Ravi Vaidyanathan*

University of Southampton, Southampton, UK
Naval Postgraduate School, Monterey, CA, USA
Case Western Reserve University, OH, USA
rvaidyan@nps.edu

Monique P. Fargues
R. Serdar Kurcan

Naval Postgraduate School, Monterey, CA, USA

Lalit Gupta
Srinivas Kota

Southern Illinois University, Carbondale, IL, USA

Roger D. Quinn

Case Western Reserve University, OH, USA

Dong Lin

Think-A-Move, Ltd, Beachwood, OH, USA

A Dual Mode Human-Robot Teleoperation Interface Based on Airflow in the Aural Cavity

Abstract

Robot teleoperation systems have been limited in their utility due to the need for operator motion, lack of portability and limitation to singular input modalities. In this article, the design and construction of a dual-mode human–machine interface system for robot teleoperation addressing all these issues is presented. The interface is capable of directing robotic devices in response to tongue movement and/or speech without insertion of any device in the vicinity of the oral cavity. The interface is centered on the unique properties of the human ear as an acoustic output device. Specifically, we present: (1) an analysis of the sensitivity of human ear canals as acoustic output device; (2) the design of a new sensor for monitoring airflow in the aural canal; (3) pattern recognition procedures for recognition of both speech and tongue movement by monitoring aural flow across several human test subjects; and (4) a conceptual design and simulation of the machine interface system. We believe this work will lay the foundation for a new generation of human machine interface systems for all manner of robotic applications.

The International Journal of Robotics Research
Vol. 00, No. 00, Xxxxxxxx 2007, pp. 000–000
DOI: 10.1177/0278364907082612

Figures 1, 3, 13–15 appear in color online: http://ijr.sagepub.com

KEY WORDS—robot teleoperation, human–machine interface, decision fusion, multi-modal control, physiological signal recognition

1. Introduction

Sensors, actuators and onboard intelligence have not yet reached a level where robotic vehicles may function with complete autonomy. Human operation and command is still necessary for utility in unstructured environments, tasks where individual action is undesirable or infeasible and in situations where a robot must assist and/or interface with people. The fundamental goal of a teleoperation system is to facilitate such control through uniting a human operator as the supervisor with a robot as the task executor. Research in the field has highlighted applications ranging from space robotics, terrestrial and undersea exploration, handling of hazardous materials, surgical robots, dismantling of explosives, use in nuclear facilities, operating in inaccessible sites in rescue, surveillance, home assistance, industrial mining and (military) scouting (Melchiorri and Eusebi 1996). Given its breadth, robot teleoperation has benefited from the confluence of research efforts in an array of fields including engineering, psychology, education and medicine (Goldberg 2000).

* Corresponding author

1.1. Human–machine interfaces for telerobotics

Almost any robotic teleoperation system will consist of three fundamental components (Cui et. al. 2003): (1) the remote robot; (2) the communication link; and (3) the human–robot interface. While all three have been extensively researched[1], almost all existing systems rely on joysticks (operated by the hands or more rarely with the head, chin and/or neck), computer mouse devices or other mechanisms based on external operator (physical) movement for the human–robot interface (Cui et al. 2003; Kuan and Kuu 2003).

A number of researchers have recognized the importance of creating smoother and more natural human interfaces for robot teleoperation. Concepts for body-worn devices such as exoskeleton mechanical devices (Chang et al. 1999), instrumented gloves (Tezuka et al. 1994; Harada et al. 2000), inertial (Yun and Bachmann 2006) or electromagnetic (Bashashati et al. 2006) motion tracking sensors on the arms, head or legs, electroencephalographic (EEG) brain activity sensors (Millan et al. 2004; Tanaka et al. 2005), and electromyographic (EMG) muscular activity sensors (Ferguson and Dunlop 2002; Fukuda et al. 2003) have all been explored. Additionally, camera-based (vision) and laser interfaces (Hu et al. 2003; Kofman et al. 2005) have been developed to recognize hand gestures, track arm motions or measure eye movement (Chen and Newman 2004) for assessment of operator intent and generation of robotic control signals.

As an alternative to external movement, a great deal of work has also highlighted the potential of the human oral cavity as a source for control input (primarily in rehabilitative applications). Contemporary examples include inserting a track-ball, joystick, plastic palate or 'sip-and-puff' straw into the mouth of an individual with the tongue or lips providing control input. Voice recognition software, arguably a subset of these oral-cavity interfaces, also offers a promising technique for human-robot interface. Although external noise can disrupt and mask operator commands, several research groups have successfully implemented voice recognition systems as an interface for robot teleoperation (Marin et al. 2002; Liu et al. 2005).

1.2. Limitations of current systems

While these developments offer a wealth of future promise for robotic control, their utility at present has been restricted largely to laboratory demonstrations in controlled environments. Major challenges which have limited the realization of natural human–machine interface systems functioning in real-world situations include:

- Portability and robustness: Most existing interfaces consist of components that are not robust enough, too awkward or too bulky for human use outside of controlled environments. Many sensors (e.g. cameras oriented for motion capture) can be used only in indoor spaces which have been prepared *a priori* and do not have external interference (e.g. darkness or obstacles occluding vision sensors, background noise disrupting voice recognition, etc.). Furthermore, systems that require extensive supporting and/or processing equipment (e.g. instrumented gloves, physiological sensing, etc.) can be difficult to transport and use in uncontrolled environments (outdoors, unfamiliar locations, etc.), and may be unwieldy for a human operator.
- Motion constraints: Most existing interfaces may be classified as mechanical input devices; i.e. the user physically moves some component in order to generate a control input signal. In assessing such systems, Kofman et al. (2005) stated: '…contacting devices may hinder dexterous human motion due to the presence of the devices, sensors or attached cables', and further observed '...mechanical robot-arm replicas, dials, joysticks, computer mouse, and computer graphical interfaces require operator motions that may be unnatural and must be learned'. Constant bodily movements are also not feasible in many situations (e.g. when the hands are occupied; Richardson and Rodgers 2001; Karlsen 2004[2]), and clearly exclude individuals with extremity impairments who need to control assistive devices.
- Input modality: Interfaces typically allow for only a single mode of interaction. This precludes the possibility of interacting with the robotic device on more than one level. Consider, for example, low-level versus high-level input to maneuver a robot through a dense environment. One could directly drive the robot, give higher-level directions ('move forward *X* meters, turn right,...'), or designate waypoints on a map (Wang and Liu 2004). Transition between different situations, different robots and different environments is not feasible unless the interface can accommodate this changing level of interaction. A small body of research has attempted to address this issue (Lim et al. 2003; Raneda et al. 2003; Wang and Liu 2004; Marin et al. 2005; Urban and Bajcsy 2005; Galindo et al. 2006) by combining voice recognition with other modalities (Marin et al. 2005; Urban and Bajcsy 2005). However, these solutions demand the integration of several disparate input devices

1. Communications in particular has been the focus of a proliferation of recent work due to internet and telecommunications advances (Siegwart and Goldberg 2000).

2. In a recent document summarizing their needs (Karlsen 2004), the US Army Tank and Automotive Command (TACOM) stated 'Teleoperation is currently the most reliable method for operating an unmanned ground vehicle. However, there are a number of disadvantages to standard methods of teleoperation, including the requirement for the soldier to give up his weapon in exchange for a control device'.

(e.g. cameras, microphones and/or joysticks; Wang and Liu 2004), lasers and pressure sensors (Lim et al. 2003) which incur further motion constraints, portability and robustness issues.

Oral interface mechanisms offer some potential for addressing these challenges. However, they can be intrusive, may irritate the mouth, impair verbal communication, present hygiene issues and are limited in signal generation capacity. External noise also limits voice recognition systems to controlled environments. Although microphone arrays designed to monitor and filter out environmental noise offer some potential to address this issue, these systems still suffer from the fact that the speech capture microphone has no direct shielding since it must be placed near or in front of a user's mouth. In-ear microphones (Westerlund et al. 2001) are also available for collecting speech data in high noise environments, yet this work has seen no application in robotics and relies on a custom-made device that may not be desirable when dealing with multiple users or field environment constraints. Finally all these devices still offer only one input modality. In summary, we are not aware of any human–machine interface system available for robot teleoperation that is not hindered by at least one of the aforementioned limitations.

1.3. Aim of research

The goal of our ongoing research is to develop a human–robot interface which can overcome these challenges for seamless teleoperation of robotic platforms in any real-world environment. In past work (Vaidyanathan et al. 2004; 2006; 2007) we reported the development of a non-intrusive tongue-movement machine interface. In particular, we demonstrated that tongue movements within the human oral cavity create unique, subtle pressure signals in the ear (referred to as tongue-movement-ear-pressure or TMEP signals) that can be recognized with a range of pattern recognition strategies (Vaidyanathan et al. 2007).

In this work, we introduce a dual mode human-robot teleoperation interface based on monitoring airflow in the aural cavity. The interface is easily portable and requires no physical movement from the human operator. It is capable of detecting both tongue movement and speech for multiple levels of control input using nothing more than a microphone-earpiece housing. We present research results demonstrating the veracity and operability of the system including: (1) an analysis of the sensitivity of the human ear canals as acoustic output device; (2) the design of a new sensor for monitoring airflow in the aural canal; (3) pattern classification algorithms and implementations for recognition of both speech and tongue movement by monitoring aural flow across several human test subjects; and (4) a conceptual design and simulation results on a candidate robotic platform of each mode of the human-robot interface.

2. Modeling of Air Flow Within the Human Ear Canal

We have developed a model of the ear canal to establish the veracity of using the ear as an output device and to dictate proper sensor design. The acoustic sensitivity of the human ear acting as an output platform has been studied from two aspects: a static aspect and dynamic aspect. The former studies the sensitivity created by the change of ear canal volume, while the latter studies the sensitivity created by the airflow velocity in the ear canal.

2.1. Ear canal pressure change due to volume variation

We have modeled the ear canal as a 2 cm^3 volume cylinder as specified by the American National Standards Institute (ANSI) S3.7 and IEC 711 standard. When the tongue or cheek moves, forces will be created around the walls of the ear canal, which in turn changes the volume of the ear. The whole process is approximated as adiabatic. Therefore, according to thermodynamics theory, we have

$$PV^{\gamma} = \mathrm{C}, \tag{1}$$

where P is the air pressure in the ear canal, the summation of atmospheric pressure P_0 and induced acoustic pressure p; V is the volume of the ear canal, the summation of the static volume and the variation due to tongue movements; $\gamma = 1.4$ is the specific heat ratio of the air; and C is a constant. The relationship between the induced acoustic pressure and the variation of ear canal volume can be obtained from Equation 1 as:

$$p = -\gamma P \frac{\delta V}{V} \approx -\gamma P_0 \frac{\delta V}{V_0}, \tag{2}$$

where δV is the volume variation of the canal and V_0 is the static ear canal volume. Based on Equation 2, a relative change in canal volume as small 10^{-6} introduces an acoustic pressure of 77 dB (from a reference of 20 μPa). Thus, with a volume variation of only 0.002 mm^3 in our model, a significant acoustic pressure is created inside the ear canal, justifying the premise of its sensitivity as acoustic output device.

2.2. Ear canal acoustic pressure due to volume speed of airflow

In Section 2.1, the pressure variation within the static ear canal was evaluated. In actuality, the volume change within the ear canal varies dynamically according to the airflow velocity. In the following analysis, the airflow velocity is treated as a harmonic signal with amplitude $v/2$ and angular frequency ω. At low frequency, the equivalent circuit for the 2 cm^3 canal is

Fig. 1. Sensor/earpiece housing for signal capture.

treated as an acoustic capacitor. Therefore, the absolute induced acoustic pressure inside the ear canal due to airflow velocity is:

$$|p| = \frac{v\gamma P_0}{\omega V}, \tag{3}$$

where v is the velocity, V = 2 cm^3 and $\dot{u}$ is the angular frequency of the airflow when treated as harmonic vibration. According to Equation 3, an airflow velocity v of only 1 mm^3 s^{-1} at 10 Hz results in an induced acoustic pressure of p = 1.13 Pa, which corresponds to 95.0 dB (from a reference of 20 μPa.). Both of these models demonstrate the sensitivity of the ear canal as an acoustic output device for human machine interface.

3. Sensor (Earpiece) Design for Aural Flow Monitoring

Our research team has designed and, through iterative prototypes, significantly improved the performance of the earpiece sensor housing to detect pressure fluctuations in the ear canal. In previous experiments, the pressure sensor and circuitry were housed in a custom-designed and molded earplug housing. We have completed the design and fabrication of a new physical housing suitable for use with any subject with no custom-made components. The result is the earpiece shown in Figure 1.

The earpiece system is separated into two components. The portion of the device that is actually inserted in the ear to pick up pressure fluctuations is a soft foam shell with a tube that connects the ear canal to the sensor and electronics housing. Studies conducted for sensor placement (based on the acoustic air flow models) dictated the shape and depth of insertion of the microphone-earpiece housing. The tube capturing airflow input to the microphone resides on the interior portion of the housing within the ear canal. The tube capturing airflow in this device resides within the ear at around 10 mm from the opening of the ear canal. The sensor and electronics housing are formed into a small molded shell, which is then fitted over the back of the ear. The system has been demonstrated to provide comparable performance and comfort to the first generation system and is easily adaptable to a wide range of users. Furthermore, due to the compliant soft foam insertion, the new earpiece enjoys greater benefits with respect to shielding pressure signals from environmental noise.

4. Measure of Aural Flow Resulting from Initiating Actions

4.1. Speech

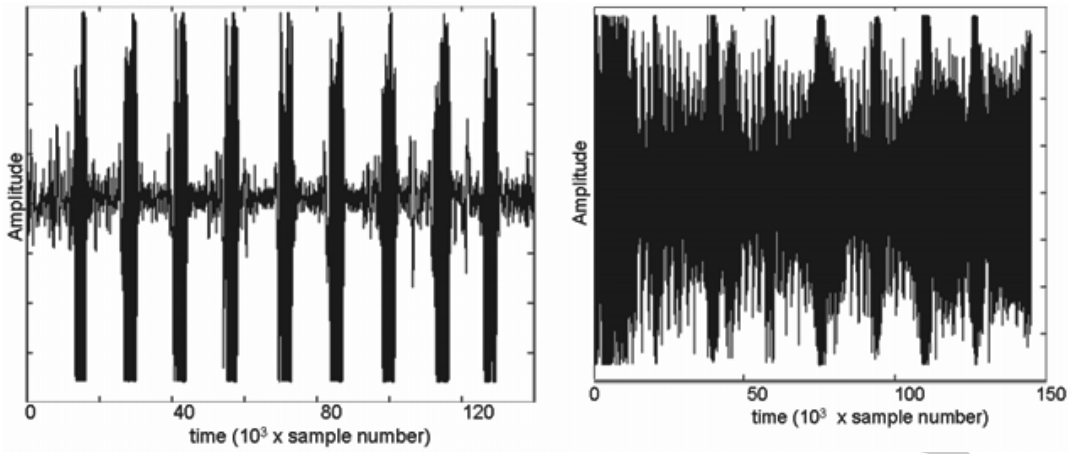
Figure 2 shows speech data collected by the sensor earpiece housing in a high noise environment. Figure 2a shows data collected with the sensor in Figure 1 located in the ear, while Figure 2b shows data collected with the same sensor located in front of the mouth. Both experiments were performed with the same word and same background noise for comparison to traditional speech recognition.

The two plots clearly illustrate the external shielding capability of the device when inserted in the ear, which highlights the superiority of the device in noisy environments compared to other speech recognition systems. While other research groups have investigated speech capture in the aural cavity (Westerlund et al. 2001; 2002), this is the only work we are aware of that has made use of a non-customized sensor.

4.2. Tongue movement

Tongue movement signals are more difficult to generate than speech, and normally take several hours of practice for new users. Based upon extensive feedback from test subjects, we have defined four basic tongue movements for robotic interface, which nearly all operators should be capable of generating (Vaidyanathan et al. 2007). These are: touching the tongue to the top/front center of the roof of the mouth, and flicking it gently forward (forward movement), touching the tongue to the bottom/front center of the mouth, the front/right side of the mouth, or the front/left side of the mouth and flicking it gently up from any of these positions (backward, right and left movements). Backwards, right and left tongue movements are illustrated graphically in Figure 3. While a broad range of actions are possible and may be tailored to individual user preference, most subjects have been comfortable with these movements; we therefore refer to this set of 4 movements as the standard interface.

Figure 4 shows a sample of raw data gathered from a microphone embedded in the housing described above and inserted in the ear of a subject as shown in Figure 1. The subject was asked to make a right movement as previously described. Not only does the trace offer a very clear indication of the onset of



A B

Fig. 2. Aural speech data displaying nine trials of the word 'one' within a high noise environment. (a) Sensor located in the ear and (b) sensor located in front of the mouth.

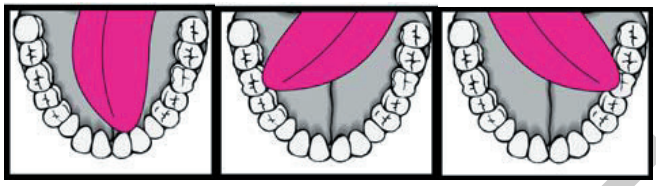


Fig. 3. Three tongue initiating movements.

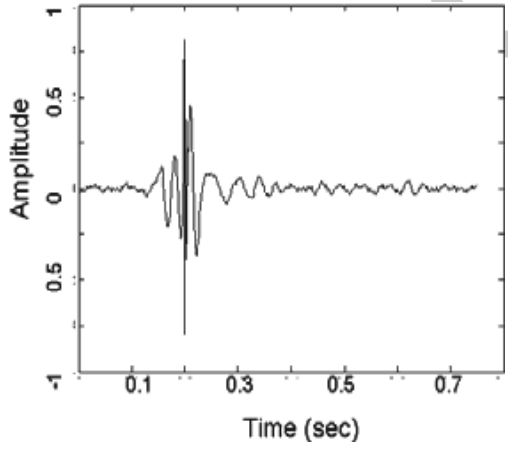


Fig. 4. Tongue movement data.

the motion, but its termination is also visibly evident, as nearly all residual traces of the motion are gone after only 0.2 seconds. The short time frame in which the activities occur allows for quite rapid control input.

5. Signal Recognition Procedure

5.1. Speech

5.1.1. Characteristics of in-ear speech

The microphone-earpiece housing was used to collect speech data from several subjects (Newton 2005). The speech database was collected in an office environment and consisted of twenty adult subjects (16 males and 4 females). Each subject repeated a set of seven short isolated words for robotic command (down, up, right, left, pan, move, kill; based on feedback from TACOM for a soldier to control a mobile robot in the field) fifty times, resulting in 7000 trials. Data were collected with an 8 kHz sampling frequency to ensure spectral information of the quality of a telephone. Short words were selected for the study as they were thought to be a better fit for the robotic control. Words selected contained both voiced (such as /i/ or /ay/) and unvoiced (such as /t/ or /f/) sounds, high (such as vowels) and low (/t/ or /f/) energy sounds which made the speech endpoint detection process challenging (Qiang and Youwei 1998).

Although speech data normally contains frequency content above 2 kHz, we have observed that the ear canal environment acts as a low-pass filter and significantly dampens speech information above 2000–2500 kHz. Figure 5 displays spectrograms obtained for one representative trial of the word 'right' collected with the ear microphone placed inside the ear canal and in front of the mouth. Shading in the spectrogram plots represents low and high energy levels of the speech signal. Similar low-pass behavior was observed for all words (Bulbuller et al. 2006). We also noted that some of the recorded data included bodily-created noises such as gulps, tongue clicks, lip smacks or coughs picked up by the sensitive

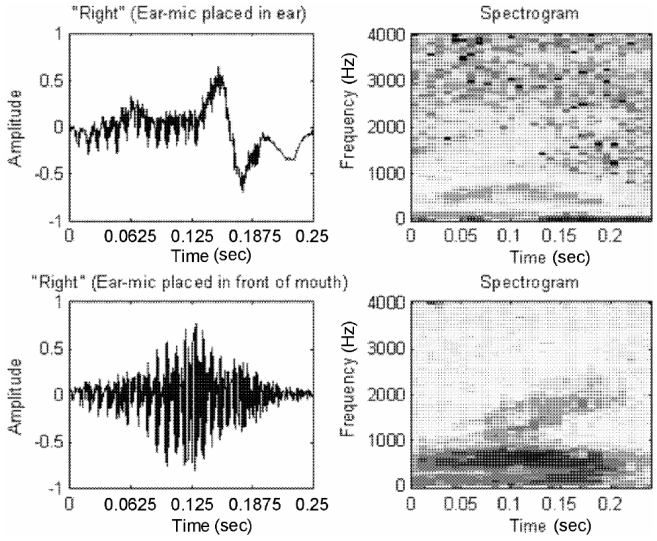


Fig. 5. Word 'right' recorded via ear microphone.

microphone device. In addition, the microphone and the A/D converter used also introduced a 50–60 Hz humming noise from the collection equipment, and a DC bias.

5.1.2. Data pre-processing

In order to prepare the data for processing, the DC offset was first removed. Next, the data was passed through a 6^{th} order infinite impulse response (IIR) elliptic high-pass filter to ensure a sharp cutoff behavior would remove the low frequency equipment noise and preserve speech information above 100 Hz. The frequency specifications of the selected filter configuration were a stopband equal to (0, 60 Hz), passband equal to (100, 4000 Hz) and passband ripple equal to 0.5 dB. We noted that the filter nonlinear phase characteristics did not result in perceivable speech distortion.

5.1.3. Speech endpoint detection

Determining the beginning and the termination of speech in the presence of background noise is a complicated problem, except in cases of high signal-to-noise ratios (∼30 dB or better) which are rarely seen in real world applications (Deller et al. 2000). Accuracy in the speech segmentation step is essential as numerous results have shown that the efficiency of accurate endpoint detection has a significant and direct effect on the performance of the associated recognition system (Qiang and Youwei 1998; Ying et al. 1993). In addition, a noise-robust endpoint detection algorithm must also be capable of dealing with speaker-dependent disturbances like coughs, gulps, tongue clicks, lip-smacks, etc. (Srydal et al. 1995), which we have observed to produce an acoustic signature in the ear. The collected speech also included mechanical and bodily-created noises.

Two simple quantities are normally used in speech detection algorithms: short-term energy and zero-crossing rates (Deller et al. 2000). The zero crossing rate (ZCR) leads to a simple scheme which allows users to track rough changes in a signal frequency behavior by computing the rate at which a zero-mean signal changes sign. Low frequency signals have samples which tend to stay of the same sign longer than high frequency signals. Thus, changes in the signal frequency content may be tracked by monitoring changes in the ZCR. Rabiner and Sambur (1975) proposed a speech endpoint detection scheme based on a combination of the short-term energy and zero-crossing rates. The main advantages of this scheme are that it is computationally inexpensive and can be implemented for online speech segmentation; numerous variants have been proposed to detect the speech endpoints over the years.

This algorithm was customized to our in-ear speech endpoint detection problem (Bulbuller et al. 2006). The resulting scheme is a two-step search algorithm where the short-time energy quantity is first applied for a coarse segmentation task. Second, the zero-crossing measure refines the coarse boundaries by considering the signal behavior around the initial endpoint estimates. The zero-crossing measure applied in the second search is designed to help detect low-energy phonemes at the beginning or end of the word, especially when dealing with weak fricatives (such as /f/, /th/, /h/), plosive bursts (such as /p/, /t/, /k/) or final nasals (such as /m/, /n/, /ng/). Figure 6 shows short-term energy and zero-crossing measures for a typical utterance of the word 'left' from ear-microphone data. In this study we selected rectangular frames of 10 ms with 50% overlap for the speech segmentation phase.

Coarse endpoint detection phase. The mean and the standard deviation of the short-term energy and zero-crossing measures are first computed during the first 50 ms of recording, assuming there is only background noise in that interval, to establish a noise floor reference for each recording. Upper and lower threshold values (shown as T_u and T_l on the upper plot of Figure 6) for the short-term energy and a threshold (shown as T_{zc} on the lower plot of Figure 6) for the zero-crossing measure are set based on the noise-only segment statistics and constants determined by trial and error on the database under study, as follows:

$$T_l = 8 \times \mathrm{STE}_{\min}, \quad T_u = 32 \times \mathrm{STE}_{\min} \tag{4}$$

$$\mathrm{STE}_{\min} = \min[0.25, \mathrm{mean(STE)} + \mathrm{std(STE)}] \tag{5}$$

$$T_{zc} = \min[0.25N, \mathrm{mean(ZCR)} + \mathrm{std(ZCR)}] \tag{6}$$

where N is the frame length; STE and ZCR represent short-term energy and zero-crossing rate, and std(x) represents the standard deviation operation, respectively.

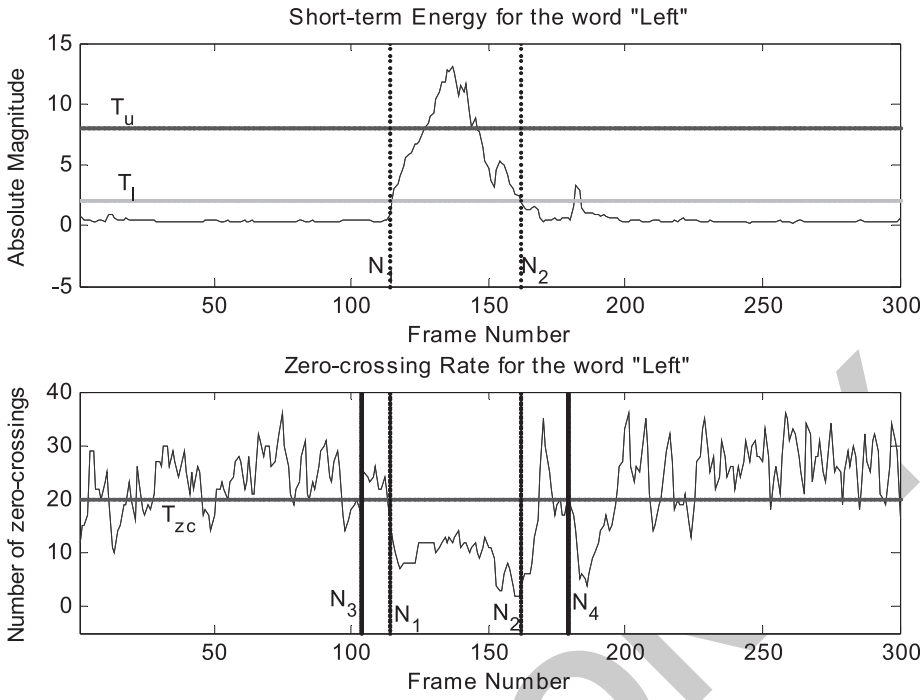


Fig. 6. Typical example of the use of short-term energy and zero-crossing rate in endpoint detection.

First, the short-term energy parameter is monitored to find the first and second successive crossings of the upper threshold T_u. Next, the scheme searches around these first and second crossings to seek the coarse endpoints (shown as N_1 and N_2), where the lower threshold T_1 is first exceeded. This initial search yields the tentative endpoints N_1 and N_2 shown as dotted lines in Figure 6.

5.1.4. Endpoint detection refinement phase

In the next step of the algorithm, a fine search on the zero-crossing plot is performed, moving toward the ends from N_1 and N_2 for no more than 10 frames, by examining the zero-crossing rate to find three occurrences of counts above the threshold T_{zc}. Finally, the final endpoint estimate is moved backward and/or forward to the first threshold location crossing if three such occurrences are found, as is illustrated with N_3 and N_4 shown as solid lines in the bottom plot included in Figure 6. Final endpoints remain at the initial estimates N_1 and N_2 when three such occurrences over the zero-crossing threshold are not detected. Finally, detected segments of duration $<$ 100 ms are considered as false alarms and dismissed during the search, as the words considered in this study have a longer duration.

The overall endpoint detection scheme performs well on average with the main exception that it produces some errors detecting the word ending for 'left'. This is most likely due to the soft /t/ sound present as a result of the low-pass filtering effect, resulting from the in-ear microphone setup which was noticed in most of the recordings.

5.1.5. Speech feature extraction

Feature extraction is designed to convert the signal into a compact set of parameters while preserving speech signal information. First, the signal is divided into frames of 256 samples (corresponding to 32 ms) with an overlap of 100 samples (roughly corresponding to 40% overlap) from frame to frame and a Hamming window is applied to each frame. Next, features are extracted from each frame. We selected Mel-Frequency Cepstral Coefficients (MFCCs), as they have been extensively used as features in speech recognition (Davis and Mermelstein 1980) and have been shown to outperform other parameter types in speech recognition applications.

However, these features do not contain information regarding the speech signal dynamic evolution, which also carries relevant information in speech recognition (Becchetti and Ricotti 1999). Further improvements in recognition performance can therefore also be obtained by taking into account the dynamic characteristics of the MFCC features (Deng and O'Shaughnessy 2003). The simplest approach to obtain these dynamic features takes the basic difference of coefficients between consecutive frames. The resulting delta-MFCC coefficients reflect cepstral changes over time. However, researchers have also argued that the basic differencing operation

Table 1. Confusion matrix for average word recognition rates (percent): average classification = 92.77%

	Up	Down	Left	Right	Kill	Pan	Move
Up	92.5500	3.2125	2.3375	0.4375	0.3458	0.8208	0.2958
Down	0.3042	92.3333	1.7958	0.3250	2.3292	2.3500	0.5625
Left	2.1375	1.6875	87.6833	7.0792	0.6542	0.3625	0.3958
Right	0.4875	1.0750	2.6417	94.6708	0.4000	0.3792	0.3458
Kill	0.3125	1.7250	0.7917	0.4667	94.2292	2.2500	0.2250
Pan	0.0917	3.5500	0.9375	0.3375	1.8708	91.9458	1.2667
Move	0.6875	0.1000	1.4542	1.2875	0.2917	0.2167	95.9625

is too sensitive to random inter-frame variations and should be replaced by a smoother estimate of the local time derivative (Deng and O'Shaughnessy 2003). As a result, we used the following regression on the set of MFCC coefficients (Westerlund et al. 2002) to generate the set of delta-MFCC coefficients:

$$d_k = \frac{\sum_{\alpha=1}^{M} \alpha (c_{k+\alpha} - c_{k+\alpha})}{2 \sum_{\alpha=1}^{M} \alpha^2} \tag{7}$$

which is equivalent to passing the static MFCCs through a linear differential filter. Finally, delta MFCC parameters were rescaled to the range of the MFCCs to ensure better-conditioned vector-quantized features.

5.1.6. Speech recognizer

Designing an isolated word recognition system first involves extracting a set of characteristic speech feature parameters from each word and tuning a specific classifier type. Hidden Markov Models (HMMs) are used extensively in today's modern automatic speech, and were selected as a tool for the first analysis of in-ear speech data. This study implemented a discrete observation left-to-right HMM (DHMM); model sizes of 5–8 states were considered. Our experiments indicated that 8 states were sufficient to model the linguistic units of phonemes present in the vocabulary.

Figure 7 summarizes the overall steps involved in the design of the HMM classifier considered in our study. First, the speech signal was split into overlapping frames by applying a Hamming window of length $N = 256$ samples (corresponding to 32 ms for a sampling frequency equal to 8 KHz) with an overlap of 100 samples (corresponding to about 40%). Second, the first 12 MFCCs and 12 delta-MFCCs were extracted from each frame of the segmented speech, resulting in feature vectors of length equal to 24. Third, DC bias from the MFCCs was removed and delta-MFCCs rescaled to the range of the MFCCs. Two-thirds of the data were used in the training phase to estimate the HMM parameters, while the last third was used in the testing phase to evaluate the recognizer performance. Thus, the codebook was generated from feature vectors contained in the training set and used to generate the HMM parameters.

Table 2. 95% confidence intervals for average recognition rates shown in Table 1

Word	95% confidence intervals (%)
Up	86.00 – 96.00
Down	86.00 – 95.33
Left	79.33 – 93.33
Right	92.33 – 97.00
Kill	91.00 – 97.33
Pan	86.67 – 96.00
Move	93.00 – 98.00
Overall Classification	91.10 – 94.29

Recognizer results. Tables 1 and 2 show average classification performances obtained and associated 95% confidence intervals obtained after 80 experiments for the testing sets. Results show an overall recognition rate of 92.77%. Results also show the worst performances are obtained for the utterance 'left' with an average recognition 4.3% below the next higher word recognition rate. We noted that misclassified 'left' utterances occurred for trials where ending boundaries did not include the low energy 't' sound which was supposed to be present at the end of that word. As a result, we surmise these errors were mostly caused by the incorrect detection of the speech boundaries, due to the unvoiced low-energy ending sound 't' in that specific word (Kurcan 2006).

5.2. Tongue movement recognition

At this time, we have developed a strategy to accurately detect and classify, in real time, changes in the air flow pressure that occur in the ear canal caused by tongue movements,

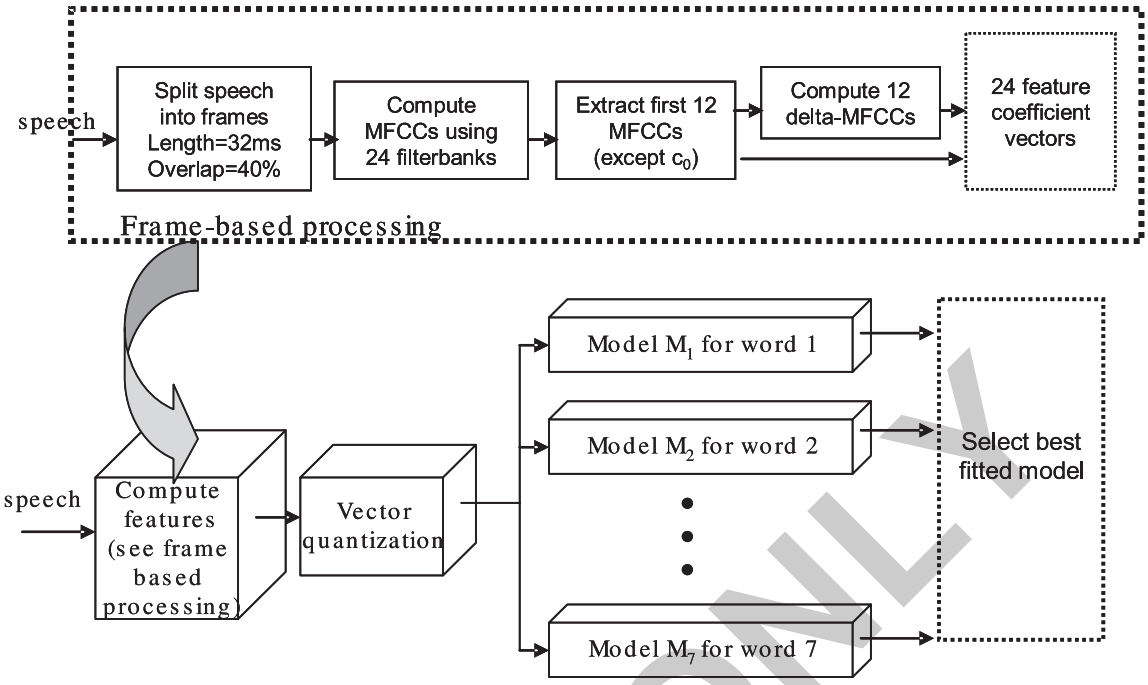


Fig. 7. Average distortion versus codebook size.

or tongue-movement ear-pressure (TMEP) signals. The strategy to first detect the presence of an unknown TMEP signal and then classify the TMEP signal is shown in Figure 8. We have developed a new Decision Fusion classification strategy which improves upon the unique strategy described previously by Vaidyanathan et al. (2006; 2007). Our past work used nearest mean classifiers, whereas the results presented in this paper uses Gaussian classifiers. The decision fusion classification strategy is based upon classifying the TMEP signals at the time-instants in which the maximum *M*-class discrimination occurs, fusing the resulting classification decisions into a discrete decision vector and classifying the decision vector. The time-instants of the TMEP signals are first ranked according to their individual classification accuracies. During testing, the time-instants with the highest ranks are classified independently and the decisions are fused into a single decision fusion vector. The resulting decision fusion vector is classified using a discrete Bayes classifier.

The averaged classification results using our new Gaussian classifier for 8 test subjects with ages ranging from 19 to 54 (who were given several hours to practice the signals), are presented in Table 3, which shows a confusion matrix enumerating classification accuracies. The 4 TMEP signal classes – left, right, up and down – are represented. The confusion matrix part of the results can be interpreted by examining the first row which shows that on average, 97.39% of the test TMEP signals of class 'left' was classified correctly as belonging to that class, 0.9% were misclassified as 'right', 0.84% were mis-

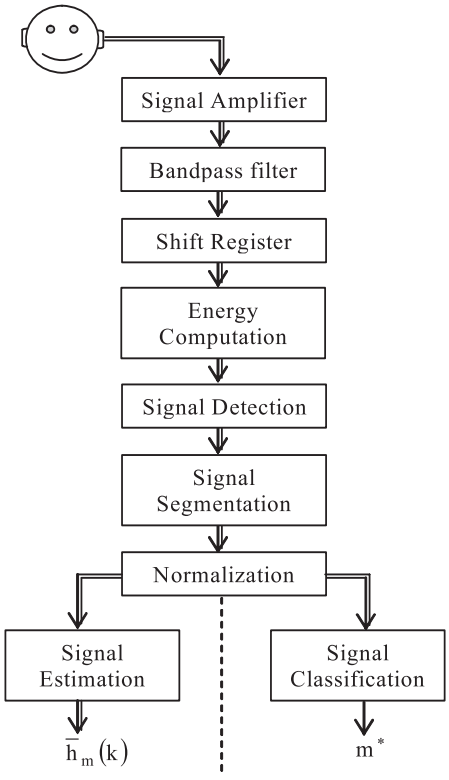


Fig. 8. Block diagram of the TMEP signal estimation and classifiction strategy.

Table 3. Gaussian decision fussion classifier accuracy: total accuracy = 97.51%

	Left	Right	Up	Down
Left	97.39	0.90	0.84	0.87
Right	0.84	97.37	0.53	1.26
Up	0.22	0.97	97.71	1.09
Down	0.38	1.52	0.52	97.57

classified as 'up' and 0.84% were misclassified as 'down'. The total accuracy is a compilation of each class accuracy.

5.2.1. Simple and compound tongue movements

The four simple tongue movements (left, right, up, and down) were selected because they can be made quite easily by most individuals and intuitively correspond to maneuvers to steer a robot. We have assumed the robot being controlled lacks the on-board intelligence to distinguish context dependent signals. Consequently, each tongue action initiates a unique command signal and four tongue movements can only result in only four robot actions. The most obvious approach to increase the number of command signals is to increase the number of simple tongue movements, for example, 45 degrees to the left-up, 45 degrees right-up, 45 degrees left-down and 45 degrees right-down. The drawback of this approach is that the performance will deteriorate due to the higher overlap in the resulting ear pressure signals. Furthermore, the performance will also deteriorate because it becomes increasingly difficult for an operator to repeat, in a consistent fashion, tongue movements with smaller differences.

We introduce an alternative strategy in which the number of command signals is increased without increasing the number of tongue movements. The most important reason for selecting the four simple tongue movements is that the number of command signals can be increased by forming compound tongue movements consisting of 2 simple tongue movements. For example, the number of command signals can be doubled by repeating each tongue movement twice with a brief pause (maximum = Δ) between the 2 tongue movements. As an example, the compound tongue action (Left/Left) ($m = 1$)/($m = 1$) can be formed by flicking the tongue twice to the left with a pause less than $t = \Delta$. With additional practice, cross compound tongue movements such as Left/Right (m = 1)/(m = 2), L/U (m = 1)/(m = 3), L/D (m = 1)/(m = 4), …., D/U (m = 4)/(m = 4) can also be included to increase the number of command signals to a total of 20 for more complex human–machine interface control and communication. The total number of commands that can be generated from M simple tongue movements in this manner is ($M + M^2$), significantly more than a traditional joystick.

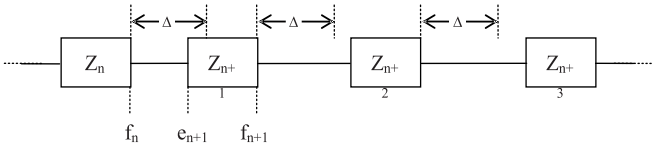


Fig. 9. Stream of simple and compound TMEP signals shown as blocks.

Classification of compound TMEP signals. A compound TMEP signal $h_{m/n}(k)$ can be classified by first detecting and then classifying the compound signal. The number of classes of the TMEP signals will, therefore, be ($M + M^2$). When M = 4, the total number of signal classes is 20, which is quite high for any pattern classification problem. It is well-known in pattern recognition theory that the classifier performance deteriorates when the number of pattern classes increase. We propose a novel M-class signal classification strategy for classifying the ($M + M^2$) TMEP signals, each simple TMEP signal and each component of a compound TMEP signal. Note that although the number of simple and compound TMEP signals is ($M + M^2$), the number of distinct components is only M. Classification, therefore, involves first detecting whether a simple TMEP signal or a compound TMEP signal movement was generated.

To illustrate the procedure, consider a time segment containing a stream of TMEP signals which includes both simple and compound TMEP signals. Figure 9 shows a stream of one compound TMEP signal Z_n/Z_{n+1} and two simple TMEP signals Z_{n+2} and Z_{n+3}. The signals, shown as blocks, are used to mark the start and end points of the TMEP signals. Let δ be the time required to classify a signal after the end point f of the signal is detected, Δ be the maximum permissible time between the two simple TMEP signals of a compound signal, and k_n be the nth sample after the end-point f_n of Z_n is detected. Z_n is detected as a simple TMEP signal only if

$$(k_n - f_n) > \Delta. \tag{8}$$

Z_n and Z_{n+1} are detected as a composite signal if

$$k_n = e_{n+1} \quad \text{and} \quad \delta < (k_n - f_n) < \Delta. \tag{9}$$

The advantage of increasing the number of commands by forming compound tongue movements is that the same detection and classification algorithms designed for simple tongue movements can be used to detect and classify each component of the compound tongue movement. Most importantly, the number of signals to be detected and classified does not increase. Consequently, the performance will not deteriorate with an increase in the number of command signals. If there are no detection errors, the performance of a machine interface using simple and compound TMEP signals will be identical to that using only simple TMEP signals. The only drawback

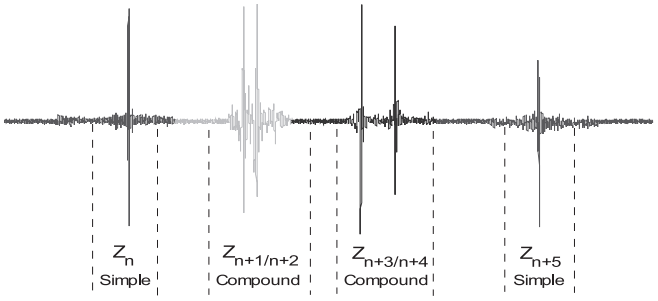


Fig. 10. A stream of real simple and compound TMEP signals.

is the introduction of a time delay for converting a simple or compound tongue movement into a command signal after the simple or compound tongue movement has occurred. The delay is equal to Δ for a simple tongue action and a maximum of Δ for a compound tongue movement. In practice, Δ is a parameter that depends on how comfortably a particular user can make compound tongue movements. The additional delay of Δ will therefore be quite acceptable given the rapid detection and classification of the TMEP signals. Figure 10 shows an example of a stream of real simple and real compound TMEP signals which were detected and classified correctly using the parameters δ and Δ.

5.2.2. Multi-channel tongue movement recognition strategy

In order to improve the classification accuracy, we have also developed a new dual-channel classification strategy which exploits information from the TMEP signals of the left and right ear pressure channels. Clearly, this strategy will only be effective if the signals from the left and right channels, corresponding to the same tongue movement, carry different aspects of the tongue movement. Furthermore, the information from the two channels must also be complimentary. Figure 11 shows pairwise averaged estimates of several trials of the left and right channel TMEP signals corresponding to the 'down' tongue movement for a test subject. It is interesting to note that the signals in each ear are quite different for the same tongue movement. It could, therefore, be concluded that the TMEP signals of the left and right channels do indeed reflect different aspects of a given tongue movement. The challenge is to develop models that combine the complementary information from the two channels in order to improve the classification accuracies of TMEP signals.

Rather than presenting a formulation for a limited two-channel strategy, we describe a more general C-channel formulation for which the two-channel strategy is a special case ($C = 2$). A multi-channel decision fusion model is introduced for combining the information from the multiple channels. In future work, this multi-channel strategy will be implemented to combine tongue movement with other physiological signals (EEG, etc.) to synergize with and augment existing human–machine interface mechanisms.

Multi-channel decision fusion strategy. The multi-channel decision fusion classification strategy is summarized in Figure 12. The N samples of the TMEP signal of channel c are represented by: $h^c(k)$, $k = 1, 2, \ldots, N$; $c = 1, 2, \ldots, C$.

The classification of the individual channel signals is similar to the single channel decision fusion strategy described by Vaidyanathan et al. (2007). The formulation presented in this paper will therefore begin at the multi-channel decision fusion stage. Let

$$D(k) = \bigtriangledown_{c=1}^{C} d^c(k) \tag{10}$$

where $d^c(k)$ is the decision of the c^{th} channel classifier at time instant k and ∇ represents the concatenation operation. That is,

$$D(k) = \left[d^1(k), d^2(k), \ldots, d^C(k)\right]^T \tag{11}$$

is the fusion vector formed by concatenating the decisions of the C channel classifiers at the time k^{th} instant. Note that each channel makes an independent decision; therefore, $d^c(k)$, $c = 1, 2, \ldots, C$ are independent. The decision fusion vector $D(k)$ is a discrete random vector in which each element can take one of M values. Let the probability density function (PDF) of $D(k)$ under category m be $P[D(k)/m]$. Then, the Bayes decision function for class m can be written as

$$g_m[D(k)] = \ln P_m + \ln P[D(k)/m] \tag{12}$$

where P_m is the *a priori* probability of class m. The final decision m_k at the k^{th} time instant resulting from the fusion of the C decisions at the k^{th} time instant is given by

$$m_k = \arg\max_m [g_m(D^c(k)], \quad k = 1, 2, ..., N. \tag{13}$$

The discriminant functions can be derived explicitly by setting

$$p^c_{k;a/m} = P[d^c(k) = a/m], \quad c = 1, 2, ..., C. \tag{14}$$

The left side of Equation 13 is the probability that $d^c(k) = a$ when the true class is m. The PDF of $D(k)$ under the class m, $m = 1, 2, \ldots, M$, can then be written as

$$P[D(k)/m] = \Pi c = 1C \left(p^c_{k;1/m}\right)^{\delta[d^c(k)-1]} \times \left(p^c_{k;2/m}\right)^{\delta[d^c(k)-2]} \cdots \left(p^c_{k;M/m}\right)^{\delta[d^c(k)-M]} \tag{15}$$

where

$$\delta(x-a) = \begin{cases} 1 & \text{if} \quad x = a \\ 0 & \text{if} \quad x \neq a \end{cases}. \tag{16}$$

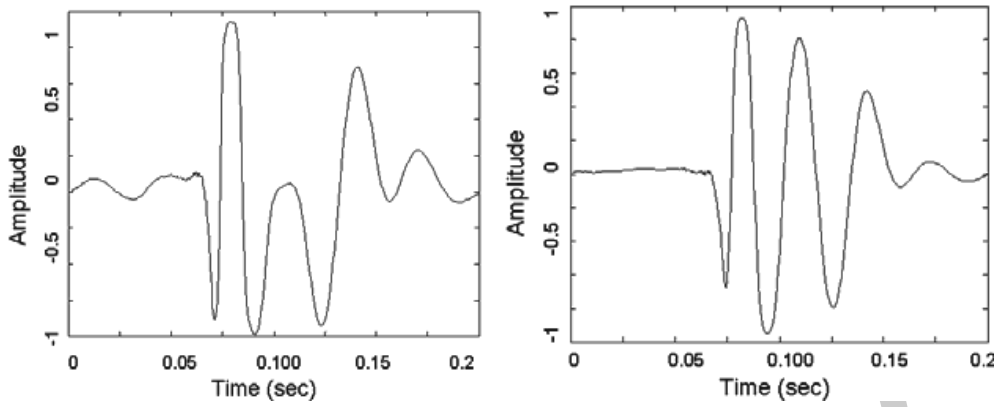


Fig. 11. Pairwise averaging estimates of the left and right TMEP signals corresponding to the 'Down' tongue movement.

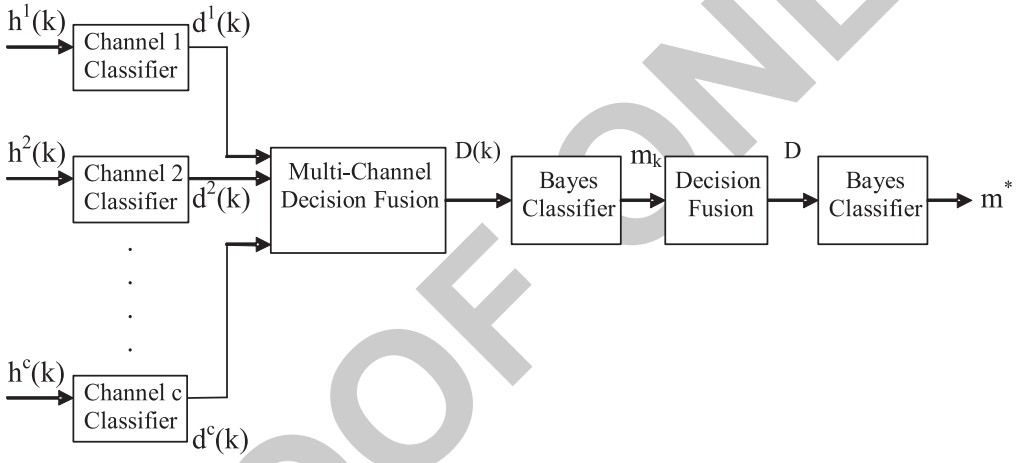


Fig. 12. The multi-channel decision fusion strategy.

By substituting the PDFs into Equation 11, it can be shown that the discriminant function for category m can be written as

$$g_m[D(k)] = \sum_{c=1}^{C} \begin{bmatrix} \delta[d^c(k)-1]\ln(p^c_{k;1/m}) \\ +\delta[d^c(k)-2]\ln(p^c_{k;2/m})+ \\ \ldots+\delta[d^c(k)-M]\ln(p^c_{k;M/m}) \end{bmatrix} + \ln P_m. \quad (17)$$

In the following step, the N decisions are fused into an N-dimensional decision fusion vector D. That is,

$$D = \nabla k = 1Nm_k. \quad (18)$$

In a manner similar to the derivation of the Bayes discriminant function for $D(k)$, it can be shown that the Bayes discriminant function of D under class m is given by:

$$g_m(D) = \sum_{k=1}^{N} \begin{bmatrix} \delta[m_k-1]\ln(p_{k;1/m}) \\ +\delta[m_k-2]\ln(p_{k;2/m})+ \\ \ldots+\delta[m_k-M]\ln(p_{k;M/m}) \end{bmatrix} + \ln P_m, \quad (19)$$

where the left side of Equation 13 is the probability that $m_k = a$ when the true class is m. The final multi-channel decision is given by

$$m^* = \arg\max_m [g_m(D)]. \quad (20)$$

To summarize, for each channel, a decision is made at each sampling instant using a scalar Gaussian classifier and the C decisions are fused into a multi-channel decision fusion vector of dimension C. The decision fusion vector is classified using a discrete Bayes classifier in order to determine the class of the TMEP signal at each sampling instant. Finally, the N de-

Table 4. Single-channel left ear

(a)	Up	Down	(b)	Up	Down
Up	98.12	1.88	Up	96.22	3.78
Down	7.78	92.22	Down	10.34	89.66
Class Acc = 95.17%			Class. Acc = 92.94%		

Table 5. Dual-channel decision fusion results

	Up	Down
Up	99.14	0.86
Down	1.32	98.68
Class. Acc = 98.91%		

cisions are fused into an N-dimensional decision fusion vector which is classified by a Bayes classifier to determine the class of the TMEP signal across all the sampling instant. A ranking strategy (Vaidyanathan et al. 2007) can also be used to select only those L ($L < N$) time instants that yield the best decisions. Consequently, this strategy requires $L \times C$ scalar Gaussian classifiers and $L + 1$ Bayes classifiers.

Multi-channel decision fusion algorithm test results. In this study, dual-channel data was collected from a 27-year-old male subject (with no training whatsoever) for two tongue movements: up and down. A total of 100 signals per tongue movement were collected from each channel. Using the random partitioning method for dividing the signals into training and test sets (Vaidyanathan et al. 2007), results were averaged over at least 100 partitions.

The averaged classification results are presented in Tables 4 and 5, which show confusion matrices enumerating classification rates and accuracies. Table 4 shows the single-channel results for the (a) left and (b) right channels. Table 5 shows the dual-channel results for the multi-channel decision fusion strategy. It is clear that the two multi-channel strategies are superior when compared with the single best channel in this preliminary study. We are confident that the improvement would be more dramatic in a larger study. Furthermore, the generalized formulation makes the strategies applicable to numerous problems involving the classification of multi-sensor, multi-category signals. We can also expect an improvement in performance if information can be exploited from a larger number of channels.

6. Design of Machine Interface

6.1. Introduction

In order to definitively establish the veracity of our system for robotic operation, we designed and implemented a range of simulation experiments to characterize its performance and provide quantitative data on how error rates and misclassifications in speech and tongue movement translate directly into measurements of robot performance. The purpose of the simulations was to address questions related to the performance of the system for general robot maneuvering, assess the incidence of collision avoidance or mishap due to mistaken recognition, measure system performance in the presence of degraded recognition rates and compare the two operating modes over several thousand trials.

Although the system has successfully been implemented in real-time (Think-A-Move Ltd 2005, 2006; Koliousis 2007), the collection of data for rigorous performance assessment is not practical for a significant number of tests. Furthermore, it is virtually impossible to distinguish between a human error and a classification error with an actual operator in a virtual or actual robotic test. A candidate platform was therefore selected to parameterize the simulation based on envisioned field use (for robotic scouting and reconnaissance) with simulated performance based on data recognition rates from all test subjects.

6.2. Robotic platform

The mobile robot called Whegs™ II (Quinn et al. 2002) was selected as a candidate device for testing the performance of the dual mode human–robot teleoperation interface. While the utility of our interface lends itself to a breadth of platforms for several applications, Whegs™ II is an ideal demonstration robot for the interface given that it has shown the autonomy necessary to perform tasks based on high-level commands and has exceptional mobility with few low-level control inputs. Specifically, the robot's passive mechanisms lend it the mobility to move over irregular and rugged terrain without the need for force feedback control (which is not feasible for our system) using only three low-level inputs (speed, heading and body flexion angle) and without the need for complex control software.

Whegs™ II has the mobility and autonomy necessary to take advantage of both the high-level and low-level commands of the dual mode interface. Whegs™ II with tactile antennae has been shown capable of autonomously flexing its body and varying its speed appropriately to surmount rectangular steps (Figure 13) (Lewinger et al. 2005). In that experiment, the path was narrow and the operator controlled the heading of the vehicle with a joystick. With an interface possessing the capabilities of our dual-mode system, an operator could give Whegs™ II the high level speech command 'climb the stairs' and then use tongue movements to steer the vehicle while monitoring its progress remotely. In this way, an operator using our dual mode interface could use Whegs™ to help perform meaningful tasks such as searching a burning building for survivors while the operator' hands are free to carry survivors to

Fig. 13. Whegs™ II flexing its body as it climbs over a kerb.

safety. Whegs™ II is also an excellent platform for performance assessment for future scouting and exploration applications (Karlsen 2004) which are being commercially pursued at this time.

The footprint of Whegs™ II is 47 cm long by 36 cm wide and it weighs 3.86 kg. It has a two-piece aluminum frame and it can flex 30 degrees up and down about its middle axle. It has torsionally compliant devices in all six of its axles. The whegs have internal linear springs (2280 N m^{-1}) that permit them to comply radially. Its radial wheel-leg-spoke length is 10 cm when no load is applied. It uses a 90 W Maxon motor with a 26:1 integral transmission to propel it, two small hobby servos for steering, and a larger hobby servo to activate the body joint. Its two 7.2 V battery packs are placed on its rear body segment such that its center of mass is in the rear and it can lift its front body half. Speed, steering and body joint motion are controlled via a hobby RC system. Whegs™ II can run at 3 body-lengths per second. Using its body flexion joint, it can readily climb a series of obstacles that are 1.38 spoke-lengths high and 0.8 body-lengths deep. It can also run as a quadruped on its middle and rear whegs while holding its front airborne.

6.3. Simulation parameterization

Two simulated missions were tested with the interface controlling the Whegs™ II robot, each suited to a different mode of interface. The first was an open environment search where the robot was required to reach a set of waypoints akin to an outdoor mapping mission (suited to higher-level interaction with speech commands) while the second directed the robot to move through a dense obstacle-rich indoor environment (suited to lower-level interaction with tongue commands). Parameters of the robot programmed into the simulation included its footprint (size), speed, acceleration and turning capacity (ability to change heading), which were gathered experimentally. Terrain was specifically not considered in the simulations due to the capacity of Whegs™ II to passively adapt its gait to various substrates and move over obstacles. Parameters of the interface included in the simulation were based on data gathered from test subjects enumerated in Sections 3, 4 and 5. Specific information included: operator time to generate a signal (t = 0.2 s for tongue movement, t = 0.25 s for speech), minimum operator reaction time between signals (Δt = 0.2 s for tongue movement, Δt = 0.5 s for speech), recognition accuracy and rate of misclassification. Note that recognition accuracy for tongue movement was based on data from individual users for whom the device had been calibrated, while speech data was taken as an average for the entire data set for all users assuming no specific calibration for each user.

The processing time of the pattern classifier for both tongue movement and speech was judged to be negligible in the simulations. For tongue movement, the scalar classifiers (which can be implemented in parallel), are simple univariate classifiers. Each discriminant function, one for each class, requires only one simple multiplication and one difference operation during use. Furthermore, the decision fusion classifier described in Section 5.2 was specifically designed to make a recognition decision based on only a small segment of data. For the decision fusion part, there is only one discriminant function which requires multiplying L terms and summing $L - 1$ terms. Testing has shown typical values of L = 50 for most users. For speech, HMM classifiers have a well-established history of working in real-time, specifically in the case of our commands which were limited to monosyllabic words with a 0.5 s pause between user inputs. Consequently, neither command mode requires enough floating point operations to add significant delay in real-time operation.

6.4. Robot control through speech interface

A first generation conceptual design of the human–machine interface system for the first mode of operation, control of the Whegs™ II robot based on aural speech recognition, has been completed. We propose a straightforward system designed around four words for motion control. These are centered on the words ‘up’, ‘down’, ‘left’, and ‘right’. These four words can be coupled to create an intuitive interface such that a ‘right’ command corresponds to a right movement, with ‘left’ following naturally. Note that the passive mechanisms of the robot are expected to adjust to handle terrain fluctuations. For control, each left/right command may indicate half a cycle of a sinusoidal input to the steering angle of the vehicle. The device will thus turn a fixed increment on each left/right input, and subsequently resume a straight path in accordance with the new heading. Repeated commands may increase the direction of the turn. Forward and reverse motions are controlled with ‘up’ and ‘down’ commands (which were selected rather than ‘forward’ and ‘backward’ in order to keep monosyllabic commands). An ‘up’ command inputs a forward velocity signal while a ‘down’ movement results in a backwards velocity

input. The forward/backward velocity is altered in a fashion where each additional command would increase or decrease speed by a fixed increment. Finally, in the proposed interface, a 'kill' command executes an all stop for the robot. All enumerated commands would be very straightforward to implement in a standard communication setup.

Note that seven words were actually recognized in our study: up, down, left, right, move, pan, kill. These words were chosen as a set of commands a robotic operator may give to a Whegs robot moving in the field for missions suggested by Karlsen (2004). The commands 'kill', 'move' and 'pan' would allow even greater versatility by allowing an all stop command as well as the potential to switch control devices (e.g. 'pan' may switch venues to control a camera on the robot with commands of 'up', 'down', etc.).

6.4.1. Open environment control simulation in speech mode

A version of this control interface was implemented in simulation to prove that the current speech recognition accuracies are sufficient for robotic control. Speech recognition errors were included based on the accuracies presented in Table 2. For example, if a 'right' command was performed, a 94.7% probability of the robot receiving this signal was assumed, with a 2.6% probability of the robot receiving a 'left' command and a 0.4% probability of a 'down' command. If the recognition returned a directive that was not used in the simulation (e.g. 'pan'), the robot did not acknowledge the command and continued on its previous course of action.

Figure 14 shows the results of a simple simulation where this interface was implemented to direct a robot to reach a series of (20) waypoints in a planar (200 m × 200 m) work space. The + symbols represent the waypoints with the path of the robot shown. The waypoints were spaced arbitrarily across a 200 m amplitude sinusoidal path with a period of 80 m. In each case, the control logic steering the robot or virtual operator was provided the planar position of each successive waypoint and the robot's position. The virtual operator maneuvered the robot towards the next waypoint based on the previously discussed commands to align the robot direction with a line-of-sight vector to the next waypoint as an actual operator would. The reason this was done in simulation was to generate performance data over several thousand repetitions of the task and due to the difficulty of separating operator and classification errors.

A waypoint was considered reached if the robot successfully executed a stop within 20 cm of that waypoint. It should be noted that the location of the next waypoint was withheld from the virtual operator until the current waypoint had been reached. Although the Whegs™ robot turning radius and the small time delay for speech commands did not allow straight line motion to some waypoints (which would have been possible for a robot with differential steering), as can be seen from

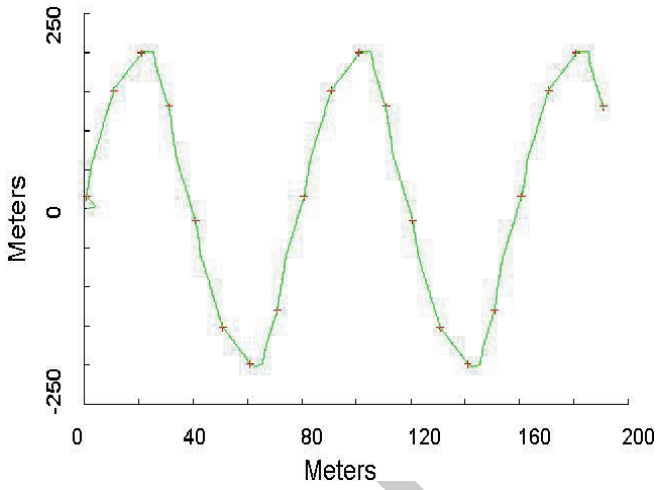


Fig. 14. Open environment robot waypoint navigation.

the figure, the robot successfully reached all 20 waypoints. In order to assess the impact of any erroneous operator commands, this particular simulation was repeated 1000 times. In every case, the robot successfully reached all waypoints without fail. While some commands were mistaken by the interface, the time span and accuracy at which commands may be given allowed rapid correction. Note that we have also controlled the robot for the same simulation using tongue movements (Vaidyanathan et al. 2004), but this required a higher degree of operator interaction.

6.5. Robot control through tongue movements

At present, we have identified more than ten distinct repeatable movements of the tongue that provide traceable pressure signature which can be captured by the microphone-earpiece housing. The control interface can thus be tailored to any set of movements appropriate to the robot being controlled. In practice, however, larger amounts of movement involve higher levels of complexity and a steeper learning curve for the operator.

We propose a straightforward system designed around the standard interface described earlier (four movements) for control of the Whegs™ II platform. The up/down/left/right tongue movements in this interface controlled the robot in a parallel fashion to the up/down/left/right voice commands, but with shorter command input corresponding to the speed of tongue movement. Beyond forward/reverse and left/right motions, additional commands may be necessary to control the robot. In order to correlate robot actions to additional movements, we propose the use of compound tongue movements as described in Section 5.2.1. Three inputs beyond motion commands are necessary to complete the Whegs™ II control interface. In order for the robot to navigate harsh terrain, the operator must

be able to specify 'flex up' and 'flex down' commands to the body flexion joint and, of course, an 'all stop' command to halt robot motion. In the proposed interface, two compound bottom movements execute an all stop command, while bottom/top and top/bottom tongue motions correlate to up and down body flexion respectively. Finally, it is very important to note that the tongue movements to be used are very gentle, and will not tire or fatigue the tongue any more than speech.

6.5.1. Constrained environment maneuvering simulation

The second implemented simulation consisted of the robot moving through an obstacle-rich indoor environment where speed of movement and clutter would necessitate lower-level tongue control. While collisions with obstacles occasionally occur with virtually all existing interfaces, it is critical that they be kept to an absolute minimum. Thus, we conducted a series of simulations using performance parameters of the Whegs™ II robot in order to test the ability of our interface for maneuvering in cramped environments (e.g. such as hallways or between furniture). The same virtual operator was tasked to maneuver the robot through such an environment using tongue movements with the aural robotic interface. Signal recognition accuracies for each movement were measured for eight individual test subjects using the Gaussian decision fusion classifier to provide a realistic appraisal of the robot's performance.

Figure 15 shows the results of one such simulation. In the simulation shown, the Whegs™ II robot, under the control of a test subject, was placed in an environment comprised of a variety of obstacles forming a narrow canyon only slightly wider than the vehicle itself. The robot begins the simulation in the lower left portion of the figure (at the origin) with the goal of maneuvering out of the constrained environment through the narrow passage on the lower right portion of the figure. The path of the robot as it moved through the room is shown, with the robot itself illustrated at key positions along the path. The arrows overlaying the robot show its heading at the illustrated positions along its path. In addition to maneuvering through the room, the virtual operator was also tasked to stop the robot completely at each of the illustrated points to ensure the ability of the system to stop the robot at a desired location was incorporated into the simulation.

Of critical importance to note is that while some tongue movement commands were mistaken by the system (approximately 20 commands were identified incorrectly in the simulation shown), and despite the very narrow corridors the robot maneuvered through, no collisions between the robot and obstacles were recorded in the simulation**.** The high rate of recognition accuracy and speed at which tongue commands may be given allow for immediate correction, thus all potential collisions may be avoided. For a collision to occur, three or more

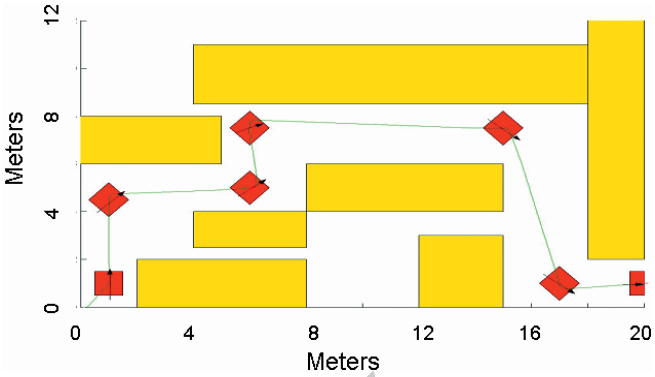


Fig. 15. Robot constrained environment navigation.

commands would likely have to be mistaken in sequence at a very specific moment, which is less than a 0.001% probability for most subjects. This particular simulation was repeated 1000 times with data from all of our test subjects. Collisions were recorded less than once per 1000 runs for every test subject. The nature of the interface system coupled with the tiny possibility of any repeated error allows for a virtually error-free operation, even in restrictive environments. Furthermore, in the very rare event of a collision, resuming the original path is a very easy task.

As a final test aimed at understanding the control system's ability to correct erroneous commands should signal recognition degrade, a series of simulations were run with induced errors in the pattern recognition strategy. For example, in one case, errors were induced to reduce the pattern recognition of a 'left' movement to 80% accuracy with the principle recognition error being a 'right' movement. When this system was implemented in the same simulation shown in Figure 15, approximately 10% of the trials resulted in at least one collision. Thus, even with radically reduced recognition accuracies, the consequences of misrecognized commands still rarely result in a collision.

7. Conclusions

The goal of this paper was to demonstrate the utility of a new concept for human–machine interface in robot teleoperation. While extensive research has been performed in robot teleoperation to date, nearly all interfaces are limited in their utility outside controlled environments due to the need for operator motion, lack of portability and singular input modalities. We introduce the first system we are aware of that addresses all these issues. The system is capable of tracking tongue movement and speech to indicate operator desire by monitoring airflow in the ear canal, thus no external operator movement

is required. The system is unobtrusive, trivial to carry[3] and leaves the operator free to execute any other activity while wearing/using the system. The only sensor necessary is a simple microphone and earpiece housing which is small enough and comfortable enough to be worn in the ear indefinitely. Finally, the system allows for multiple levels of operator input in one device with only one sensor (i.e. without the addition of any sensing or processing equipment). To our knowledge, our research team is the only group that has investigated the aural cavity as a monitoring venue for machine interface, has proposed the only system whereby both tongue movement and speech may be tracked without insertion of any device in the oral cavity and has developed the only machine interface with multiple input modalities that requires only a single sensor in a single device.

Speech and tongue movement each has complementary strengths which could be synergized in a comprehensive system. We have observed tongue movement to be faster, quieter, and (in most cases) more intuitive to the user for direct device motion control when compared to speech. Aural speech capture provides the benefits of trivial calibration and no training on the part of the user, yet demands a higher level of robot autonomy.

In conclusion, new contributions in this paper include:

- an analysis of the sensitivity of human ear canal as acoustic output device;
- the design of a new sensor for monitoring airflow in the aural cavity;
- expansion of our tongue-movement control concept to include speech recognition through monitoring of airflow in the aural cavity;
- implementation of signal capture and recognition algorithms to accurately identify and classify speech through monitoring of airflow in the aural cavity;
- a strategy for detecting and classifying simple and compound tongue movements based on airflow in the aural cavity for robust hands-free robot teleoperation;
- a multi-channel decision fusion pattern classification algorithm to accurately identify tongue movements in one or both ears based on aural flow monitoring; and
- simulation results on a mobile robot system demonstrating the feasibility of hands-free teleoperation by detecting both speech and tongue movements through monitoring of airflow in the aural cavity.

3. While the current system utilizes a light (< 2.5 kg) laptop computer to host data processing and pattern recognition algorithms, a small circuit board (less than 5 cm $\times$ 5 cm) has been designed capable of hosting all processing hardware for complete man-portability.

Future work involves synergizing both the speech and tongue movement modes of interface to develop a cohesive, robust human/robot interface that will allow one to control and task robotic platforms without causing additional weight, and without the addition of any bulky or encumbering equipment. In the longer term, two distinct modes of operation with the device are envisioned whereby several devices (e.g. a power wheelchair, household appliances, stationary mechanical assist devices, etc.) may all be directed given the breadth of possibilities for control input.

At this time, functional prototypes working in real-time have been constructed for both speech (Koliousis 2007) and tongue movement (Think-A-Move Ltd 2005, 2006). Commercial applications being pursued based on this work include military scouting robots (Karlsen 2004) and rehabilitation/assist equipment, including interfaces for power wheelchair control. Although the device is not universally applicable for any situation (e.g. when force feedback is required) we believe it represents a significant contribution to human–machine interface and has the potential to lay the foundation for an entirely new generation of robot teleoperation systems.

Acknowledgements

We would like to express our gratitude to Dr Massood Tabib-Azar and Joseph Zarycki for construction of the data acquisition system, Mica Newton for speech data collection, Thomas Allen for gathering performance data on the Whegs™ II robot, Robert Karlsen and the US Army TACOM for support and military field performance/application insights, Think-A-Move, Ltd. and our test subjects at the Naval Postgraduate School and Case Western Reserve University.

References

Bashashati, A. et al. (2006). An experimental study to investigate the effects of a motion tracking electromagnetic sensor during EEG data acquisition. *IEEE Transactions on Biomedical Engineering*, **53**(3), 559–563.

Becchetti, C. and Ricotti, L. P. (1999). *Speech recognition theory and C++ implementation*. John Wiley & Sons, West Sussex, UK.

Bulbuller, G., Fargues, M. P. and Vaidyanathan, R. (2006). In-ear microphone speech data segmentation and recognition using neural networks. *Proceedings of the 2006 IEEE Workshop on Digital Signal Processing*.

Chang, S., Kim, I. and Borm, J. H. (1999). KIST teleoperation system for humanoid robot. *Proceedings of the 1999 IEEE International Conference on Intelligent Robots and Systems (IROS)*.

Chen, Y. and Newman, W. S. (2004). A human-robot interface based on electrooculography. *Proceedings of the 2004*

IEEE International Conference on Robotics and Automation (ICRA).

Cui, J. et al. (2003). A review of teleoperation system control. *Proceedings of the 2006 Florida Conference Recent Advances in Robotics (FCRAR)*.

Davis, S. and Mermelstein, P. (1980). Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *Proceedings of the 2006 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

Deller, J. R., Hansen, J. H. L. and Proakis, J. G. (2000). *Discrete-time processing of speech signals.* IEEE Press, New York, USA.

Deng, L. and O'Shaughnessy, D. (2003). *Speech processing: a dynamic and optimization-oriented approach*. Marcel Dekker, New York, USA.

Ferguson, S. and Dunlop, G. R. (2002). Grasp recognition from myoelectric signals. *Proceedings of the Australasian conference robotics and automation*.

Fukuda, O. et al. (2003). A human-assisting manipulator teleoperated by EMG signals and arm motions. *IEEE Transactions on Robotics and Automation*, 19(2), 210–222.

Galindo, C., Gonzalez, J. and Fernandez-Madrigal, J. (2006). Control architecture for human-robot integration: application to a robot wheelchair. *IEEE Transactions on Systems, Man, and Cybernetics B*, **36**(5), 1053–1067.

Goldberg, K. (2000). *The Robot in the Garden*. MIT Press, US.

Harada, T., Sato, T. and Mori, T. (2000). Human motion tracking system based on skeleton and surface integration model using pressure sensors distribution bed. *Proceedings of the 2000 Workshop Human Motion (HUMO '00)*.

Hu, C. et al. (2003). Visual gesture recognition for human-machine interface of robot teleoperation. *Proceedings of the 2003 IEEE International Conference on Intelligent Robots and Systems (IROS)*.

Karlsen, R. (2004). *Hands free teleoperation via physiological signal recognition*. US Army Tank and Automotive Command (TACOM).

Kofman, J. et al. (2005). Teleoperation of a robot manipulator using vision-based human-robot interface. *IEEE Transactions on Industrial Electronics*, **52**(5), 1206–1219.

Koliousis, D. S. (2007). *Real-time speech recognition system for robotic control applications using an in-ear microphone*. M.Sc. Thesis, Electrical Engineering, Naval Postgraduate School, Monterey, CA, USA.

Kuan, C. and Kuu, Y. (2003). Challenges in VR-based robot teleoperation. *Proceedings of the 2003 IEEE International Conference on Robotics and Automation (ICRA)*.

Kurcan, R. S. (2006). *Isolated word recognition from in-ear microphone data using hidden markov models (HMM)*. M.Sc. Thesis, Electrical Engineering, Naval Postgraduate School, Monterey, CA, USA.

Lewinger, W. A. et al. (2005). Insect-like antennal sensing for climbing and tunneling behavior in a biologically-inspired mobile robot. *Proceedings of the 2005 IEEE International Conference on Robotics and Automation (ICRA)*.

Lim, S., Lee, K. and Kwon, D. (2003). Human friendly interfaces of robotic manipulator control system for handicapped persons. *Proceedings of the 2003 IEEE/ASME International Conference on Advanced Intelligent Mechatroncis (AIM)*.

Liu, P. X. et al. (2005). Voice based robot control. *Proceedings of the 2005 IEEE International Conference on Information Acquisition*.

Marin, R. et al. (2002). Automatic speech recognition to teleoperate a robot via web. *Proceedings of the 2002 IEEE International Conference on Intelligent Robots and Systems (IROS)*.

Marin, R. et al. (2005). A multimodal interface to control a robot arm via the web: a case study on remote programming. *IEEE Transactions on Industrial Electronics*, **52**(6), 1506–1520.

Melchiorri, C. and Eusebi, A. (1996). Telemanipulation: system aspects and control issues. *Proceedings of Model Control Mechanisms in Robotics.*

Millan, J. R. et al. (2004). Noninvasive brain-actuated control of a mobile robot by human EEG. *IEEE Transactions on Biomedical Engineering*, **51**(6), 1026–1033.

Newton, M. (2005). *In-ear speech data collection*. US Naval Postgraduate School, Monterey, California, USA.

Qiang, H. and Youwei, Z. (1998). On prefiltering and endpoint detection of speech signal. *Proceedings of the 1998 International Conference on Signal Processing (ICSP '98)*.

Quinn, R. D. et al. (2002). Improved mobility through abstracted biological principles. *Proceedings of the 2002 IEEE International Conference on Intelligent Robots and Systems (IROS)*.

Rabiner, L. R. and Sambur, M. R. (1975). An algorithm for determining the endpoints of isolated utterances. *The Bell System Technical Journal*.

Raneda, A., Vilenius, J. and Huhtala, K. (2003). Teleoperation interfaces for a remote controlled hydraulic mobile machine. *Proceedings of the 2003 IEEE/ASME International Conference on Advanced Intelligent Mechatronics (AIM)*.

Richardson, A. and Rodgers, M. (2001). Vision-based semi-autonomous outdoor robot system to reduce soldier workload. *Proceedings of the 2001 SPIE, Unmanned Ground Vehicles III*.

Siegwart, R. and Goldberg, K (eds.) (2000). Robots on the Web. *IEEE Robotics and Automation Magazine*, 7(1).

Srydal, A., Bennett, R. and Greenspan, S. (1995). *Applied Speech Technology*. CRC Press, Florida.

Tanaka, K., Matsunaga, K. and Wang, H. (2005). Electroencephalogram-based control of an electric wheelchair. *IEEE Transactions on Robotics*, **21**(4), 762–766.

Tezuka, T. et al. (1994). A study on space interface for teleoperation system. *Proceedings of the 2003 IEEE International Workshop on Robot and Human Communication*.

Think-A-Move Ltd. (2005). *A hands-free human/robot interface for soldiers in the field.* Phase I SBIR final report. US Army Tank and Automotive Command (TACOM).

Think-A-Move Ltd. (2006). http://www.think-a-move.com/videodemos.html.

Urban, M. and Bajcsy, P. (2005). Fusion of voice, gesture, and human-computer interface controls for remotely operated robot. *Proceedings of the 2003 International Conference on Information Fusion*.

Vaidyanathan, R. et al. (2004). Human-machine interface for tele-robotic operation: mapping of tongue movements based on aural flow monitoring. *Proceedings of the 2004 IEEE International Conference on Intelligent Robots and Systems (IROS)*.

Vaidyanathan, R. et al. (2006). A dual mode human-machine interface for robotic control based on acoustic sensitivity of the aural cavity. *Proceedings of the 2006 IEEE/RAS-EMBS International Conference on Biomedical Robotics and Biomechatronics (BioRob)*.

Vaidyanathan, R. et al. (2007). A tongue movement communication and control concept for hands-free human-machine interfaces. *IEEE Transactions on Systems, Man, and Cybernetics*, **37**(4), 533–546.

Wang, M. and Liu, J. N. K. (2004). A novel teleoperation paradigm for human-robot interaction. *IEEE Conference on Robotics, Automation and Mechatroncis*.

Westerlund, N., M., D., and I., C. (2001). In-ear microphone equalization exploiting an active noise control. *Proceedings of Internoise* 2001.

Westerlund, N., Dahl M. and Claesson I. (2002). Speech recognition in severely disturbed environments combining earmic and active noise control. *Proceedings of Internoise* 2002.

Ying, G. S., Mitchell, C. D. and Jamieson, L. H. (1993). Endpoint detection of isolated utterances based on a modified Teager energy measurement. *Proceedings of the 1993 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*.

Yun, X. and Bachmann, E. R. (2006). Design, implementation, and experimental results of a quaternion-based kalman filter for human body motion tracking. *IEEE Transactions on Robotics and Automation*, **22**(6), 1216–1227.