

Hands-free device control using sound picked up in the ear canal

Siddharth R. Chhatpar^a, Lester Ngia^a, Chris Vlach^a, Dong Lin^a, Craig Birkhimer^a, Amit Juneja^a,
Tarun Pruthi^a, Orin Hoffman^b, Tristan Lewis^b

^aThink-A-Move, Ltd., 23715 Mercantile Road, Suite 100, Beachwood, OH, USA 44122-5931;

^biRobot Corp., 63, South Avenue, Burlington, MA, USA 01803

ABSTRACT

Hands-free control of unmanned ground vehicles is essential for soldiers, bomb disposal squads, and first responders. Having their hands free for other equipment and tasks allows them to be safer and more mobile. Currently, the most successful hands-free control devices are speech-command based. However, these devices use external microphones, and in field environments, e.g., war zones and fire sites, their performance suffers because of loud ambient noise: typically above 90dBA. This paper describes the development of technology using the ear as an output source that can provide excellent command recognition accuracy even in noisy environments. Instead of picking up speech radiating from the mouth, this technology detects speech transmitted internally through the ear canal. Discreet tongue movements also create air pressure changes within the ear canal, and can be used for stealth control. A patented earpiece was developed with a microphone pointed into the ear canal that captures these signals generated by tongue movements and speech. The signals are transmitted from the earpiece to an Ultra-Mobile Personal Computer (UMPC) through a wired connection. The UMPC processes the signals and utilizes them for device control. The processing can include command recognition, ambient noise cancellation, acoustic echo cancellation, and speech equalization. Successful control of an iRobot PackBot has been demonstrated with both speech (13 discrete commands) and tongue (5 discrete commands) signals. In preliminary tests, command recognition accuracy was 95% with speech control and 85% with tongue control.

Keywords: Human-robot interaction, robot control, robust speech recognition, tongue movement recognition, ear canal

1. INTRODUCTION

This paper describes the development of hands-free device control technology using speech and tongue movements. The novel development behind this technology is that the speech and tongue movement signals are captured in the ear canal, which can be sealed, so that the captured signals are not impaired by ambient noise. Hence, the technology is well suited for hands-free device control in noisy environments.

There are many applications with inherently noisy environments where hands-free device control is a requirement. One such rapidly growing application is the control of robots, particularly with the rise in the number of Unmanned Ground Vehicles (UGV) employed by the military [1]. Moreover, as the Army's Future Combat Systems (FCS) program—their principal modernization program—is implemented, the number of Small UGVs (SUGV) employed by the Army is going to increase significantly. On the domestic scene, robots are used for homeland defense, including bomb-disposal squads, SWAT teams, and border patrol. Besides this, potential future applications include relief operations in disaster zones, e.g., fire sites, flood-stricken areas, nuclear accident sites, etc, that are hazardous for human rescuers.

Traditionally, these robots are teleoperated through a joystick-based Operator Control Unit (OCU), which tends to be big and bulky, requiring the operator to be confined to a military vehicle or some other stationary position. Moreover, joystick control of the robot requires the operator to employ both of his hands, leaving him defenseless in case of an attack. Hence, there has been a recent push to: 1) Reduce the size and weight of the OCU, making it wearable, allowing the operator to be mobile, and 2) Find a hands-free means of controlling the robot leaving the operator free to use his hands for carrying weapons or other tactical devices. These requirements are assimilated into the concept of the *Warfighter's Associate* developed by researchers at the Space and Naval Warfare Systems Command San Diego (SPAWAR) [2]. This concept calls for a significant effort to introduce novel control techniques to provide the robot with adequate autonomy to function as a *Warfighter's Associate*, as well as novel command input schemes to allow the operator to interact with the robot as he/she would any other fellow soldier.

Currently, the most successful hands-free control devices are speech-command based. However, these devices use external microphones—mounted on a boom to place them close to the user’s mouth, and in field environments their performance suffers because of high decibel ambient noise: typically above 90dBA with impulses of up to 110dBA [3].

The technology described in this paper uses the ear as an output source and can provide excellent command recognition accuracy even in noisy environments. Instead of picking up speech radiating from the mouth, this technology captures speech from the ear canal as shown in Figure 1. Movements of the tongue in the oral cavity and other articulators in the vocal tract are also transmitted through the flesh and bones in the human skull, and eventually detected as pressure changes in the ear canal. These tongue movements create no audible sound, and thus provide an excellent method for stealth control. Tongue movements have been employed previously for device control by researchers at the Naval Postgraduate School, Monterey, in conjunction with Think-A-Move [9]. However, the previous implementation was not able to achieve real-time execution, which was a major drawback. The system described here is real-time.

A patented earpiece, called the TAM earpiece, has been developed with a microphone pointed into the ear canal that captures the signal from speech and tongue movements [4][5][6]. A wired connection transmits the signal between the TAM earpiece and an Ultra-Mobile Personal Computer (UMPC). On the UMPC, the signal is processed using specially developed signal processing algorithms, which can include ambient noise cancellation, acoustic echo cancellation, and speech equalization. The processed signal is then fed to speech and tongue command recognition engines that identify the user command and output it in text format. This system was developed and integrated with the control system of an iRobot PackBot. Successful demonstrations using both speech and tongue signals to control the PackBot have been carried out at several military locations and other public sites. Video demonstrations can be viewed online at:

<http://www.think-a-move.com/video.html>.

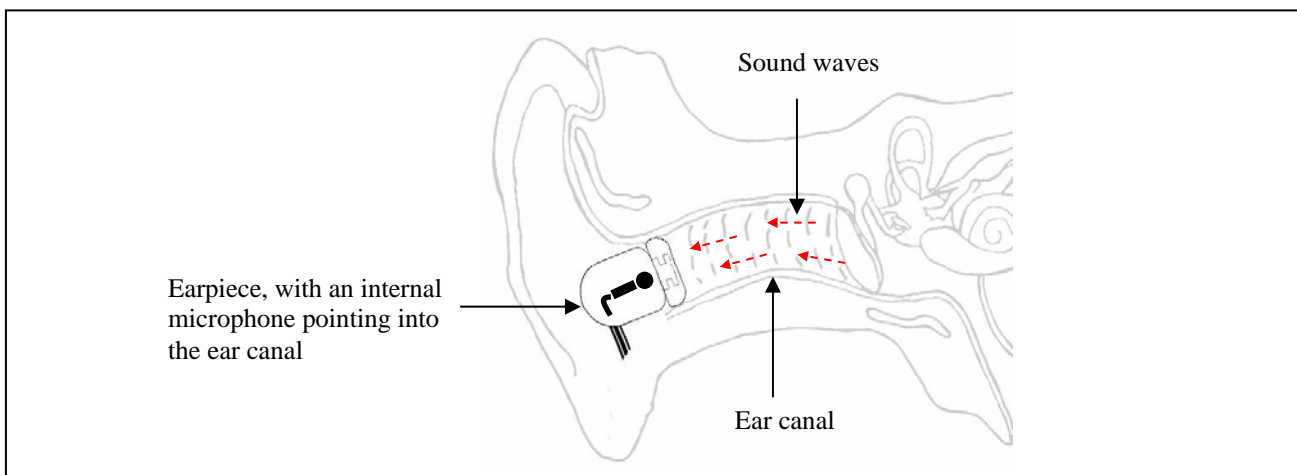


Figure 1. TAM earpiece to pick up speech or tongue motion-induced pressure waves inside the ear canal. These vibrations originate in the voice box or oral cavity and travel through the head to the ear canal.

The rest of the paper is organized as follows. TAM’s in-ear sound capture technology is described in detail in Section 2, along with the characteristics of in-ear speech and tongue movement acoustic signals. Section 3 describes the system configuration for TAM’s command input technology used for device control with focus on robot control. Signal processing algorithms and command recognition engines for speech and tongue commands signals are described in Sections 4 and 5, respectively. The overall command input system was integrated with the iRobot PackBot platform for prototype demonstration. The integrated system is described in Section 6. Section 7 provides results, and conclusions and future work are stated in Section 8.

2. IN-EAR SOUND CAPTURE TECHNOLOGY

As shown in Figure 1, the technology developed by Think-A-Move (TAM) is based on capturing sound waves in the ear canal. These sound waves can be generated by speech or tongue flicks inside the mouth.

2.1 Earpiece for in-ear sound capture

The earpiece used to capture the sound waves, shown in **Error! Reference source not found.**, consists of a molded-plastic shell fitted with a foam ear-tip at its end. The shell houses the microphone, which is installed in such a way that it points into the ear canal when the earpiece is inserted into the ear. The foam ear tips shown here are Commercial-Off-The-Shelf (COTS) products, called Comply Canal Tips, manufactured by Hearing Components [7][8]. Each foam tip is provided with internal threading for easy mounting, and screws onto the end of the earpiece. It is designed to be squeezed, inserted in the outer ear canal, and allowed to expand slowly at body temperature, conforming precisely to the outer ear canal of each user. The foam maintains its excellent seal throughout a variety of movements and activities like talking, laughing, or even chewing. The ear canal is very well shielded from ambient noise allowing the in-ear microphone to pick up speech and tongue signals with a high Signal-to-Noise Ratio (SNR) even in very noisy environments.

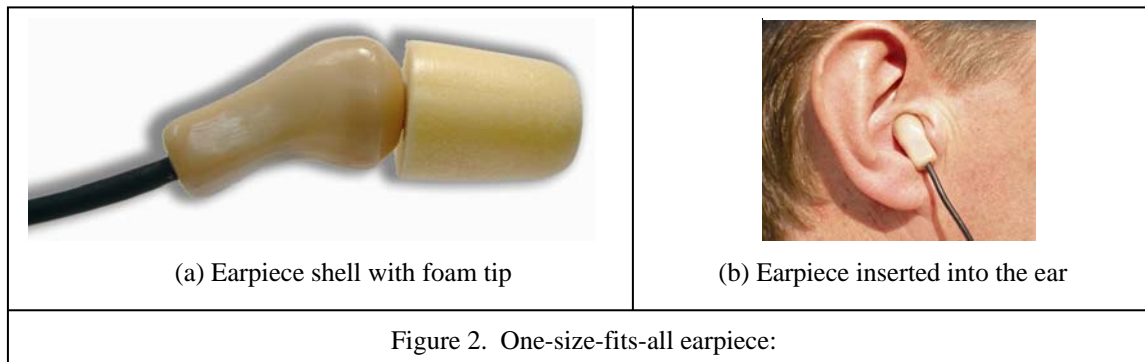


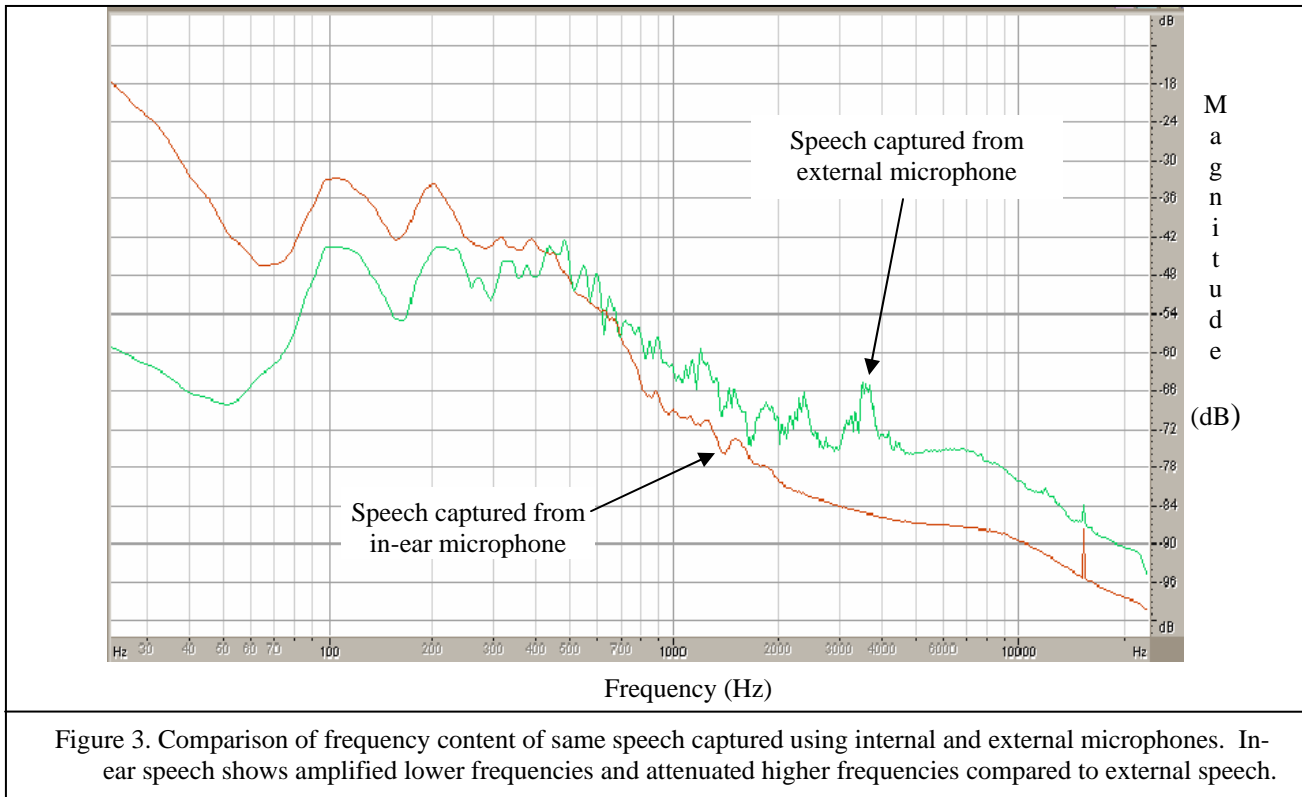
Figure 2. One-size-fits-all earpiece:

The desired characteristics of the microphone used in the earpiece are small size and high sensitivity. Several microphones were compared for size and sensitivity. Finally, a Star-Micronics (MAA series) microphone was selected for its smaller size (3mm diameter), since space is a critical issue with the compact earpiece.

2.2 Acoustic characteristics of in-ear speech and tongue flicks

The acoustic waves produced by speech and tongue flicks reside in disjoint parts of the spectrum. While speech resides in the broad frequency range of 200Hz to above 10,000Hz, the acoustic vibrations produced by tongue flicks are below 200Hz. Speech picked up in the ear canal is the result of bone conduction, and does not have the same characteristics as speech radiating from the mouth. The muscles and tissues covering the skull have a low-pass filtering effect on this signal [10]. This causes the low frequencies of speech to be amplified, intermediate frequencies to be dampened, and higher frequencies to be greatly attenuated. Further, because this speech reaches the ear canal through bone conduction, it does not incorporate the high-pass filtering effect due to the radiation load at the mouth [11][10]. These effects together make the speech picked up by the internal microphone sound muffled. Figure 3 shows the average spectra of speech signals recorded over a 2 minute-interval using an in-ear microphone and an external microphone. It can be seen that the in-ear speech signal is amplified in the lower frequency range and attenuated in the higher frequency range compared to the external speech signal. This muffled speech can be rectified to produce natural-sounding speech using a speech equalization filter as described in Section 4.1.

Sharp movements of the tongue also produce vibrations that can be picked up in the ear canal. These vibrations are at a lower frequency—under 100Hz—compared to speech. To produce strong and consistent signals, an effective method is flicking the tongue lightly against the gum line of the lower jaw. Flicking the tongue along different parts of the lower jaw—as shown in Figure 4—can produce repeatable yet distinct signals. Once a user is trained in generating two or three different signals consistently, combinations of the signals produced in rapid succession can be used for generating a wide variety of control commands. Because generating each tongue signal takes less than 0.4s, combinations can be produced without introducing any significant time delay.



3. DEVICE CONTROL

TAM's speech and tongue movement based command input technology can be used for controlling various devices, particularly those used in loud noise environments. Currently, TAM is focused on the control of robots and their payloads, e.g., manipulator arms, cameras, and other tools and sensors.

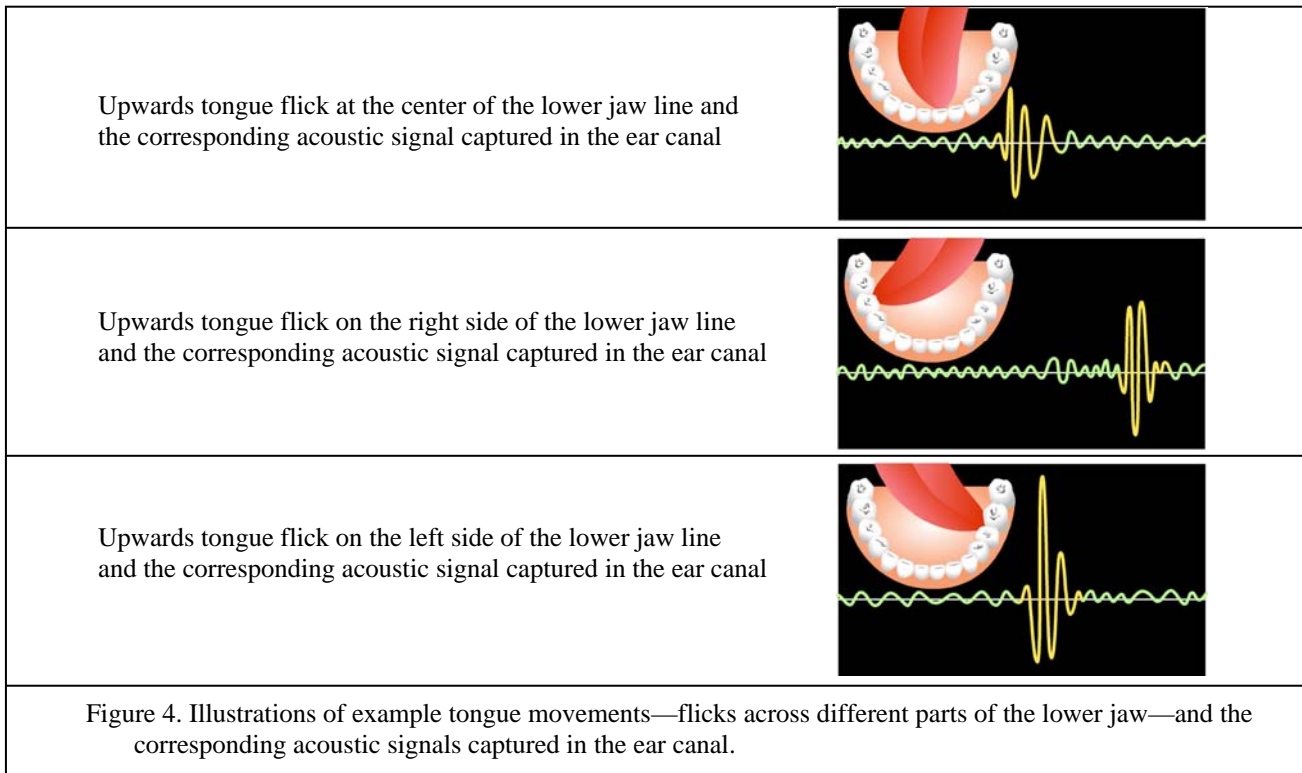
3.1 System configuration

Speech and tongue movement based commands are inherently discrete¹ as opposed to a joystick, which generates commands regularly at very short intervals—a few milliseconds. For continuous control, a discrete command can be latched, so that the command is executed until a new command is received. However, discrete commands are most beneficial when used with an autonomous or semi-autonomous device, i.e., a device that has adequate autonomy to execute a command *intelligently* to achieve a goal. The autonomy can be rudimentary, e.g., a robot that possesses enough autonomy to be able to independently execute a command to move by a fixed distance, or advanced, e.g., a robot that can move to a specified location through path-planning and obstacle avoidance.

An example system configuration for using TAM's speech and tongue movement based command input technology is shown in Figure 5². A wired connection relays the in-ear microphone signal to the Ultra Mobile Personal Computer (UMPC). Because the operator is given the freedom to deliver a speech or tongue movement signal without having to select either command input mode, the captured signal is processed through both speech and tongue modules simultaneously. The processed signals are then sent to the speech and tongue movement command recognition modules, respectively. Depending on whether a speech command or a tongue motion command was issued by the operator, one of the two recognition modules will output a text command. This text command is then passed to the robot control system for execution. Both speech and tongue command input modes can be run simultaneously, because, as stated earlier, speech and tongue motion acoustic signals are in disjoint parts of the spectrum. Therefore, command delivery in one mode does not trigger a false positive in the other mode.

¹ TAM is currently conducting research on humming based continuous control, described briefly in Section 8.

² This configuration was used for the integration with iRobot's PackBot, and will be described here in that context.



TAM also developed a Manager subsystem that handles inter-process communications and displays video and telemetry feedback from the robot in a graphical user interface (GUI). This subsystem was designed to be generic so that it could work with different robot controllers with minimal changes. It also provides a user-friendly system for mapping spoken commands to robot control commands. This mapping is different for different robots based on the specific control commands used by its controller. A library of these robot-specific mappings can be stored on the system, and the appropriate mapping can be loaded depending on the robot.

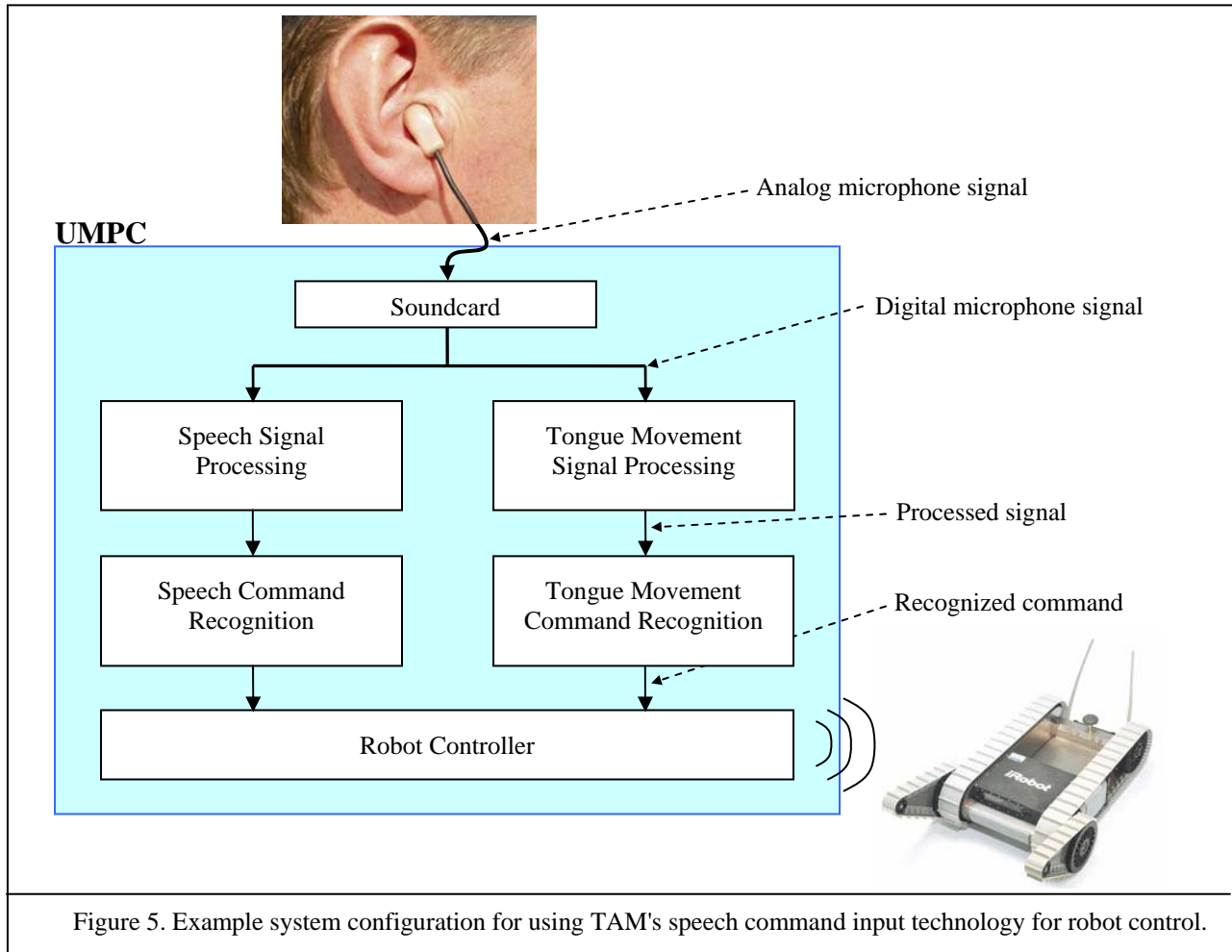
3.2 Speech commands

There are several benefits to using speech for control command input, such as hands-free and heads-up operation and intuitiveness of command delivery. When these benefits are paired with a semi-autonomous robot to produce a *Warfighter's Associate*, the major advantage that emerges is that a minimally trained operator can assume control of the robot and use it effectively. However, very often, the speech commands used are selected by speech recognition experts based on their potential for higher recognition accuracies. This can greatly reduce their intuitiveness, and can require rigorous training for the operator to remember the correct spoken commands. Hence, TAM undertook preliminary work to establish a library of speech commands designed to provide the operator with options in choosing intuitive spoken words for each control command. To make the speech commands truly intuitive, appropriate spoken words were selected with input from end-users. Potential for command recognition accuracy was balanced with intuitiveness. The ultimate goal of this project is to establish standard mission-specific speech command libraries that can be used by other robot control applications using speech commands for similar control tasks.

3.3 Tongue motion commands

An average user can train him/herself to produce two distinct, consistent tongue movement signals within two hours. Usually, a left-flick and a center-flick (see Figure 4) provide the best results. These two signals can be used in combinations to generate several different control commands. Using combinations also reduces the probability of tongue commands being executed without intention. For example, when the user is speaking, some audio can get interpreted as a tongue movement signal, albeit with a very low probability. However, the probability of two or more such false

interpretations being produced within a very short duration is almost zero. Therefore, using combinations, a false control command is almost never produced.



4. SIGNAL PROCESSING

The speech and tongue movement signals captured through the in-ear microphone need to be processed before they are relayed to the respective speech and tongue command recognition engines. The kind of initial processing required is different for both engines.

4.1 Signal processing for speech

The techniques used for processing speech signals include, speech quality enhancement to improve voice quality, ambient noise cancellation for further noise removal, and echo cancellation to remove the echo caused by the speaker in the earpiece.

Speech quality enhancement: As explained above, user speech, captured in the ear canal, sounds muffled. There are two reasons for the muffling effect: 1) Low frequencies are amplified and high frequencies are greatly attenuated, and 2) the bandwidths of the formants increase leading to reduced distinctiveness of the formant peaks [10]. Both of these effects can be seen in the comparison of speech captured through in-ear and external microphones in Figure 3. The muffled speech from bone-conduction can be partially rectified by using a speech morphing filter that will dampen the lower frequencies and amplify the intermediate frequencies in the right proportion. Two different filters were developed and tested. The first was a proprietary filter that produces a high-pass filtering effect, and the second was hand-crafted based on the comparison of the frequency content of speech captured using in-ear and external microphones. Both filters

improved speech quality considerably, making the speech more natural-sounding and improving intelligibility. Ultimately, the high-pass filter was selected for speech enhancement, since it provides a more efficient implementation.

Further improvement in speech quality can be obtained through spectral sharpening. Spectral sharpening [12] refers to reducing the bandwidths of speech formants. This algorithm is still under development.

Another algorithm still under development is called Speech High-band Reclamation (SHR). The goal of SHR is to synthesize speech in the 3kHz-4kHz frequency range based on frequency content below 3kHz. The SHR is based on a proprietary algorithm using the vector quantization method, and should produce better results in this range than simple filtering, which tends to also amplify electronic and background noise in the target range.

Speech quality enhancement allows much improved user command recognition accuracy, and opens up the possibility of using commercially available speech recognition software that are not designed for muffled, in-ear speech.

Ambient noise cancellation: The mainstay of TAM's technology is the capture of audio signals with an in-ear microphone. Part of the challenge is to obtain signals that have most of the ambient noise removed. This is especially critical in a battlefield environment, and is a key part of TAM's research effort. The seal provided by the foam ear tip cancels out much of the ambient noise. However, to achieve greater noise rejection, active noise cancellation was implemented in software using a Spectral Subtraction algorithm [13].

Echo cancellation: A potential feature, targeted by TAM, is the addition of a speaker in the earpiece for audio feedback to the user including audio acknowledgements from the device for command confirmation or audio feed from a microphone mounted on a robot for situational awareness. An earpiece has already been produced that contains both a microphone and a speaker. However, since the internal microphone is positioned next to the speaker, the audio output from the speaker feeds back into the microphone producing an echo. This echo can be very loud because the speaker and microphone are in close proximity (typically, about 2 mm apart), and the ear canal presents a small enclosure. Hence, the sound picked up in the user's internal microphone is first processed to remove this echo. TAM's proprietary acoustic echo cancellation (AEC) module involves various functions including an adaptive filter (ADF), a double-talk detector (DTD), a non-linear processor (NLP), and an echo cancellation controller (ECC). The functions are described below.

- **Adaptive filter (ADF)**

The adaptive filter maintains, and continuously adapts, a transfer function that filters the speaker signal to approximate the speaker output that feeds back into the internal microphone. The transfer function incorporates the speaker response, the acoustic properties of the ear canal, and microphone characteristics. The filter is designed to be adaptive, since the acoustic properties of the ear canal are not constant over time.

The performance of the filter is evaluated based on its Echo Return Loss Enhancement (ERLE). The current version of AEC achieves a filter performance of 20 dB ERLE after 1 s of convergence. The long-term goal is to achieve an ERLE of at least 40 dB. The filter has a relatively high convergence rate, with a convergence time of less than 100 ms. The ADF is a relatively low-complexity frequency domain adaptive filter. While the AEC filter is constantly active, its adaptation should only be activated when there is no local talk, i.e., the proximal user is not speaking, and therefore, only the speaker feedback contributes to the microphone signal. The condition where the proximal user is speaking is called double-talk, and the function used to detect this condition is called a double-talk detector.

- **Double-talk detector (DTD)**

The function of the double-talk detector is to detect whether the proximal user is speaking. If so, then the adaptation of the ADF should be turned off. Double-talk detection is challenging, because, it requires distinguishing between speaker echo and user speech. The signal that is output to the speaker can be accessed by the double-talk detector, and is utilized to perform the detection. The DTD is a low-complexity energy-based method.

- **Non-linear processor (NLP)**

The adaptive filter has its limitations and is unable to cancel out 100% of the echo from the microphone signal. Several non-linear methods are available that reduce the residual echo. The TAM system includes a non-linear function that processes the microphone signal, after the adaptive filter, to eliminate this residual echo.

- **Echo cancellation controller (ECC)**

The ECC is responsible for coordinating the switching on and off of the various AEC modules. For example, when the DTD function detects double talk, the ECC switches off the ADF function to turn off adaptation.

4.2 Signal processing for tongue movement

As mentioned above, the acoustic signals generated in the ear canal through tongue flicks are in the frequency range of 0-100Hz. Hence, the first step in processing the microphone signal for tongue movement detection is low-pass filtering with a cut-off at 100Hz. This removes speech and other noise in the higher frequency range. Some low frequency noise, e.g., the user's pulse, remains in the processed signal, and can be removed with further specialized processing.

5. COMMAND RECOGNITION

As described above, the in-ear microphone signal undergoes specific processing for speech and tongue movement signals, respectively. These processed signals are then relayed to the respective command recognition engines.

5.1 Speech command recognition

A speech command recognition (SCR) system is used to identify the command spoken by the user. The SCR system collects the processed speech signal and outputs a recognized command. Although, several speech recognition systems are available commercially, most of these systems do not allow customized training of their speech models, and hence cannot be used for muffled, in-ear speech. The systems that do allow customized training are usually computationally too intensive to be used for command and control applications on low-cost hardware. TAM has experimented with several commercial-off-the-shelf (COTS) speech recognition systems, including Nuance Dragon NaturallySpeaking, IBM ViaVoice, and Microsoft Speech Recognition. These experiments have yielded encouraging results, providing more than 95% recognition accuracy with in-ear speech. However, in order to achieve better accuracy and maintain high accuracy numbers while engaged in physical activity as well as in the presence of ambient noise, TAM has developed a proprietary SCR system.

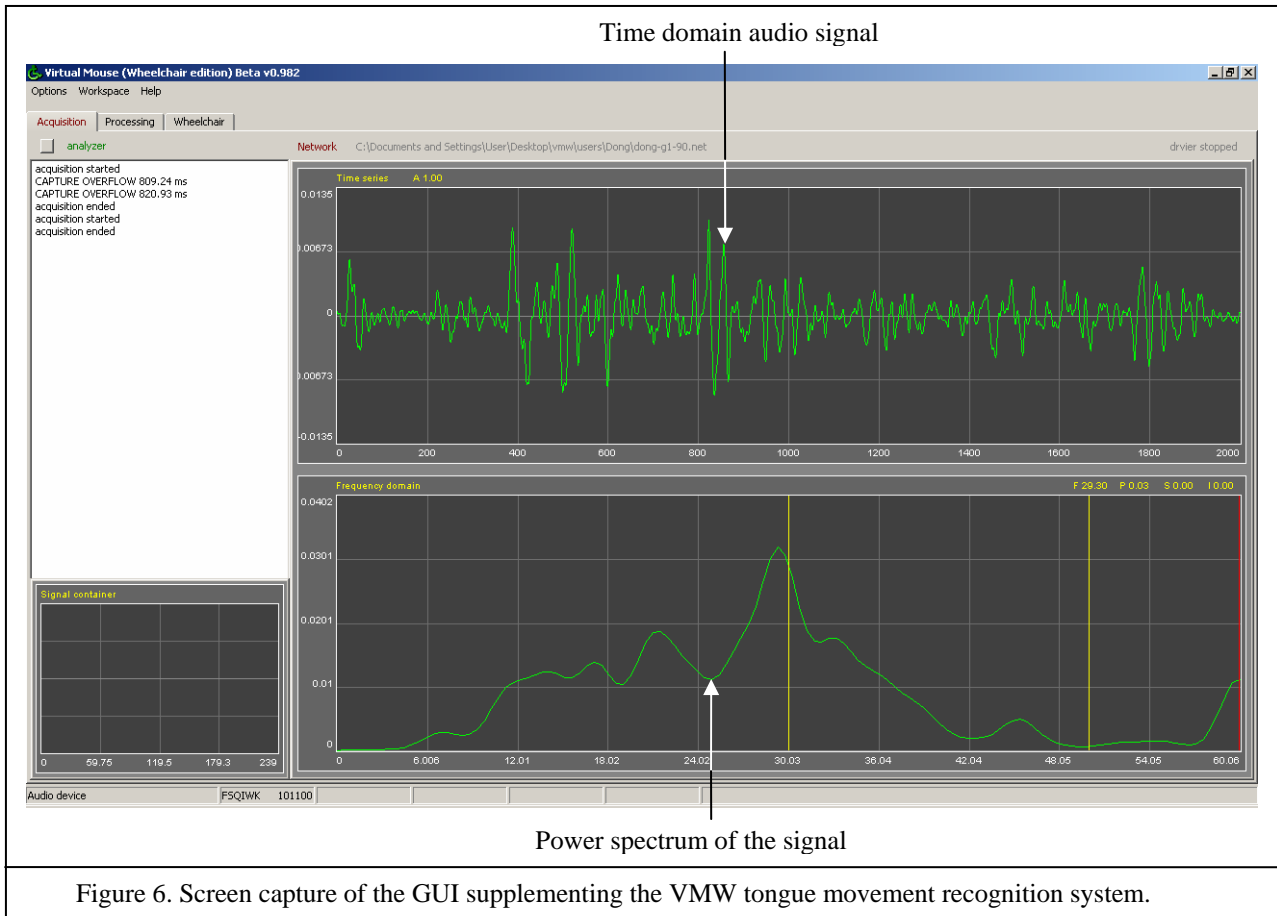
TAM's SCR system has been developed for use on a Linux platform, which is the preferred platform for the military community and the future platform for iRobot's Dismounted Operator Control Unit (DOCU). The SCR engine is a Hidden Markov Model (HMM)-based system, and incorporates speech parameterization amenable to in-ear speech characteristics. In-ear speech above 4kHz is highly attenuated even after the speech enhancement procedures discussed above are applied. Therefore, speech features—mel-frequency cepstral coefficients—are computed such that only the spectrum below 4kHz is targeted. TAM has trained specialized models tuned for recognition of muffled, in-ear speech, and the recognition accuracy target has been set to more than 95% even in extreme conditions, which include battlefield conditions where the operator is involved in intense physical activity like hiking and running, with loud noise in the background.

Testing and Results

The following observations were noted in comparing the COTS speech recognition software: Dragon NaturallySpeaking (DNS) and Microsoft SDK. 1) DNS does a better job with "word-spotting" compared to Microsoft SDK. Word-spotting refers to recognizing commands amidst conversation, and rejecting non-command conversational words. 2) However, in the presence of ambient noise, DNS has lower recognition accuracy compared to Microsoft SDK and requires significantly more processing time than Microsoft SDK.

5.2 Tongue movement based command recognition

Proprietary software, called VMW, was developed by TAM for the processing and recognition of tongue movement signals in real-time for device control. VMW is also equipped with a well-developed graphical user interface (GUI) that can be used to not only set and change parameter values, but also to visualize the captured audio signal with signal-portions identified as tongue commands highlighted. A screen capture of the GUI is shown in Figure 6.



The implementation of tongue movement signal recognition by the VMW software can be divided into two parts: training and execution. The training part consists of extracting *generalized shapes* from a database of processed signals and then training a neural network to recognize these generalized shapes. Once the neural network has been trained, in the execution part, the algorithm uses the neural network in real-time to detect and classify tongue motion signals. These two parts are described below in more detail.

a. Training: Signal Capture, Processing, and Neural Training

The training part is executed in the following sequence:

1. The first step in building a training database is capturing training signals. On average 100 signals (also termed *vectors*) are recorded for each *action*. An action is a particular type of tongue movement, as shown in Figure 4.
2. The signals are then processed to extract a generalized signal shape for each action. Processing involves signal segmentation based on energy, followed by signal alignment using maximum cross-correlation. The aligned signals are then averaged to find the generalized shape.
3. Corresponding to each action, a number of *stable points* are identified on its generalized shape. The stable points are the points on the generalized shape whose variance, with respect to all the vectors corresponding to that action, is under a set threshold. Besides the generalized shape, the processing also yields other signal detection parameters such as signal energy, signal duration, and frequency range of the signal. These parameters can be used in the real-time signal detection routine for a quick check to eliminate signals that do not satisfy the requirements.
4. Once the stable points for all the actions have been identified, they are employed to train the neural network. The network can be constructed by the user by specifying the number of cluster neurons to be used. The number of inputs is automatically set equal to the total number of stable points, and the number of outputs is set

equal to the number of actions. The trained neural network can then be employed for signal recognition in real-time.

b. Execution: Real-time Signal Identification and Recognition

During the execution phase of the VMW software, the audio signal corresponding to a tongue movement captured in the in-ear microphone is processed in real-time to detect the presence of a known signal and then to identify and classify the signal. The signal detection parameters found in the training phase are employed to select promising portions of the incoming signal for further processing. If a portion of the signal satisfies signal energy, duration, and frequency range requirements, then it is processed through the neural network to classify the action that it corresponds to. This processing is carried out by finding the signal values at the pre-determined stable points and feeding the values into the neural network. The output of the neural network identifies the action that the signal corresponds to. Once the action has been identified, the control command corresponding to that action is output.

A preliminary version of the VMW software was designed such that it required considerable expertise on the part of the user to carry out processing of the database and training the neural network. This was contrary to our goal of developing a system that can be used by a minimally trained operator without extensive knowledge of the system. Hence, one of the main improvements achieved in the current version is the automation of these procedures, so that, after the signal database has been recorded, processing the signals and training the neural network can be carried out by pressing a single button. Thus, the system set-up has been greatly simplified. It is important to mention here that the system still allows an expert to step through the database processing and neural network training procedures, and to customize parameters, if required.

Testing and Results

Preliminary tests were carried out on the VMW software to evaluate ease of training and recognition accuracy for tongue movement based control. A subject who had never undergone tongue motion training or used the software prior to the test was recruited. The subject first undertook a practice phase where he repeatedly made two distinct tongue motions: left flick and right flick. While making the tongue motions, the subject was watching the captured signals in the GUI. The visual feedback helped the subject hone his skills in producing consistent signals. This practice routine was executed in two sessions of two hours each. After practicing, the subject recorded a database of signals: 250 vectors per action (a total of 500 signals) in about an hour. Once the neural network was trained on this database, the tongue command recognition system was tested. The subject produced 200 signals of the two tongue motions in random order. Only 5 of the 200 signals were identified incorrectly, and there were no false positives, providing a recognition accuracy of 97.5%. More extensive testing of the VMW system has been planned.

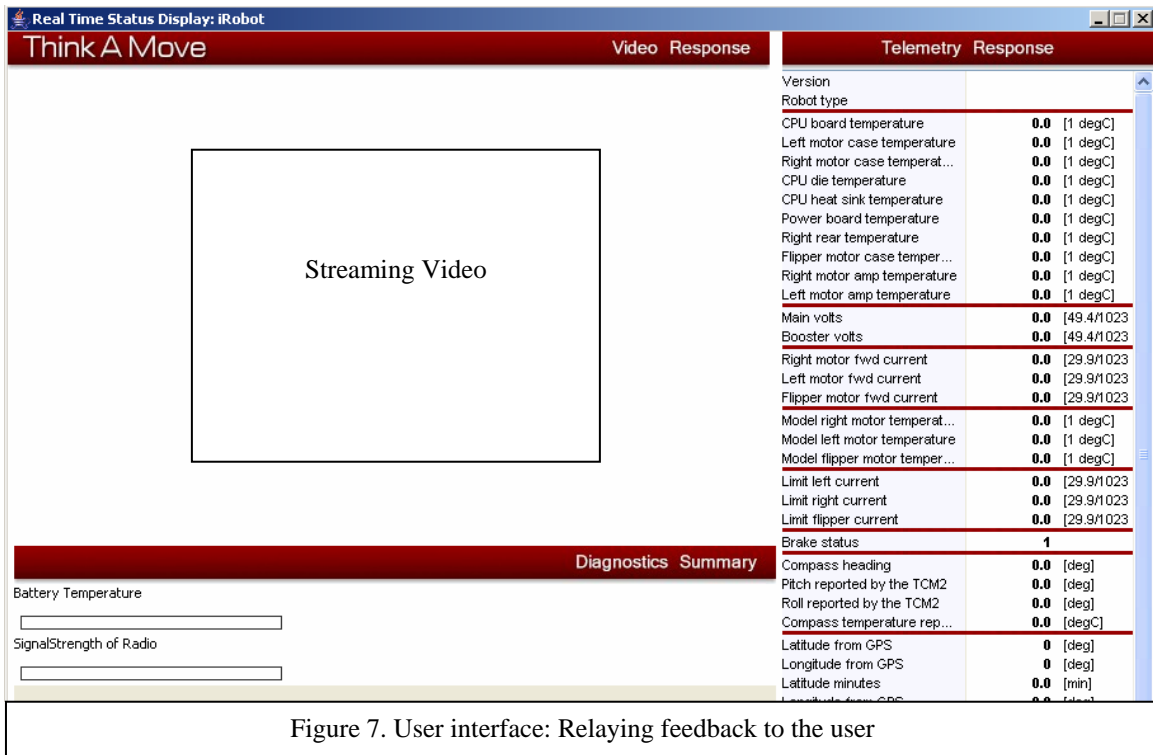
6. SYSTEM INTEGRATION WITH ROBOT CONTROLLER

TAM's speech and tongue movement based command input system, described above, was integrated with iRobot's PackBot platform. The TAM system, running on a Sony Vaio UX280P MicroPC, accepted speech and tongue command signals from the earpiece placed in the user's ear canal, interpreted and mapped the speech commands onto control commands, and communicated the control commands to the robot. Commercially available speech recognition software, Dragon NaturallySpeaking, distributed by Nuance, was used as the speech command recognition engine. This software can be run under "Command Mode," where a user can add custom speech commands and program the action to be taken on recognition of the commands. The action required was communicating a text message—uniquely corresponding to the command recognized—to the Manager subsystem. This *text command* was then passed to the PackBot control system. A comprehensive user interface to display live robot information was also developed for the demonstration. The interface displays streaming video and telemetry data, and a diagnostics summary. Figure 7 shows a screenshot of the interface. The interface was designed based on ergonomic considerations and input/feedback from prospective users.

The control commands were designed to command distance/angle motions to the robot and not velocity. The total number of motion commands used was 13 for speech and 5 for tongue. The speech commands used are listed in Table 1. For speech, separate motion commands for large motions and small corrections were incorporated. This required the controller to be capable of keeping account of robot motion, so that the motion could be stopped upon successful completion. This was accomplished by using the telemetry data reported back to the OCU. While this data does not provide a highly accurate prediction of robot motion, it was adequate for the purposes of a prototype demonstration. The five tongue commands were mapped to "Forward big," "Reverse big," "Left big," "Right big," and "Stop."

Table 1. Speech commands user for controlling the iRobot PackBot

Speech command	Action
Forward big	Move forward by 10m
Forward small	Move forward by 3m
Forward left	Move forward by 3m and turn left by 15°
Forward right	Move forward by 3m and turn right by 15°
Reverse big	Move reverse by 10m
Reverse small	Move reverse by 3m
Reverse left	Move reverse by 3m and turn left by 15°
Reverse right	Move reverse by 3m and turn right by 15°
Left big	Turn left by 30°
Left small	Turn left by 15°
Right big	Turn right by 30°
Right small	Turn right by 15°
Stop	Stop all motion



7. RESULTS

The system described above, TAM's speech and tongue command input technology integrated with the iRobot PackBot, was demonstrated at the 25th Army Science Conference in Orlando, 27-30 November, 2006. The demonstration was successful, as the system achieved nearly 100% command recognition accuracy in spite of the conversational and other noises on the exhibit floor. The hands-free command input technology was also showcased by iRobot, among other leading technologies, to General Cartwright at the iRobot headquarters in Burlington, MA. In this demonstration, iRobot integrated the command input system with the Future Combat Systems (FCS) Small Unmanned Ground Vehicle (SUGV). Unlike the PackBot, the SUGV requires JAUS compliance. Hence, the TAM system was provided with rudimentary JAUS compliance to be able to control the SUGV. The demo was very successful as the TAM system was operated by a freshly-trained user, and recognized all the commands delivered.

8. CONCLUSIONS AND FUTURE WORK

This paper described the development of hands-free device control technology using speech and tongue movements. The novel development behind this technology was the capture of speech and tongue motion signals in the ear canal, which provides a high Signal-to-Noise ratio even in very noisy environments. This gives the system an advantage over speech command input systems using external microphones whose performance degrades rapidly in noisy environments. Moreover, the tongue movements create no audible sound, and thus provide an excellent method for stealth control in situations when speaking is not an option.

Currently, the TAM system is at a Technology Readiness Level of TRL-6. TAM has carried out prototype demonstrations in relevant environments. In the immediate future, TAM has planned extensive testing of its command input system for robust operation with consistent command recognition accuracy even in extreme conditions. This testing will be paired with further improvements in the core technology to achieve TRL-8.

TAM is working on an enhanced version of its technology that includes a speaker in the earpiece. The addition of the speaker allows the operator to receive discreet audio feedback from the robot. This feedback can be from the microphone on the robot allowing the operator to listen to the robot's surroundings, or from the robot's controller providing the operator with a variety of vital information, e.g., urgent communication or emergency alert generated by a sensory system mounted on the robot, command confirmation, and task status update. An earpiece has already been produced that contains both a microphone and a speaker. TAM is in the process of developing an echo cancellation function, which will then allow the feature to be integrated into TAM's command input system. With the addition of the speaker, the system can also be used to provide the operator with a communications system. This solution, converging device control with communications, provides great benefit to the dismounted operator by reducing the number of different gadgets that the operator has to carry and handle.

The major drawback of speech and tongue motion based input systems is that they are limited to producing discrete commands, making it difficult to adapt them to joystick-type directional robot control. Hence, TAM is pursuing a humming-based solution to emulate a joystick, allowing the operator to command a sustained continuous action.

REFERENCES

- [1] Office of the Secretary of Defense, "Unmanned Systems Roadmap 2007-2032," 2007.
- [2] Everett, H., Pacis, E., Kogut, G., Farrington, N., Khurana, S., "Towards a Warfighter's Associate: Eliminating the Operator Control Unit," SPIE Proceedings 5609: Mobile Robots XVII, Philadelphia, PA, 2004.
- [3] Henry, P., Mermagen, T., Letowski, T., "An Evaluation of a Spoken Language Interface," Technical Report ARL-TR-3477, Army Research Laboratory, MD, 2005.
- [4] Nemirovski, G., "System and method for detecting an action of the head and generating an output in response thereto," U.S. Patent No. 6,503,197, Jan. 7, 2003.
- [5] Nemirovski, G., "Sensor pair for detecting changes within a human ear and producing a signal corresponding to thought, movement, biological function and/or speech," U.S. Patent No. 6,647,368, Nov. 11, 2003.
- [6] Nemirovski, G., "Ear microphone apparatus and method," U.S. Patent No. 6,671,379, Dec. 30, 2003.
- [7] Ahlberg, C., Chamberlin, D., Bushong, J., Oliveira, R., Kolpe, V., "Hearing aid ear piece having disposable compressible polymeric foam sleeve," U.S. Patent No. 4,880,076, Dec. 5, 1986.
- [8] Oliveira, R., Chamberlin, D., Babcock, M., "Ear piece having disposable compressible polymeric foam sleeve," U.S. Patent No. 5,002,151, Oct. 4, 1989.
- [9] Vaidyanathan, R., Chung, B., Gupta, L., Kook, H., Kota, S., West, J., "Tongue-movement communication and control concept for hands-free human-machine interfaces," IEEE Transactions on Systems, Man, and Cybernetics, Part A, 37(4), 533—546, July 2007.
- [10] Westerlund, N., Dahl, M., and Claesson, I., "In-ear microphone techniques for severe noise situations," Blekinge Institute of Technology, Research Report No. 2005:12.
- [11] Rabiner L. and Schafer R., [Digital Processing of Speech Signals], Prentice Hall, 1978.
- [12] Schaub, A. and Straub, P., "Spectral sharpening for speech enhancement/noise reduction," Proc. of the International Conference on Acoustics, Speech and Signal Processing, pp. 993-996, 1991.
- [13] Berouti, M., Schwartz, R., Makhoul, J., "Enhancement of speech corrupted by acoustic noise," Proc. of ICASSP, 208-211, 1979.