

# Disambiguating Speech Commands using Physical Context

Katherine M. Everitt<sup>1,3</sup>, Susumu Harada<sup>1</sup>, Jeff Bilmes<sup>2</sup>, James A. Landay<sup>1,3</sup>

DUB Group<sup>1</sup>  
Dept. of Computer Science,  
University of Washington  
Seattle WA, USA 98195  
everitt@cs.washington.edu,  
harada@cs.washington.edu

Dept. of Electrical Engineering<sup>2</sup>  
University of Washington  
Seattle WA, USA 98195  
bilmes@ee.washington.edu

Intel Research Seattle<sup>3</sup>  
1100 45<sup>th</sup> St  
Seattle, WA  
landay@cs.washington.edu

## ABSTRACT

Speech has great potential as an input mechanism for ubiquitous computing. However, the current requirements necessary for accurate speech recognition, such as a quiet environment and a well-positioned and high-quality microphone, are unreasonable to expect in a realistic setting. In a physical environment, there is often contextual information which can be sensed and used to augment the speech signal. We investigated improving speech recognition rates for an electronic personal trainer using knowledge about what equipment was in use as context. We performed an experiment with participants speaking in an instrumented apartment environment and compared the recognition rates of a larger grammar with those of a smaller grammar that is determined by the context.

## Categories and Subject Descriptors

H.5.2 [User Interfaces]: Voice I/O

## General Terms

Design, Human Factors, Measurement, Experimentation

## Keywords

Speech recognition, context, fitness, exercise

## 1. INTRODUCTION

Weiser's vision of ubiquitous computing provides the potential for a whole new way of interacting with the devices in our environment in a calm and integrated manner [18]. One of the central tenants of ubiquitous computing is that the computing devices will disappear into the background [19]. This has implications for user interface design, as it precludes requiring users to interact with computers using a traditional keyboard or mouse. Speech seems a natural choice for controlling such devices in homes or offices, as it can work from anywhere in the environment without requiring the user to know the location of

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

ICMI'07, November 12–15, 2007, Nagoya, Aichi, Japan.

Copyright 2007 ACM 978-1-59593-817-6/07/0011...\$5.00.



Figure 1: We tagged gym objects with modified RFID tags to provide context to the speech recognizer.

the device. Speech itself is commonly used and so requires little additional training to use. However, current speech recognizers are often inaccurate in non-controlled conditions due to ambient noise and improperly located microphones. Recognition error is a major problem for speech recognition, and solving it is crucial to supporting speech in this new paradigm.

Humans regularly communicate using speech in non-controlled conditions with few problems. One advantage humans have is that they can use context in addition to the sound itself to understand each other's utterances. Context helps people to disambiguate between words that sound alike and can provide information about which words and phrases are more likely to occur. Relevant context may include where you are, who is around, what activity you are doing, body language, recent conversational topics, and what objects are nearby or in use.

Previous work has explored the use of dialog history [12,14] or specialized domain knowledge [8,17] to reduce recognition errors. These approaches do not utilize the context that can be sensed in our environment [7].

Our approach is to leverage the manipulation of physical objects as additional context for the speech recognizer. As people move around an environment, they interact with devices and objects. This project seeks to use this context to determine the smallest grammar that is still relevant to the current user's activity. The assumption is that smaller grammar will lead to improved recognition rates for the voice commands.

As an exploratory step toward using context to improve recognition, we have developed ActiveSpeech, a speech-controlled application to assist with tracking health and fitness. We focus on improving command detection rates for this

electronic personal trainer by using information about of which gym equipment is in use (see Figure 1).

Keeping track of health and fitness is a difficult problem. Many people currently keep track of their workouts on paper, but this is a laborious process that requires the user to stop their workout and write down the information manually. Many people give up journaling after a short time. Lately, systems such as MySportTraining [11] have developed ways to enter information about activities and even heart rate on a PDA. However, this requires the gym user to interrupt their workout to enter data into the small device using a stylus. Storing the data electronically allows the users to more easily see summaries and graphs. However, it is disruptive to require users to focus on a small PDA device rather than their activity.

Related work has explored using sophisticated sensing devices to detect the activity. For example, Consolvo et al. [5] uses data from a wearable sensor to determine if people are walking, running, or using a small set of gym equipment, and Chang et al. [3] use accelerometer data to detect and count gym exercises people are doing from among several options. However, this latter approach requires placing accelerometers or other devices on all the objects, which involves added expense in hardware costs and battery maintenance. In addition, any errors in sensing the exact number of movements are significant enough to have impact on the usability of the application we have built. Both approaches require the user to wear sensors. Our goal is to track context with minimal sensing and maintenance. Accordingly, we use RFID tags, which are inexpensive, small, and do not require a battery.

This paper describes the design and implementation of ActiveSpeech, an electronic personal trainer that uses knowledge of what objects are in use in the gym to reduce recognition error with the speech-based assistant. We use speech recognition as the primary input to ActiveSpeech. In addition, speech supports functionality beyond simply logging, such as explaining how to carry out a particular exercise, displaying feedback, obtaining information on how the user is feeling, and displaying the user's progress. The system also provides access to data summaries that the user can use to gauge her progress.

In the remainder of this paper, we will describe the ActiveSpeech approach, how we gathered speech and object movement data, the grammars we used, the methods of changing grammar sizes and the results of switching grammars. Finally, we will discuss lessons learned from the implementation of this system and finish with related work and conclusions.

## 2. APPROACH

The ActiveSpeech system consists of four parts: RFID sensing of object use, the speech recognition engine, phrase understanding, and the personal trainer output.

### 2.1 Sensing Object Use

To determine context, we have tagged dumbbells, tension grips, resistance bands, and mats with Radio Frequency Identification (RFID) tags.

The specific tags we are using are WISPs (Wireless Identification and Sensing Platform). The WISP is a battery-free platform for sensing and computation [16]. We used the most primitive WISP, known as an  $\alpha$ -WISP [13]. It consists of two RFID chips, which

are connected to a single antenna by means of anti-parallel mercury switches (see Figure 2). The mercury switches connect only one of the RFID chips to the antenna at a time, depending on which way the WISP is oriented. This means only one tag ID will be active at any time. Thus, if we see both IDs of a WISP's two chips within a short period of time (e.g., 20 seconds), we know that the object it is attached to has been moved.

We chose to use the  $\alpha$ -WISP tags instead of a traditional RFID tag because we want to detect when the object is being manipulated, and a traditional tag would only say whether it was in range. This approach does not require the user to wear any sensing equipment. For our experiment, we used an Alien Technologies RFID reader with two antennas, one at the front and one placed to the side of the environment. The reader could also be placed in the ceiling, but at present the range on the tags means that this results in fewer reads of the  $\alpha$ -WISPs.

We expect RFID technology to improve significantly in the next several years, which will allow for the antenna to be placed in the ceiling or wall and be less disruptive to the environment. In addition, RFID tags are currently a few cents to buy, so their cost is already reasonable. A final advantage of RFID tags is that they do not require a battery, which reduces the maintenance cost of the tags. Also, the more current versions of WISPs are smaller than the  $\alpha$ -WISPs pictured, and we eventually expect smaller tags or the RFID tags to be integrated into the objects themselves.

The RFID reader antennas are connected to a dedicated machine that is installed in the ceiling of an instrumented apartment in our lab. When the RFID detects an object tag, it sends the current active object name via sockets to the machine running speech recognition. At present, the system only considers one object to be in use at any one time; extending this to multiple active object models is left to future work.

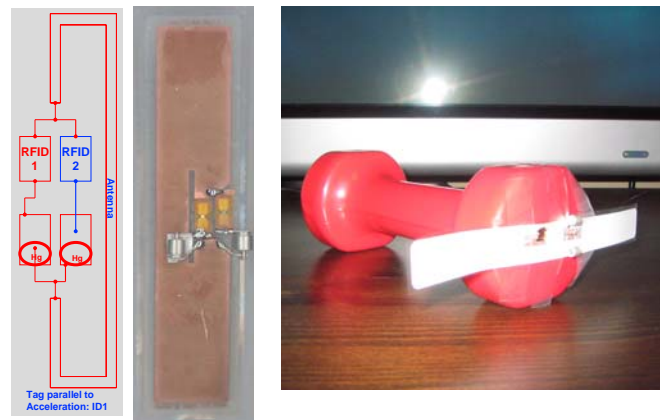


Figure 2: Left:  $\alpha$ -WISP detail [13], Right:  $\alpha$ -WISP tagged dumbbell

### 2.2 Speech Recognition

The speech recognizer we use is Microsoft Speech SDK 5.1 using Microsoft Speech API 5 (SAPI 5) [10]. The grammars are specified as command and control grammars in XML. Command and Control grammars consist of a set of potential phrases that can be recognized. Partial utterances and phrases that do not appear to match the list of phrases are ignored.

Each object has its own associated grammar that contains utterances that are expected to occur while it is in use. The grammars were manually constructed based on the types of exercise and actions that the trainer application supports. Table 1 shows example phrases from the dumbbell object.

When the object is moved, we deactivate the previous grammar and activate the new one. For example, when the dumbbell grammar is active, the phrase “Log 15 bicep curls” is active, but “Log a sit up” is not. For our experiment, we use this simple on/off grammar strategy, which turns a grammar on when the related object is manipulated, and off when a different object is detected. We leave more sophisticated grammar strategies such as weighting grammars to future work.

**Table 1. Subset of the Dumbbell Command and Control Object Grammar. This grammar consists of utterances we expect to occur when the dumbbells are in use.**

Show workout
Show my progress
I feel tired
how do I do a bicep curl
how do I do a lateral dumbbell raise
how do I do a hammer curl
how do I do an overhead dumbbell press
log <NUMBER> bicep curls
log <NUMBER> lateral dumbbell raises
...

### 2.3 Phrase Understanding & Trainer Output

Each possible verbal phrase is linked to an appropriate output in the trainer application. The current output of the trainer is a visual display that could appear on a home television or screen in a gym. Figure 3 shows examples of the ActiveSpeech output screens. ActiveSpeech can list the current workout, show how to do specific exercises, log an exercise (e.g., record the number of repetitions), acknowledge when an exercise has been logged, and show progress and trends over time. Example phrases include: “How do I do an overhead dumbbell press”, “Record 20 sit-ups”, “Show workout”, “I feel tired”, and “Show progress”. When a phrase is recognized, the trainer displays relevant information on the screen.

The trainer also includes the capability to respond via audio, and thus could easily use a Bluetooth headset and phone or other portable device when a screen is not available. However, we did not test this type of output during our experiment.



**Figure 3: Digital Personal Trainer Screenshots**

### 3. GRAMMARS

Our main hypothesis is that instead of using a large grammar that contains all possible gym and exercise utterances, we can use context to determine a smaller grammar that will constrain what is allowed to be said, and that the smaller grammar will have higher recognition rates.

Theoretically, correctly determining a smaller and more specific grammar that is a subset of the larger grammar will have a higher recognition accuracy rate because it has lower perplexity, a measure of confusability [15]. The broadest grammar available to our recognizer is the English language dictation grammar, which consists of tens of thousands of words. Knowledge of a specific domain, such as exercise, can reduce the number of potential words to hundreds, and thus we would expect the fitness grammar to do better than the broader dictation grammar. Similarly, we would expect a context-specific subset of the fitness grammar to do better than both of the larger grammars.

However, introducing context also has the potential to introduce errors. A smaller grammar will not perform well if it is too specific and used on phrases it does not contain. Contextual sensing has errors of its own, which may contribute to speech recognition errors by choosing the wrong subset grammar. Also, because of the cost and errors associated with external sensing, any improvement in speech recognition accuracy must be worth the extra cost. If the size of the grammar is cut in half but the phrases removed are acoustically very different than those that remain, it is possible that the reduced grammar size will have little effect on recognition rates. In contrast, if grammar phrases are removed that could easily be confused with the intended words, the smaller grammar could provide a significant benefit.

There is a tradeoff between the improvement in recognition rates that the reduced perplexity of a smaller grammar provides, and the potential errors that can be introduced if the context is incorrectly sensed or the relationship between the context and utterances is incorrect.

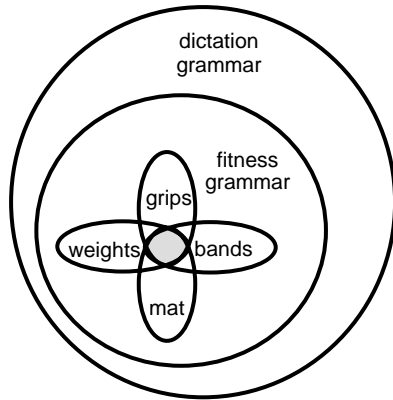


Figure 4: Venn diagram showing the relationships between the grammars.

### 3.1 Grammar Size

For our experiment, we created a series of grammars of decreasing size and increasing specificity (see Figure 4). The broadest grammar is the dictation grammar, which is designed to recognize the entire English language. The most specific grammars are the object-based grammars, which only recognize the phrases centered on the objects used in our study. We assume that when people are interacting with an object, they will say things about that object. We also assume that they will not say things about other objects. The grammar sizes we explored in this experiment (in order of decreasing size) are:

**Dictation Grammar:** This is a general grammar optimized for English dictation. It includes tens of thousands of words.

**Fitness Grammar:** This grammar is a command and control grammar which contains approximately 205 commands that are related to gym activities. It is a superset of all the phrases in the various object grammars.

**Object Grammars:** Each object grammar is a subset of the fitness grammar, and is the set of commands related to a specific object as well as some of the general informational commands (e.g., “show workout”). In our experiment, we use four object grammars: *Dumbbells*, *Resistance Bands*, *Mat*, and *Tension Grips*. Each object grammar contains about 14 commands.

As we expect smaller grammars to do better, we would expect that the object grammar would have fewer errors than the fitness grammar, and the fitness grammar would have fewer errors than the dictation grammar.

### 3.2 Switching Policy

To use a more specific object grammar we must determine which object grammar should be active at any one time. We chose to explore three different kinds of switching:

**Ideal Switching:** To test our assumptions as to the improved accuracy of a smaller grammar, we use an oracle to switch to the correct object grammar at least 1 second before the related utterance.

**Observed Timed Switching:** Observed Time Switching is based on the time the participants in the study actually picked up an object as recorded by a human observer. This switching shows the

accuracy rate if the RFID system could detect object manipulation 100% correctly.

**RFID Timed Switching:** RFID switching is based on the detection of the RFID of an object changing state within 20 seconds. For our  $\alpha$ -WISP tags, this means both IDs of the tag needed to be detected within this time period to be counted. As the  $\alpha$ -WISP tags are research prototypes, we do not expect the tags to detect object manipulation 100% of the time.

## 4. USER STUDY

We performed a user study with the ActiveSpeech personal trainer to gather data on how the objects are actually used in relation to the speech, how accurate the current RFID system is, and how the speech recognizer with different grammar sizes will perform on data from real users.

### 4.1 Task

5 participants (1 female, 4 male) between the ages of 20 and 30 participated in a one hour user study to gather speech and object use data. Two of them had accents. They were each compensated with a \$10 coffee gift certificate. We used an Andrea close-talking tethered microphone with a USB digital audio adaptor.

Participants were first asked to complete the Microsoft English Recognizer 5.1 speech training wizard, which consists of an adjustment of microphone volume, a test for microphone positioning, a dialog box for gender and age, and approximately 5-10 minutes of training. Participants then completed a practice session of 5 exercises. For the experimental task, users were asked to complete 2 sets of 20 exercises. For each exercise, they were asked to say “How do I do <exercise>”, do the exercise, and then say “Log <number of exercises> <exercise>”. They also said “show workout” or “what’s next” to display their next exercise. At regular intervals, they were prompted to say “I feel <how they felt>”.

We had three major types of commands in our experiment. *General* commands such as “show workout” are not object specific. *Information* commands were of the form “How do I do a leg press” and *Recording* commands were of the form “Log 16 leg presses”. We assumed that the Information and Recording commands would be spoken when people were interacting with the related objects.

The exercises came from five categories representing five different gym objects: 5 lb Dumbbells, 3 lb Dumbbells, Resistance Bands, Tension Grips, and a Gym Mat. The exercises were balanced to prevent fatigue and require the user to switch objects before each exercise.

Although the system can run from voice commands, it occasionally misses utterances or misrecognizes them. To remove this factor, we used a Wizard of Oz (human operated) version of the ActiveSpeech digital trainer during the study. The ActiveSpeech screens are loaded into SketchWizard [6], a tool that supports the display of images on a remote screen. During the study, the wizard displayed the appropriate screen based on the vocal commands of the participant. Object movement was recorded by the RFID system and also manually logged by the Wizard to later evaluate RFID accuracy. The speech data from the study was gathered for post-analysis.

## 4.2 Error Measurement

A common error metric for speech recognition is Word Error Rate (WER). This measures the number of incorrect words per total number of words in the correct transcript. Word Error Rate is a function of the number of insertions (extraneous sounds recognized), substitutions (incorrectly identified words) and deletions (missed words.)

$$WER = \frac{100 \times (\text{insertions} + \text{substitutions} + \text{deletions})}{\text{total \# of words in correct transcript}}$$

However, from a user's point of view, a more relevant error metric is Task Error Rate (TER). This is a measure of how often the system does not perform as expected. As our system responds to phrases, we calculate TER using the number of incorrect phrases over the number of phrases in the correct transcript. This corresponds to how often the ActiveSpeech system did not perform as expected.

A user perceives three kinds of errors; the system failing to respond to a command (ignore), the system incorrectly responding to a command (wrong), and the system incorrectly acting on a non-command sound (superfluous). Ignores correspond to deletions and the superfluous commands correspond to insertions in word error rate. However, in the Task Error Rate metric, a phrase substitution is considered to be an ignore error *and* a superfluous error, as the user has to correct the action taken by the system *and* re-issue the command. In TER, we treat a substitution as both an ignore and as a superfluous command because this is a better measure of how it impacts the user.

The Task Error Rate metric is not directly comparable to word error rate because not only must the system recognize *all* the words in the phrase to have it considered correct, what would be a single substitution error in WER corresponds to two errors in TER. However, multiple word errors within a phrase will only result in one incorrect phrase.

$$TER = \frac{100 \times (\text{ignored} + \text{superfluous})}{\text{total \# of phrases in correct transcript}}$$

The task error metric is intended to measure the overall success of the system, not just the speech recognizer itself.

Because we are using a set of phrases in the command and control grammar, we are leveraging some knowledge about dialog context. We only accept phrases that conform to the expected command structures. Without this constraint, the recognition engine we used has significant difficulty recognizing appropriate commands, even with a smaller word set.

## 5. RESULTS

Experimental trials were recorded in waveform audio format (.wav) files and run using the recognizer under various conditions. They include the dictation grammar, the fitness grammar, and various object grammars with different switching policies.

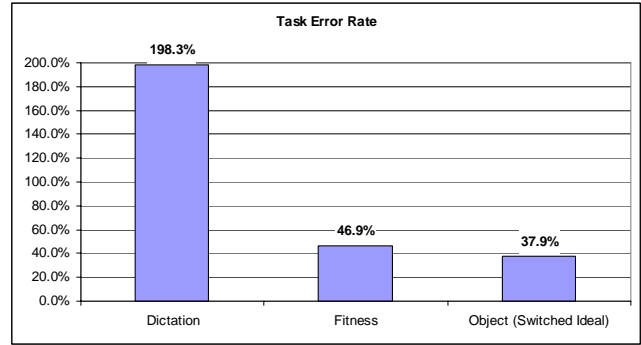


Figure 5: Task Error Rate for various grammar sizes.

### 5.1 Effect of Grammar Size

Figure 5 shows the Task Error Rate for the Dictation grammar, Fitness Grammar, and the Object grammars (with an ideal switching policy from the oracle). As expected, the Dictation grammar does quite poorly. The TER of the Dictation grammar is significantly larger than the TER of the Fitness grammar ( $p < 0.01$ ). This is the effect of the smaller grammar and also defining phrases that correspond to expected commands. The error rate for the Ideal switched object grammars is significantly smaller than the Fitness grammar error rate ( $p < 0.01$ ). This is as expected, because we expect larger grammars to have more errors than smaller grammars. Thus, using the smaller grammars results in a 19% improvement in TER over the fitness grammar.

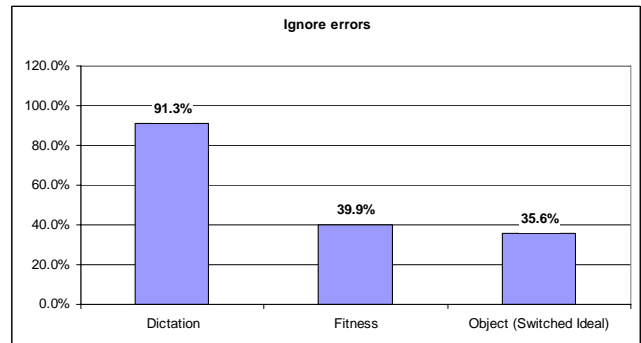
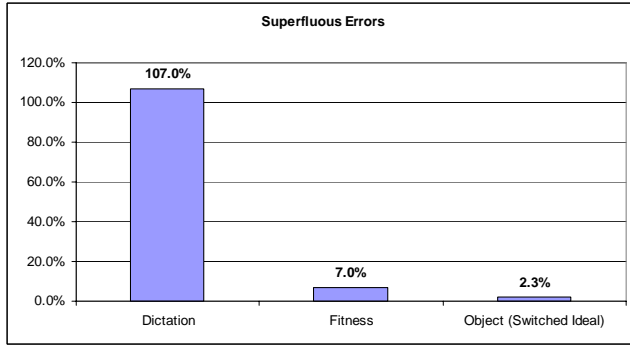


Figure 6: Ignore errors as a % of commands

Figure 6 shows ignore errors that occur, which corresponds to the number of times the system would not respond to the user the first time they uttered a command. Ignore errors cause approximately half of the dictation errors but most of the Fitness errors and almost all of the Ideal errors.



**Figure 7: Superfluous Errors as a % of commands**

Figure 7 shows the number of superfluous errors, or times non command sounds were recognized as commands. The dictation grammar had a superfluous error rate of 107%, which is a significant problem. The Fitness grammar had significantly more superfluous errors than the object grammar ( $p < 0.01$ ).

Superfluous errors are arguably more disruptive than ignore errors. Superfluous error can occur when the user is not attending to the system, and require that the user recognize what has occurred and correct it. To correct an ignore error; the user can simply re-speak the command.

These results suggest that using a smaller object grammar can have a significant advantageous effect on the number of superfluous “false positive” errors that occur with this recognizer, and this is where it is likely to be of most benefit.

## 5.2 Data Variability

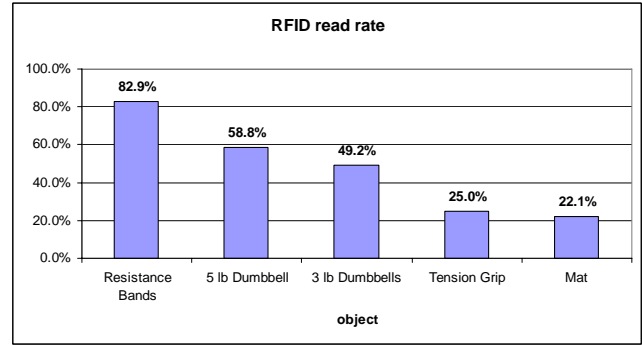
Some participants’ commands are better recognized than others. In particular, participant B had high accuracy rates, most likely because of the volume of the audio was loud and clear. Participant C was recognized at 100% in the first trial, but much lower in the second trial. Partway through the second trial, the microphone fell off participant C’s head, and afterward the volume of the audio was significantly decreased. Before the incident, the accuracy rate is 79%, and afterward 45%.

## 5.3 RFID Accuracy

Accuracy of the RFID system is important to the overall accuracy of the personal trainer. One major challenge with RFID is placing the tags on the objects in a way that will not interfere with the workout but can be read by the reader. With metal dumbbells, the metal can interfere with the readability of the tag, and the tag must be facing the reader for it to be read. For the experiment, we had one Resistance Band with a tag on each handle, two 3lb dumbbells with a tag each, one tagged 5lb dumbbell, a Tension Grip with one tag on the handle, and a mat with one tag. See Figure 1 for more detail.

Figure 8 shows the accuracy rates of the  $\alpha$ -WISPs for various objects. We found that requiring the RFID system to detect both ends of the tag within 20 seconds resulted in only about half of the object movements being detected and frequent event misses.

The Resistance Bands had the highest rate of correct detection. This is most likely because they have two WISPs, and they are moved around a lot while preparing to do the related exercises. Surprisingly, the 3 lb Dumbbells accuracy rate was lower than the



**Figure 8: RFID accuracy rates**

5 lb Dumbbell accuracy rate, even though the 3 lbs have two WISPs. This may be because of the difference in exercise motions between the 3 lb and 5 lb Dumbbells. Surprisingly, the dumbbell rates were higher than the Tension Grip rates. This goes against our expectation that the dumbbell rate would be lower because of interference from the metal. We expect that this result is related to how the Tension Grips are held, and possibly people’s hands are occluding the signal. Unsurprisingly, the Mat detection rate was lowest. This was expected because the mat was used primarily on the floor, and thus it was further from the reader antennas.

When using the RFID to switch, the average Task Error Rate was 72%, which is much higher than the ideal case. Overall, these results suggest that the type of movement can affect the detection rate of the RFID system. As the  $\alpha$ -WISP is a research prototype, it was not responsive enough to use in this application, but we expect future tags will have better detection rates.

## 6. DISCUSSION

Although theoretically switching among smaller grammars can result in a better recognition rate, there are practical issues to deploying such systems that may introduce more errors.

### 6.1 Object Use

We found during our experiment that the assumption that an object will be in use while related utterances are being said did not always hold. Our assumption was that people would always pick up the object before uttering an Information or Recording command about it. In most cases, participants did not pick up the object before they asked how to do the exercise (Information command). Recording commands were usually spoken immediately after they had put down the related object, but before they picked up the next object, so in this case the assumption did hold.

Switching object grammars based on when the wizard observed the objects in use results in a Task Error Rate of about 60%. This is significantly higher than the ideal case, and even worse than using the fitness grammar. This suggests that switching between grammars using an on/off policy is insufficient, because having the wrong grammar caused additional errors. A policy of weighting grammars based on objects would be less sensitive to this type of error. Another approach might be to differentiate the switched grammars only in the Recording commands when the context information is relevant. Preliminary analysis of the error rate for the recording commands leads us to believe that it would be closer to the ideal rate.

This issue of relating grammar and context will occur in any ubiquitous computing speech system which takes advantage of physical context. While the on/off phrase switching policy is an extreme example, any use of context may change how the system operates in different situations. This may be potentially confusing to users. If the context has such a strong effect on the speech recognizer, researchers should consider making the results of the context sensing explicit so that users can correct it. Alternatively, there could be some negotiation of grammars which could occur. If the context-specific grammar does not get the speech correct, a system could potentially consider re-evaluating the use of context and possibly using a broader grammar. More sophisticated techniques, such as increasing the probabilities of relevant grammars, might be less brittle. One question that comes up in sensing multiple objects is how long to remember that an object has been moved. One approach is to use a graduated falloff scheme where recent objects have more weight.

## 6.2 Microphone Use

The participants used a head mounted, tethered microphone. While not the ideal microphone to wear working out, it was chosen for this experiment because it is considered to provide the best quality audio for speech recognition.

The microphone works best when placed just outside the corner of the mouth. It was carefully placed there during calibration and training, but it quite naturally moved around during the exercises. At one point it fell off of one participant's head, and another got caught up in the cord and knocked it askew. Less targeted microphones such as array microphones have more difficulty with ambient noise. More work is necessary to determine how various microphones perform in practice, and whether context can help some of the problems encountered with ambient noise from array microphones.

Given that the volume of the audio and accuracy of the recording decreased after a participant knocked the microphone off his head, it might also make sense to automatically detect this sort of event and either require the user to recalibrate or reset the noise baseline expectations in the recognizer.

## 6.3 Sources of Error

There were several causes of both ignore errors and superfluous errors. Superfluous errors were commonly caused by out-of-vocabulary (OOV) words and phrases that were not in the grammar. Also, because people were exercising, they would naturally be breathing more heavily than normal and the microphone would move as they exercised. This caused non-speech sounds to be picked up by the microphone.

Microphone movement also caused problems with audio being too quiet, causing ignore errors. Ignore errors were also caused by the speaker pausing in the middle of the phrase, resulting in a mis-recognition.

Participants often re-phrased the commands, which is difficult for the system to recognize, but we still considered this to be a task error. In general, the speech recognizer had more trouble with the phrases when they did not precisely match the commands, but often recognized them as the intended command despite slight differences. Whenever possible, alternate phrasings should be included in a command and control grammar. However, it is impossible to determine a priori everything people might say.

## 7. RELATED WORK

Related work in this area falls into several categories: Physical context for speech, dialog context for speech, and using other context such as domain knowledge and combining modalities.

### 7.1 Physical Context

The most relevant related work is systems that use physical context to change a speech recognizer. CASIS [9] is a Context-Aware Speech Interface System that uses context such as used command history, seat occupancy, speaker location, brightness, and sound level to help speech recognition for the control of devices in a conference room environment. The external context sensors were simulated by an assistant. They found significant benefit from rejecting actions that were illegal, such as turning on a light when it was already on.

A Context Sensitive Natural Language Modality for the Intelligent Room [4] uses a linguistic data structure they call a recognition forest to associate context and the related grammars. We use a similar on/off grammar strategy to the one they proposed to relate grammars to context. Although the recognition forest paper presents the design goals for their grammar forest for their specific intelligent room, they do not evaluate the efficacy of the system.

While these projects use semantic knowledge of device state (a light cannot be turned on if it is already on), we use a set of heuristics that involve knowledge of what objects are associated with what utterances but do not require semantic domain knowledge of which operations are legal given the current state. Our associated approach supports a model whereby one can easily add and remove tagged objects with associated grammar phrases to an environment. Previous work requires a specific intelligent environment with management of multiple potential sensors, semantic understanding of sensors' meaning, and training with regards to expected utterances.

### 7.2 Dialog Context

Another approach to improve speech recognition is to use dialog context. While dialog context does seem like a good approach to speech recognition in the exercise domain because of set workouts etc, we did not explore it as our first step. Because the study required the participants to follow the provided workout exactly, using dialog to determine audio would have been trivial. However, other work has explored dialog context in more unbounded settings.

Project Listen [1] is a Speech Tutoring system that listens as students read pre-assigned sentences. It uses knowledge about the expected utterances and models of oral reading to determine when to correct or intervene. Using extended knowledge about the expected sounds or the reader is another way to use context to help the speech recognizer do the right thing.

Answers Anywhere [2] is a mobile answering system that supports multiple platforms and uses the current request, request history, application state and output modality to help answer the user's questions. It uses an agent network to interpret free form input and translate it into machine readable statements.

Paek and Chickering [12] use predicative models learned from a population of mobile users to predict the next command based on location and time of day.

### 7.3 Other Context

Another common approach is to use domain knowledge to get better accuracy rates. Using smaller grammars that are more tailored to the domain can result in better recognition rates. Glass et al [8] describe the development of JUPITER, which uses a corpus of utterances from people asking weather related questions to improve query recognition for a telephone based weather information system.

### 8. FUTURE WORK

This project has only scratched the surface of using context to help speech recognition. At the core is the question of whether the decreased grammar size from context is worth the extra sensing costs and recognition errors due to sensing errors.

An obvious next step is to continue to explore different grammar sizes and how they affect the recognition rates of various real world situations. Another direction is to combine the physical context with dialog context to have improved overall rates. Especially if the user has a workout plan they are working from, this could be very beneficial to solving the overall problem of recognition in this specific domain.

Also, more work can be done in looking at the error rates of various sensing devices. The on/off grammar switching is a simple approach that is not robust to context sensing errors. Modifying object probabilities rather than simply turning grammars on and off and dealing with multiple objects at once is an interesting direction for future work.

A further direction for future work is to combine the speech recognition with more sophisticated sensing such as [3] so that the speech and sensing can mutually disambiguate each other, resulting in improved Task Error Rates. We could combine speech recognition with sophisticated sensing and counting to support mutual disambiguation of sensing and phrases.

### 9. CONCLUSIONS

We have developed a digital personal trainer, ActiveSpeech, to reduce the effort needed to log and view exercise usage information. The trainer uses a combination of speech and knowledge about which objects are in use to assist the user. We carried out a preliminary experiment to explore the feasibility of combining speech and context in a physical setting, and have identified several challenges as well as promising directions for future work.

### 10. ACKNOWLEDGMENTS

Thanks to the participants for giving their time. Also thanks to Jon Malkin and Amar Subramanya for answering our questions, and Jonathan Lester for his advice.

### 11. REFERENCES

1. G.S.Aist and J. Mostow. 1997. Adapting Human Tutorial Interventions for a Reading Tutor that Listens: Using Continuous Speech Recognition in Interactive Educational Multimedia. In CALL '97 Conference on Multimedia, England.
2. Answers Anywhere [www.iAnywhere.com](http://www.iAnywhere.com)
3. Chang, K. Chen, M., Canny, J., Towards Balanced Exercise Programs: Tracking Free-weight Exercises, UbiComp 2007.
4. Coen, M.; Weisman, L; Thomas, K; Groh, M. A Context Sensitive Natural Language Modality for the Intelligent Room. In Proc. MANSE'99. Dublin, Ireland. 1999.
5. Consolvo, S., Paulos, E., Smith, I. Mobile Persuasion for Everyday Behavior Change. *Mobile Persuasion*. Stanford Captology Media. 2007
6. Richard C. Davis, T. Scott Saponas, Michael Shilman, and James A. Landay. SketchWizard: Wizard of Oz Prototyping of Pen-based User Interfaces, In submission to UIST 2007
7. A. K. Dey, Understanding and Using Context, *Personal and Ubiquitous Computing Journal*, Volume 5(1), pp 4-7, 2001.
8. J. Glass, T. J. Hazen and I. L. Hetherington, "Realtime Telephone-based Speech Recognition in the Jupiter Domain," in Proc. ICASSP '99, Phoenix, pp. 61-64, Mar. 1999.
9. Leong et al. CASIS: a context-aware speech interface system. In Proc. Intelligent user interfaces 2005
10. Microsoft Speech Server <http://microsoft.com/speech>
11. MySportTraining 3.97 by VidaOne, Inc [http://www.pocketgear.com/software\\_detail.asp?id=630](http://www.pocketgear.com/software_detail.asp?id=630)
12. T. Paek & D. Chickering. Improving command and control speech recognition on mobile devices: Using predictive user models for language modeling. *User Modeling and User-Adapted Interaction, Special Issue on Statistical and Probabilistic Methods for User Modeling*, 2007, 17(1-2): 93-117.
13. Matthai Philipose, Joshua R. Smith, Bing Jiang, Kishore Sundara-Rajan, Alexander Mamishev, Sumit Roy. Battery-Free Wireless Identification and Sensing, *IEEE Pervasive Computing*, Vol. 4, No. 1, pp. 37-45, January-March 2005.
14. R. Porzel and I. Gurevych, Contextual Coherence in Natural Language Processing, *CONTEXT 2003*, LNAI 2680, Springer-Verlag, pp 272--285, 2003.
15. Lawrence Rabiner, Biing-Hwang Juang, *Fundamentals of Speech Recognition*, Prentice-Hall, Inc., NJ, 1993
16. Joshua R. Smith, Alanson Sample, Pauline Powledge, Alexander Mamishev, Sumit Roy. A wirelessly powered platform for sensing and computation. In Proc. UbiComp 2006
17. C. Wai, R. Pieraccini, and H. M. Meng, A Dynamic Semantic Model for Re-scoring Recognition Hypotheses, In *Proceedings of ICASSP2001*, pp 589-592, 2001.
18. Mark Weiser, John Seely Brown "The Coming Age of Calm Technology", In *Beyond Calculation: The Next Fifty Years of Computing*, Peter J. Denning and Robert M. Metcalfe, New York, Springer-Verlag 1997.
19. Mark Weiser, "The Computer for the Twenty-First Century," *Scientific American*, pp. 94-10, September 1991