A TESTBED FOR VOICE-BASED ROBOT CONTROL

Joseph Rothweiler
Sensicomm LLC
Hudson, NH 03051
+1 603 882 5688
http://www.sensicomm.com/contact.html

***Abstract*—A hardware and software configuration is being developed to support research and development in speech recognition(SR) as applied to the interactive control of devices. A research goal is to investigate new signal processing techniques and parametric representations such as those used in speech coding to improve communication effectiveness for person-to-person communications [1], [2], and to apply similar types of processing to improve the performance of machine recognition.**

SR systems are often developed and trained using prerecorded databases, which often do not capture the range of variation that can be expected in reality. For example, words spoken while reading or in casual conversation sound substantially different in live environments including noise, vibration, distractions, or emotional stress. A live system is desired to set up games or other situations which introduce excitement and competition, with the resulting changes in vocal characteristics.

Therefore, the testbed includes high quality recognition software and allows a high degree of interactivity. It is intended to support testing and improvement of speech recognition algorithms and software in a realistic environment. It should be suitable for research (improved audio feature extraction algorithms, model adaptation, etc), engineering (reliable data acquisition front ends), and human factors evaluation.

The paper discusses the motivation, chosen software and hardware approach, and some results. I plan to demonstrate the robot responding to various navigation commands using either a conventional microphone or a modified throat microphone.

I. Introduction

Speech recognition is a useful input and control mechanism for specific applications. Examples are eyes-busy, hands-busy situations. Various systems are now in widespread use, from simple isolated-word chips for games and toys to conversational systems for various commercial applications.

My research interest is in developing techniques for robust operation in challenging conditions, such as high levels of background noise and with degraded signals, and non-fluent speech. Current work is focused on robust signal analysis and acquisition techniques, which involves primarily the front-end processing of a speech recognition system.

This paper describes a test and development system comprised of speech acquisition and analysis tools, speech recognition software, and a robotic component capable of responding to spoken commands. The next section presents the speech-related issues and components under investigation. Then, I describe the system configuration and provide details of key components.

II. Speech Recognition Research Goals

While speech recognition is becoming widely used, there are still areas where significant improvement is possible. Two areas of interest are the use of sensors other than conventional acoustic microphones, and signal processing techniques to better capture the information in the speech waveform. Goals are to improve the overall recognition accuracy and to achieve robust operation even in the presence of high levels of acoustic background noise.

A typical speech recognition system contains the following components;

1) Audio signal acquisition – microphone with associated filtering, amplification, and Analog to Digital conversion.
2) Feature Extraction – Transform the audio waveform to Cepstral coefficients.
3) Phoneme recognition – Match the cepstrals against a Hidden Markov Model (HMM) of the phonemes.
4) Word recognition – Combine phonemes into a word sequence, using a word dictionary.
5) Phrase recognition – Match the word sequences against a language grammar, to determine the intended sentence.
6) Output – Send the recognition results to some device which takes action based on the spoken phrase.

The process can be viewed as transforming a large amount of data with minimal structure (audio samples) to a small amount of information (the output). Higher levels of analysis can correct lower levels; for example a misrecognized phoneme can be corrected because it results in a highly unlikely word or phrase. Conversely, the higher levels depend on having good information from the lower levels. Improvements in the signal acquisition and feature extraction phases preserve more information about the speech signal and so lead to better accuracy of the output.

So goals of this research are to investigate (1) better signal acquisition, (2) better feature extraction, and (3) collection of training databases to support them.

A. Alternative sensors for noisy environments.

It is often desirable to use speech recognition even when surrounded by high levels of background noise. An example would be in the presence of machinery. Noise can be mitigated

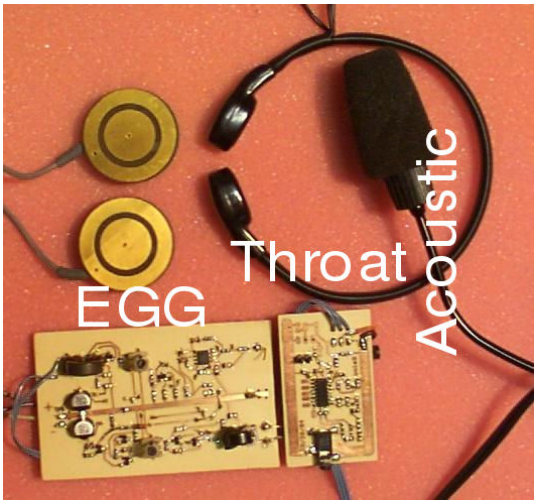

Fig. 1. Acoustic and throat microphones and EGG sensor.

Fig. 2. Testing acoustic and throat microphones with vacuum noise.

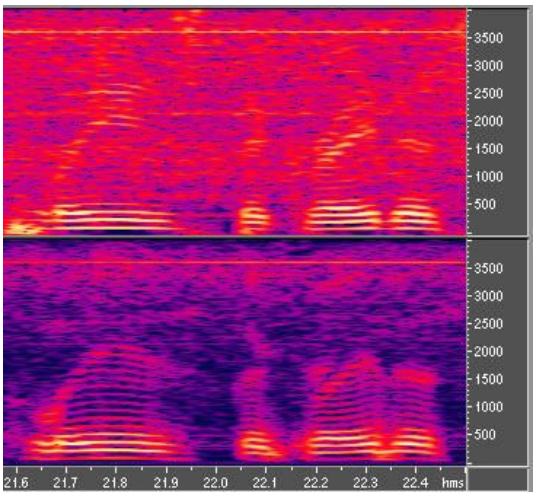

Fig. 3. Comparison of acoustic and throat microphones for vacuum noise.

by the recognition algorithm, but it's better to remove it from the audio signal as much as possible.

A very simple speech acquisition system uses a close-talking microphone. For robust operation, additional sensors can be used to acquire more information – some examples are shown in Fig. 1.

As shown in the figure, we can use (1) close-talking, noise cancelling microphones, (2) throat or bone conduction microphones (3) nonacoustic sensors such as the Electroglottograph. The basic concept is to collect as much information as possible about the sounds that are being produced by the speaker. This provides the raw information to be used in the speech recognition process. As with any deductive process, the better and more complete the raw data, the better the output decision results should be.

Noise-cancelling microphones are a well-established technology. Multiple microphones can be used to obtain both the speech signal and a measure of the acoustic background noise. This additional background information can be used by a noise-cancellation algorithm, or it can be used in later stages of the recognition process.

Throat microphones or bone-conduction microphones have been used (in military communications systems, for example), but they offer limited sound quality. With recent advances in signal processing technology, it's possible to achieve much better sound quality with these types of sensors.

To demonstrate the improvement, a recording was made while standing next to a workshop vacuum cleaner (Fig. 2). Speech was recorded using the acoustic and throat microphones of Fig. 1. Spectral analysis of the resulting signals (Fig. 3) shows that the detailed structure of the speech is far clearer in the signal from the throat mic.

A less familiar technique is the Electroglottograph or Laryngograph. This device uses resistance measurements to determine the open and closed state of the vocal folds of the larynx. So, it provides good information about whether speech is voiced (vocal folds vibrating) or unvoiced (sounds produced by turbulence in the vocal tract, like an /s/ sound). EGG devices are commercially available, and moderately large and expensive, but equivalent devices can be built very compactly.

B. Robust analysis

The audio signal is typically first transformed to Cepstral coefficients. These provide a compact representation of the speech signal, yet preserve most of the useful information. They also have good statistical properties which make them suitable for recognition.

However, they to not capture all useful information, particularly pitch and voicing information (although they can be inferred to some extent from the cepstrals). Limited testing has shown that inclusion of other non-Cepstral features can improve the accuracy of the recognition process.

In an ICASSP-99 paper[3], Atal showed that transforming a speech signal to the Cepstral feature set does not capture all of the signal characteristics that are critical to intelligibility for a human listener. Based on this observation, I performed an experiment in which the Cepstral feature set is augmented with additional features that were developed for the purpose of improving the intelligibility of vocoders in the presence of

acoustic background noise. Performance was tested using the ISIP LVCSR system, and showed a reduction of the sentence error rate from 4.8% to 3.3% when these additional features are included in both the training and decoding processes.

See [4] for more details about the signal processing used to obtain this feature set. Results were obtained using the ISIP prototyping toolkit (version 5.13)[5].

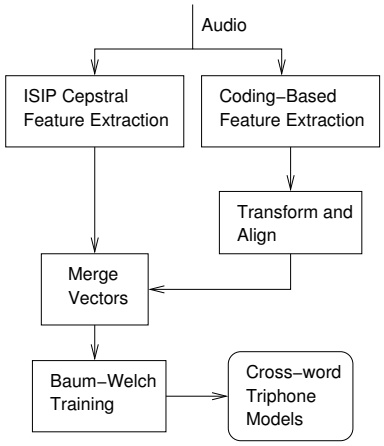

Fig. 4. The training process.

TABLE I
SENTENCE RECOGNITION PERFORMANCE

sentences	336	
	Baseline	Enhanced
Total errors	4.8%	3.3%
substitutions	0.3%	0.3%
deletions	0.0%	0.0%
insertions	4.5%	3.0%

C. Support for Training Data Collection

Development of an SR system typically involves a training phase, in which known utterances are analyzed to develop the statistical models (HMM's) of the speech sounds. This step is typically performed off-line, and requires large databases of speech samples.

Collecting sample data is a nontrivial problem. To be useful, the database needs to be representative of the speaker population, environment, conditions, etc. For example

- age, gender, accent, regional dialect.
- stress (physical and mental).
- vibration, noise.

For example, it is well known that speakers in a noisy environment tend to adapt their speech to be more intelligible by lowering their pitch and emphasizing the lower frequencies (the "Lombard" effect). Additional vocal stress also tends to change the spectral and temporal characteristics of speech.

Several well-known training databases exist such at TIMIT [6] and others available from the Linguistic Data Consortium. Obviously, if a nonstandard sensor (throat mic or EGG) is being used this canned data will be of limited value.

Therefore, the testbed includes the option to record the unprocessed audio samples while the system is in operation. These samples can then be used to refine the operation of the system in later HMM training operations.

III. System architecture

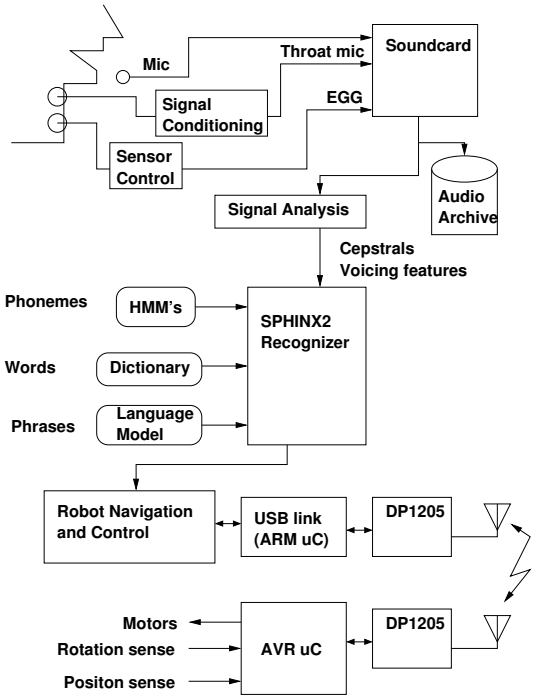

Fig. 5. Hardware and software components.

As shown in Fig. 5 the system allows for multiple sensors. Currently, a stereo input to a PC soundcard is used for data collection. This limits the operation to 2 channels. Extension to multichannel (4 or so) inputs is planned.

A Linux-based PC is used for the data collection, analysis and evaluation. All utterances can be recorded to disk while recognition is being performed. This provides a database for use in further refinements of the signal models.

The recognition software runs on the PC, and communicates with the robot via a wireless link.

The robot platform currently uses components from the Lego Mindstorms set. This set provides motors, wheels, tracks, and some simple sensors. So it provides a good starting platform. The brains of the robot is an AVR-based controller board. This was used in preference to the Lego-supplied controller because it provides a more open and flexible platform for adding additional functionality.

A. Communications

The wireless link is based on the DP-1205 module from Radiotronix. This device was chosen on the basis of cost, flexibility, ability to support multiple connections on different channels. and low power options for long battery life. It is a low-cost transmit-receive solution using FSK frequency-hopping modulation in the 915 MHz unlicensed ISM band.

The device is fully programmable for channel, bandwidth, and bit rate, so it allows tradeoffs between range and complexity. An advantage over PC-centric USB-based solutions is that the control interface is relatively simple. So a cheap uC can be used for robot control if desired.

A sample implementation of the control software is available from the Manufacturer's website. Other Open Source projects have also used this device. (http://www.raccoonrezcats.com/rfmodem.html or http://sensorscope.epfl.ch for example).

B. Recognition Software

The chosen approach uses open source components (HTK [7], ISIP [5], or Sphinx [8]) to implement the speech recognition function. These packages provide near state of the art features for feature extraction, speech model training, and online recognition. While commercial systems provide more user-friendly interfaces and many other features, the openness of the selected packages has allowed a deeper level of development, including changes at the front end (to test alternate signal processing techniques and input devices) and internally (to implement and evaluate new speaker normalization and noise mitigation techniques). It also allows the system vocabulary and operation to be tailored for the specific task.

The test platform currently uses a low-cost embedded PC board as the computing platform, with the CMU Sphinx recognition engine. A compact Finite State grammar is tailored to the specific task. Phone and word models have been designed using both the Sphinx and ISIP Hidden Markov Model (HMM) training software. Specific model sets are based on conventional microphones as well as throat microphones for noise robust operation and unobtrusiveness; other input modalities are being investigated as well.

C. Mobile Platform

The recognition outputs are communicated to a simple 2-wheel mobile robot. Platforms in use include configurations constructed of LEGO motors and other components, as well as custom hardware. Applications could include control of toy vehicles, exploring confined spaces, and similar applications.

For evaluation, the platform needs to be able to perform basic motion operations, consistent with the desired command set. The baseline platform is shown in Fig 6. The frame and drive components are from a Lego Mindstorms kit. Position sensing is obtained via rotation sensors on the wheels, the motors are driven by an Atmel AVR (board in the center), while communications is via the Radiotronix DP-1205 (small PCB to the left). Mechanical sensors on the front are standard components from the Mindstorms kit, and can be used for collision detection.

The electronics package on the mobile unit basically provide drive control and sensor monitoring. Control functionality is implemented in the PC. This partition was chosen for simplicity and to facilitate software development. Additional control could obviously be moved into the mobile unit if desired.

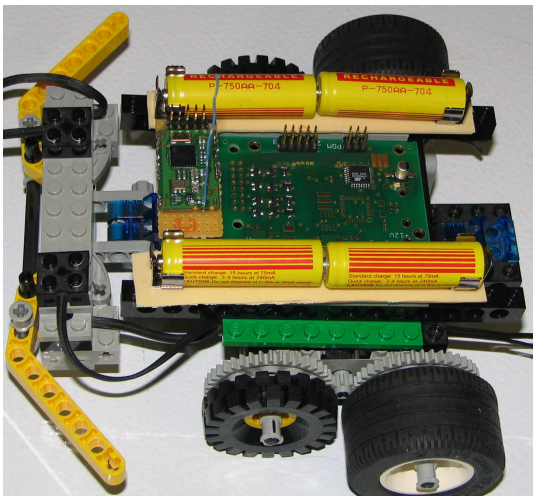

Fig. 6. The mobile robot portion of the system.

D. Vocabulary and Grammar

Recognition systems range in complexity from isolated-word systems that support simple commands ("stop","go","turn", etc) to continuous speech systems (eg, "move forward three feet") through more natural language systems ("please fetch me a beer from the refrigerator."). The current system falls in the CSR region. It is designed to accept directives that specify the motion of the platform. The common part is formed of a command, direction, and amount. So a simplified version of the grammar is

```
COMMAND DIRECTION NUMBER UNITS
```

Where COMMAND is

```
{move|turn|go|stop}
```

DIRECTION is

```
{right|left|forward|back}
```

NUMBER is a spoken value (eg, "eleven," "forty-five") and UNITS is

```
{ feet|inches|degrees}
```

This level of language was chosen because it is sufficiently complex to allow nontrivial interaction with the robot, yet the commands can be intelligently implemented using relatively simple hardware on the robot (mainly motors and wheel rotation sensors).

More complex or higher-level phrases (eg, "return to starting point") can be supported by the recognizer with suitable changes to the grammar specification. Of course, the robot would need to be provided with suitable capabilities (position sensing, navigation, etc.) to allow it to correctly respond to such a command.

IV. Software Availability

Components of the system are being made available on the Sensicomm LLC website, generally under the GNU GPL

license. Things that are currently available include

- Schematic and software to interface the DP-1205 RF link to an Atmel AVR processor.
- Schematic and software to interface the DP-1205 RF link to an NXP ARM processor. Includes an interface to a Linux PC using the LPCUSB software package.
- Codebooks and grammar files to use the Sphinx2 ASR software for control.
- Interfacing script to read audio from the sound input device run the ASR, and send the resulting commands to the robot via the wireless link.

Other items that are under active development, and will be posted in the future include

- Status monitoring software to monitor the robot's state, and report battery voltage, for example.
- ASR training scripts.

Future items in the planning stage include:

- A generic USB interface to support non-acoustic input modes (such as the EGG).

AUTHOR

For the past 6 years Joseph Rothweiler has been a consultant specializing in signal processing algorithms, software, and hardware. Projects have included sonar data acquisition, audio coding and enhancement, and signal demodulation, for clients including Sirius Satellite Radio, Agere Systems, and several other companies. Previously, he was employed by Lucent, Lockheed-Martin, and ITT in a variety of R&D applications. He has published several papers and holds 5 patents on audio and DSP techniques. He is a graduate of the University of Louisville.

REFERENCES

[1] T.S. Sun, S. Nandkumar, J. Carmody, and J. Rothweiler, "Speech enhancement using a ternary-decision based filter," in *Proc. ICASSP-95*, May 1995, pp. 820–823.

[2] J. Rothweiler, "On polynomial reduction in the computation of LSP frequencies," *IEEE Trans. on Speech and Audio Processing*, pp. 30–39, Sep 1999.

[3] Bishnu Atal, "Automatic speech recognition: A communication perspective," in *Proc. ICASSP-99*, Phoenix, AZ, May 1999, IEEE, paper 1910.

[4] J. Rothweiler, "Noise-robust 1200 bps voice coding," in *Proc. 1992 Tactical Communications Conference*, Ft. Wayne, IN, Apr 1992, pp. 65–69.

[5] Picone et al, "Institute for signal and information processing home page," URL http://www.isip.msstate.edu.

[6] William M. Fisher, George R. Doddington, and Kathleen M. Goudie-Marshall, "The DARPA speech recognition research database: Specifications and status," in *Proceedings of DARPA Workshop on Speech Recognition*, 1986, pp. 93–99.

[7] Cambridge University, "HTK homepage," http://htk.eng.cam.ac.uk.

[8] Carnegie Mellon University, "CMU Sphinx homepage," http://cmusphinx.sourceforge.net.