D. This is accomplised by 2 Hitec HS-322HD servo motors operating at 5V. Servo motor specifications listed in Table III below (only 4.8V values shown).

TABLE III
HITEC HS-322HD SPECIFICATIONS

Operating Voltage	4.8V-6.0V
Operating Temperature	-20 to +60 °C
Operating Speed(4.8V)	0.19sec/60° at no load
Stall Torque(4.8V)	42 oz.in
Current Drain(4.8V)	7.4mA/idle 160mA no load operating
Connector Wire Length	11.81"
Weight	1.52 oz

Signal wires are run from the BeagleBone Black GPIO pins on the custom header cape to the signal pin of the servo connector.

F. Power Management

Independent power is provided to components in the ANTI device. The BeagleBone Black has power supplied by the aforementioned 5V/2A power adapter purchased to be compatible with the barrel jack of the device. Servo power is provided by a second 5V/2A power adapter. This adapter feeds power to a 5V Step Up/Down voltage regulator, Figure 6 and Table IV, which feed the servos directly via a "Y" cable.

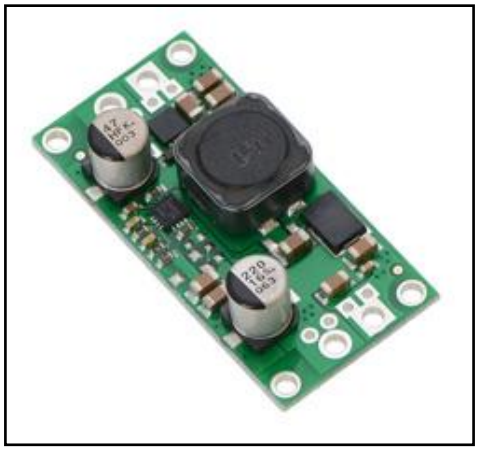


Fig. 6. 5V Step Up/Down Voltage Regulator.

TABLE IV
5V STEP UP/DOWN VOLTAGE REGULATOR SPECIFICATIONS

Operating Voltage	4.8V-6.0V
Operating Temperature	-20 to +60 °C
Operating Speed(4.8V)	0.19sec/60° at no load
Stall Torque(4.8V)	42 oz.in
Current Drain(4.8V)	7.4mA/idle 160mA no load operating
Connector Wire Length	11.81"
Weight	1.52 oz

V. EMBEDDED SOFTWARE

The purpose of this software is to identify and track a singular significant object (in this project the object will be the lecturer) in front of the camera. The secondary purpose is to keep that object in the center of the captured images in order to properly record the object's sound/speech.

The ANTI project has 2 independent software systems working together on separate platforms. The embedded software running on the BeagleBone Black tracks the presenter and moves the microphone while the Windows device runs a transcription program. This section is details the methods used on the BeagleBone Black software.

The BeagleBone Black runs an embedded version of Linux operating system named Debian. The version running on the system is named Wheezy. All code editing is done through the operating system.

Computer vision software is generally a rather memory consuming task, with relatively large secondary memory space needed. This is going to be partially remedied by adding an SD card to give the board secondary memory.

A. Computer Vision Software

Computer vision is a large part of our project and so finding libraries was a large task. OpenCV was the first and obvious choice for a solid computer vision library. Some of our teammates have had experience using the library. OpenCV is also a great environment because it has multiple language support. With it you have to know what type of masking you need and what kind of data you are processing.

The biggest reason to integrate facial tracking for our project is because we want to use a directional microphone to record and decode speech from a single individual. Facial tracking will ensure that we direct the microphone to the source of the desired sound, with the least amount of effort. With facial tracking we are assured that the microphone will be pointed in the most accurate position to insure the highest quality of sound to microphone. If we use blanket object tracking the microphone might not be guaranteed to be pointed to the mouth of the desired individual, (i.e. it could follow a moving hand or leg) and then vital sound quality would be lost and make the speech to text software's job much harder. Another reason for facial tracking is that there are many classifiers on the face that are unique from other objects. Eyes and mouths move quite differently from other objects in a classroom environment, and as such are good things to track to be able to follow a person around.

There are several methods of tracking, facial recognition algorithms are used in these algorithms to identify and to enforce new faces. There is not necessarily a direct translation from facial recognition and facial tracking, Rather, the elements are blended together so that the good points of recognition are used to track. Facial tracking is still a relatively new and developing field, there are many approaches still in development.

There are several methods of facial recognition. The simplest, but least reliable method is color detection. This method utilizes the different shaded regions of the face, i.e. the bridge of the eyebrow, eyelashes, and shading that comes from hair and light. This process is a problem conditions, and different skin tones can drastically alter the accuracy of this method and is not a reliable unit of measure.

Motions detection is another way to do facial detection. This method uses characteristics such as the fact that most blinks have the eyes closing synchronously. It also uses things like mouth movement to determine if the object is a face or not. Just like the color detection algorithm this algorithm is limited in scope and is an unreliable method of facial detection. Researchers did try to remedy this method by also including color detection. While this combined method is better than either individual method, it is still not a robust way to detect faces.

For the applications of this project we chose it use the Haar Feature-based Cascade Classifiers to identify the object to track. The Haar Cascade OpenCV library is used to identify a face from any other arbitrary object. Multiple faces are identified and their locations are determined with respect to the center of the frame. A movement vector is then reported to the servo motor control class. A class diagram details the process in Figure 7 below.

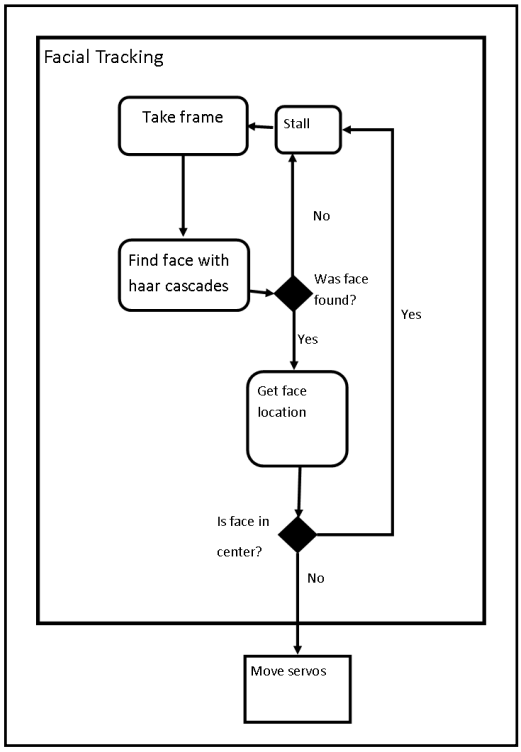


Fig. 7. Facial Tracking class diagram.

B. Servo Control

Hitec HS-322-HD servo motors are controlled using the pulse width modulation (PWM) protocol. The BeagleBone Black uses an open source library to control the servo motors over its GPIO pins.

Target information is passed to the PWM routine by the facial tracking class. Data received is in the form of a pixel coordinate to the center of the presenters face. The PWM routine adjusts the signals sent to the servo motors to center the presenter in the frame and thus point the microphone at the presenter. Feedback is presented to the routine in the form of the next coordinate passed by the facial tracking class. Figure 8 below shows the class diagram for the servo control function.

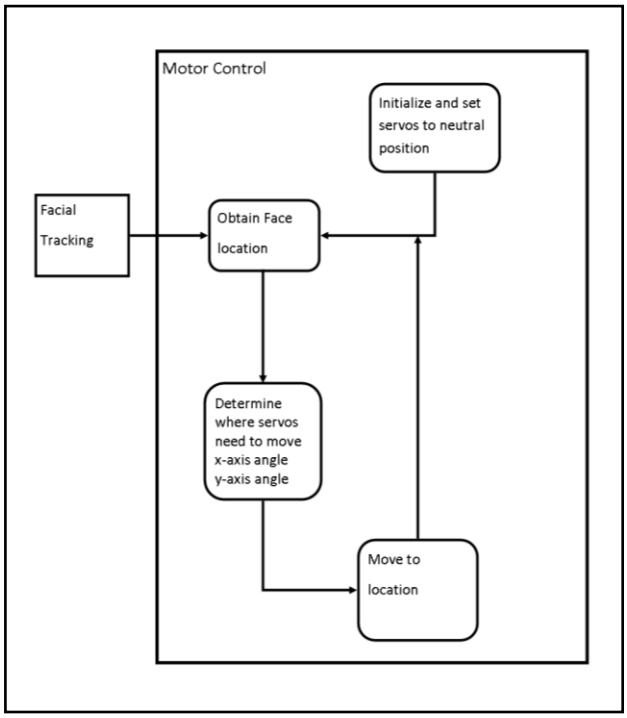


Fig. 8. Servo Control class diagram.

VI. PC SOFTWARE

The purpose of this software is to accept incoming sound from a lecturer and translate that speech into readable text for a user. The goal is to operate in an environment where someone who might have difficulty taking notes for a class has a reliable source to transcribe the class for them.

ANTI depends upon a mobile Windows device (PC) to transcribe the audio stream from the presenter. This audio stream is collected from the microphone that is being directed towards the presenter via the ANTI device. The PC runs an executable file that transcribes the audio stream.

A. Speech Recognition

In speech recognition there are two important parts of the statistical based algorithm, the acoustic and language based modeling. Speech recognition requires two files to recognize speech; an Acoustic model is created by making recoding of speech and their textual representations. Afterwards, software is used to create a statistical representation of sounds that would then make up each word. They also require a language model which is really a grammar file that contains sets of predefined combination of words. This is what a speech recognition engine uses to recognize speech.

Although there may be other avenues of approach the simple fact remains that speech to text is hardware intensive and cannot be implemented on board or hardware only. The best and so far most successful approach to implementing speech to text has been with software. This way a developer can cater to the user's needs by implementing an intuitive graphic user interface that has a low learning curve. As in our case, the user interface is very important since the user being targeted by this project will need a very basic ease of use that requires very little to no instructions what so ever. Developer can ensure the a very fast user experience without any lag and if there arise any issues they can be fix quickly an update or patch. The main goal is make sure the software looks inviting, fun to use and of course functional. The software is installed on the user's mobile Windows device. A class diagram for this software can be seen below in Figure 9.

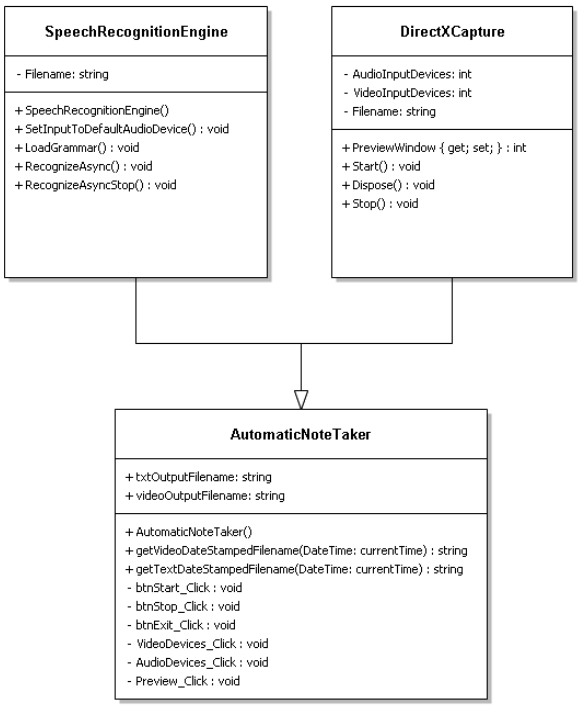


Fig. 9. Automatic Note Taker class diagram.

B. Graphic User Interface

The graphic user interface that was envisioned for this project is a very simple one that is intuitive and easy to use. Due to the fact that this project is targeted to the impaired we wanted to make the learning curve very low. The goal is that a user can intuitively find out how to use the program without any instructions or guidance. When the user presses the start button, the program will automatically start a full screen capture of the audio, video and text on the fly. The text transcription will be saved to a text file for note taking so the student can have a backup. Also in case any problem or errors occur with the speech to text, there will be audio and video saved to the same folder as the text file. Figure 10 below shows the graphic user interface.

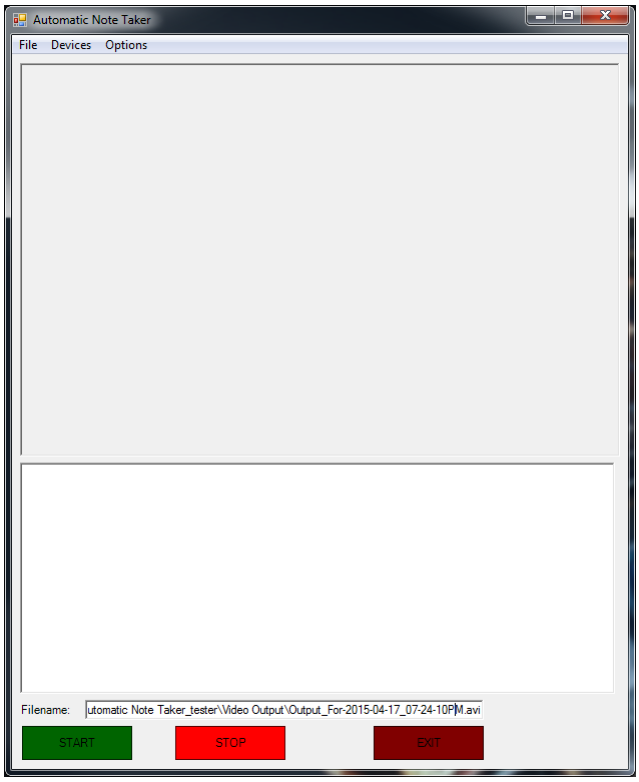


Fig. 10. Graphic User Interface.

VII. Resources

[1] “Shotgun Microphones” - http://www.bhphotovideo.com/explora/audio/buying-guides/shotgun-microphones

[2] "Acoustic Modeling." *Research.microsoft.com*. N.p., n.d. Web. 17 Nov. 2014. <http://research.microsoft.com/en-us/projects/acoustic-modeling/>.

[3] Brokish, Chuck. *U-Law Compression on the TMS320C54x*. N.p.: n.p., n.d. *Ti*. Web. 20 Nov. 2014. <www-s.ti.com/sc/psheets/spra267/spra267.pdf>.

[4] Chu, Wai C. *Speech Coding Algorithms: Foundation and Evolution of Standardized Coders*. Hoboken, NJ: J. Wiley, 2003. Print.

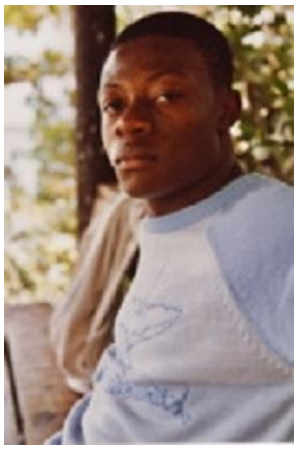
[5] Koehn, Phillip. *Statistical Machine Translation*. Cambridge: Cambridge UP, n.d. Print.

[6] Malcolm Slaney, Elizabeth Shriberg, and Jui-Ting Huang, Pitch-Gesture Modeling Using Subband Autocorrelation Change Detection, in *Proceedings of Interspeech 2013*, ISCA, 29 August 2013

[7] George Dahl, Dong Yu, Li Deng, and Alex Acero, Context-Dependent Pre-trained Deep Neural Networks for Large Vocabulary Speech Recognition, in *IEEE Transactions on Audio, Speech, and Language Processing (receiving 2013 IEEE SPS Best Paper Award)* , vol. 20, no. 1, pp. 30-42, January 2012

[8] Frank Seide, Gang Li, and Dong Yu, Conversational Speech Transcription Using Context-Dependent Deep Neural Networks, in *Interspeech 2011*, International Speech Communication Association, August 2011

VIII. Team Biographies



Roland Anderson, currently a senior at University of Central Florida will receive his Bachelors' of Science in Computer Engineering in May of 2015. He plans on continuing on to the Masters program in Computer Engineering with a focus on Intelligent System and Machine Learning. He is currently a Software Developer Level 1 for Wyle Laboratories.

Patrick Galloway is an computer engineering student planning on graduating May 2015. Patrick plans on working for a year, currently in IT at UCF, and looking for programming work. Then applying for graduate school for either machine learning or computer networks and computer security.

Casey Miville, currently a senior at University of Central Florida will receive his Bachelors' of Science in Electrical Engineering in August of 2015. He will enter the workforce with hope of obtaining an engineering position with an aerospace contractor. He also plans on continuing his education to obtain his Ph. D, become a professor, and run a research lab.