INVESTIGATING THE SUITABILITY OF SPEECH RECOGNITION FOR TRAINING SYSTEMS

Robert Rejent
Catherine Meyn

Logicon, Inc.
Tactical and Training Systems
San Diego, California

— ABSTRACT —

Speech recognition can promote enhanced training procedures and reduce operating costs in training systems. For this reason, the incorporation of speech technology into training systems is becoming more prevalent. Many users of these training systems, however, are unaware of the technical capabilities of speech recognition, and therefore have unrealistic expectations which affect trainer acceptability. To prevent this, it is important for the user and developer of any training system to probe the question: "Is speech recognition appropriate for this training application?" Logicon has integrated speech technology into air traffic control training systems for nearly 15 years. In transitioning from research and development systems to fully operational trainers, experience has been gained regarding this fundamental question. This paper identifies the issues associated with determining the suitability of speech recognition for a particular training application.

— BACKGROUND —

Speech recognition has been identified as a means for improving training capabilities and reducing operating costs. Because of this, requirements for state-of-the-art trainers are including speech technology to meet training needs. Those responsible for defining these requirements, however, are often unfamiliar with the capabilities of speech recognition. In addition, many potential users of training systems have unrealistic expectations of how the technology should perform. Much of their exposure to speech technology has been through television and movies. Robots like C-3P0 and R2D2 in "Star Wars" which understand all human speech might sell movie tickets, but do not represent the current capabilities of speech technology. The rapid advances in speech recognition devices also confuse the issue, as requirements analysts and system developers try to incorporate constantly changing capabilities.

Understanding the capabilities and limitations of speech technology permit the user to determine whether speech recognition is appropriate for a particular training application. This is most appropriate when the requirements of the training system are identified. Collaboration between the user and developer is essential during this phase. The user and developer must determine whether the incorporation of speech recognition into the trainer can support the training objectives given the current state of speech technology.

— COMMON SPEECH RECOGNITION — ASSUMPTIONS

Although it is not necessary for the user to understand speech technology in an engineering sense, there should be some understanding of how it works in general.*This is necessary to avoid misconceptions and unrealistic expectations. For example, it has been common in our experience for the user to assume that the more commands defined for the recognition system, the better the chances of a command being recognized. In fact, the larger and more complex the phraseology, the more difficult it is for the recognition system to work effectively. As it happens, this is true for the student as well, so constraining the training phraseology to the minimum number of commands required to meet the training objectives meets the dual goal of improving recognition and reducing student training requirements.

Another assumption which is commonly made is that a command which is misrecognized is a failure on the part of the system and therefore a "bad" thing. It is important to remember that a speech recognition system is designed to recognize. It is also important to remember that a speech recognition system does not know of any words outside of the training phraseology. This means that the system will always identify a match with what was spoken. While these truths constrain the technology somewhat, if properly understood they can provide significant benefits for reinforcing training. In fact, poor recognition which is related to improper use of the training phraseology or speaker inconsistency due to stress or other factors may serve as a valuable measure for the instructor to use to identify problem areas in student performance.

** Logicon is primarily involved with speaker dependent, continuous speech recognition, which permits the most robust and natural speaking style. For this reason, the focus will be on this recognition type.*

— SUITABILITY ISSUES —

Given a basic understanding of speech recognition, the issues regarding the suitability of speech recognition for a specific training application can be addressed. Questions and concerns analyzed at this point should determine whether speech recognition can support the training objectives.

Vocal Characteristics

How important are voiced commands during training? In the kind of training systems we develop, speech recognition is used to simulate human interaction. If voiced commands do not play an important role during real world operations, some other type of input device may be more suitable.

If voiced commands are necessary in the training environment, the next question regards how these commands are spoken during normal operations.

Consistency.

Current speech recognition technology requires vocal consistency. During speaker-dependent training, the recognition device establishes how each person speaks the vocabulary. Drastic deviations from this pronunciation will produce poor recognition. In the Shore Based Radar ATC Training Systems (SATS), trainees may change their pronunciation of unfamiliar words during the course of their training. For example, the word "par" has been pronounced as one syllable when beginning training. As students become more familiar with the phraseology, they change their pronunciation to "p-a-r", spoken as single letters. In order to use this new pronunciation with consistent results, the trainee must re-establish a baseline pronunciation for "p-a-r" in an offline mode before continuing with the realtime training exercise. When developing a curriculum for trainees, the user should be aware that the trainee must be taught the phraseology prior to using it to permit the successful incorporation of speech recognition.

Speech recognition is very robust in terms of speech rate. Most systems can support speech rates from half to double the rate at which the speaker trained. A speaker can speak rapidly or slowly with very good recognition, provided that consistency within the command is maintained. Wide ranges in speech rate violate the requirement for consistency. The phrase "how do you hear me" can be spoken as isolated words "how-do-you-hear-me" or as a rapid stream of syllables "ha-da-ya-hear-me". As long as the speech rate is consistent for a given speaker within the defined constraints, recognition performance will be reliable.

Is vocal consistency required of a successful trainee? With a knowledgeable instructor to monitor and correct poor speech habits, speech recognition will reinforce consistency in speech, since erratic speech will result in poor speech recognition, thereby producing inconsistent system responses. For training applications requiring consistent speech patterns of trainees, limitations of the speech technology actually promote good training. For training applications where vocal consistency is not required and is not feasible in a real world setting, speech recognition in its current state would not be a suitable input device.

Pausing.

Pausing is a natural aspect of human speech. Humans pause to separate ideas within an utterance and to think about ensuing phrases. In a training environment, this characteristic of human speech is more prevalent than in conversational speech, since trainees are in the process of learning while they are speaking. Therefore, any speech recognizer integrated into a training system must perform well with regard to speaker pausing. Some recognizers use pauses to determine the end of utterances. These types of recognizers are not optimal for trainers, since trainees may be cut off in the middle of a spoken command.

Pausing is different from verbal place holding. A trainee may hesitate before a heading by drawing out the word "heading": "fly headinggggggg .....123". This is not recognizable. A true pause is silence after a word or word group; it is not the extension of the last word spoken. Speech recognizers can accommodate true pauses; they cannot tolerate verbal place holding. In environments where verbal place holding is counter to good operational performance (such as in air traffic control), this limitation is a valuable training tool.

Tone and Prosodics.

Tone is an important element of speech in determining the recognizability of spoken utterances. Monotone speech is reliably recognized by current speech technology, but is certainly not a requirement for good recognition. Natural inflections are supported by speech recognition, and actually enhance consistent performance: a person who speaks naturally is normally very consistent in speech patterns and pronunciations. However, extreme variations in tone due to stress or excitement are not reliably recognized. Users with requirements for stressful training applications where vocal utterances can have extreme variations in tone should consider this constraint when determining the applicability of speech recognition. Naturally if vocal control during stress is one of the training objectives, this constraint can work in the instructor's favor. Note that changes in speaker tone due to nasal congestion related to allergies or colds is well accommodated by available recognizers. Speech devices are surprisingly robust in recognizing speech which, to human perception, has been drastically altered due to nasal impairment.

Training Environment

The training environment is an important area to investigate when determining whether speech recognition is appropriate within a trainer: Is training performed in an environment which can support speech recognition? Topics to evaluate include the interface between the trainee and the training system, the amount of noise in the training area, and realtime considerations.

Trainee Interface.

How the trainee communicates with the training system through vocal commands is of prime importance when determining whether speech recognition can support the training objectives. In order to incorporate speech recognition into a trainer, an interface must exist which reliably transmits voiced data to the speech device. The most common interface is a microphone. For training applications which normally include a microphone such as air traffic control or pilot training, the mobility of the trainee needs to be considered. If the trainee is required to move around the training area freely, some type of remote microphone may be required. Speech devices which allow a remote microphone interface are limited, which may effect the type of speech recognizer needed.

Background Noise.

The noise level in the training environment can impact the decision to incorporate speech recognition in the trainer. For training systems subject to varying levels of background noise, speech recognition may be more difficult than if the environment was subject to constant, low levels of noise. Since microphones transmit all sound, spoken commands as well as background noise may be present. Noise-cancelling microphones are designed to minimize background noise, so microphone selection is crucial for noisy environments. Noise-filled areas can also add to speaker stress during training operations. High stress levels may impact trainee vocal consistency, which impacts recognition. Speech recognizers differ in terms of their capability to handle varying levels of background noise and speaker stress. When determining whether speech recognition is appropriate for a training environment with these conditions, actual demonstrations are a good idea. The various recognizers being considered should be tested in the real world setting to find out whether they can support the training objectives. On the ASATS/SATS system, a recording of the noise generated during ATC operations on a carrier was used to test the robustness of the ITT voice board. In the TOTS tower trainer, it was found that excessive movement within the trainer cabs negatively affected recognition accuracy. This was due to variations in background noise between the student positions, which were open to the visual screen and the air conditioning units, and the instructor area, which was more protected. Installing glass between the student positions and the visual screen improved this situation.

Realtime Considerations.

Realtime training operations require rapid translation of spoken commands to system responses. Long delays between the trainee input and the system response can greatly impact training effectiveness. Speech recognizers differ in their capability to process recognition results in realtime. Recognizers which do operate in realtime provide results continuously with insignificant delays. Non-realtime recognizers generally buffer results for some amount of time, thus delaying the trainee-perceived system response. Although realtime training applications do not eliminate the feasibility of speech recognition, they do limit the types of recognizers under consideration.

Training Phraseology

Applicability.

Another area of analysis is the training phraseology. The training phraseology includes all possible spoken words and word combinations allowable in the training environment. Current speech technology only allows a finite number of words and phrases for a particular application. It is important, therefore, for the user to determine: Does the real world environment for which this trainer is being developed utilize a standardized, identifiable phraseology? If a standardized phraseology does not exist or cannot be developed without decreasing training effectiveness, then speech recognition cannot support training. For training systems which do incorporate a specific phraseology, the phraseology should be identified to further analyze the suitability of speech recognition. This is a critical communication point between the user and the developer. The right combination of commands will achieve high recognition and support training. The wrong combination may support training but poor recognition can make the system unusable. The sooner these tradeoffs are identified and discussed, the better and more usable the trainer will be.

Identification.

The user at this point must identify all permissible voiced commands in the training environment. This identification should include:

- rules regarding how and when each phrase is spoken. For example, all phrases which can only occur by themselves should be identified as such.
- system responses for all phrases.
- special case words. For example, the word "correction" may be spoken within each phrase in order to correct trainee mistakes. Rules regarding these types of words must identified.

Recognition requirements should also be identified at this point: What recognition accuracy is required of the speech device? As has been demonstrated, poor recognition for a particular speaker is not necessarily a bad thing, since it can flag problems in student performance. However, the accuracy for the system should be high when the student is speaking properly. Frequently the user assumes that a high recognition accuracy percentage is the sole discriminator for system performance. In fact, how a system handles a misrecognized phrase is equally important. Misrecognition of certain phrases may be unimportant if these phrases have no system response; misrecognition of other phrases may seriously decrease the effectiveness of the trainer. Recognition accuracy is really only relevant in terms of system performance. The issue of how to quantify system performance is significant, but cannot be addressed here.

Analysis.

The identified phraseology must then be analyzed to determine the feasibility of incorporating speech recognition into the trainer: Can this phraseology be reliably implemented given the speech devices available and the required recognition accuracy? This is really a question for the developer to explore, but the user is also involved, since the outcome will determine whether speech recognition can be implemented. The developer should analyze such things as:

- phraseology size: Given the specified phraseology, the developer must determine whether it will fall within the size limitations of available speech recognizers.
- recognition analysis: The phraseology should be reviewed in terms of optimizing recognition performance. Areas looked at include word competitions, poorly recognized words, and potential word combinations to improve recognition.

The results of this analysis may prove that further phraseology refinement is required. The user must decide whether certain phrases can be modified or eliminated without impacting the trainability of the system. Collaboration between the user and developer during phraseology refinement is crucial in maintaining the effectiveness of the trainer while overcoming any limitations of the speech technology. For training systems which utilize phraseologies that are not alterable and exceed the capabilities of the available devices, speech recognition is not a viable training technology.

— COSTS AND BENEFITS —

When evaluating the use of any technology within a training system, the cost of that technology must be considered. In addition, the user should be aware of benefits inherent in using speech recognition.

Outlays

Speech recognition devices range from $100 to $10,000 in cost. Lower-end devices have a very limited vocabulary capability (100 words maximum), and are generally not robust in terms of handling wide variations in speaker volume and rate. Higher-end models can handle large vocabularies (some over 5000 word limits), and perform well across a wide range of speakers. The requirements of the training application determine the type of speech recognizer needed. Selecting a device which cannot support the training objectives in order to save on device expenses will result in greater expenses during the life of the trainer, since the trainer will not be able to reliably instruct trainees.

The unique requirements of speech technology generate development expenses in addition to the general development costs of the trainer. The training phraseology must be implemented in a way that the recognizer can understand. This may range from 10 labor hours to several thousand depending on the requirements of the speech device utilized. The phraseology must also be optimized in terms of recognition performance across a wide range of speakers.

Speech recognition can incur costs over the life of the trainer. As with any hardware device, the recognizer may need occasional maintenance to replace electronic components. Changes in phraseology over time can also add to lifecycle costs, depending on the time required to translate the phraseology changes into the recognizer's format.

Benefits

If the initial costs of speech recognition are so high, why should it ever be incorporated in a trainer? There are several reasons:

1. Syntax-based speech recognition requires a fixed phraseology and consistent delivery. When a firm knowledge and use of phraseology is critical, this reinforces an essential need.

2. The direct interface between the student and the computer ensures that the system response is precisely in accordance with what was recognized with no assumptions to cause negative training. The computer can also determine the validity of a command with greater accuracy than most humans.

3. Under computer control, a more complex training scenario can be presented without the requirement for significant additional personnel resources to provide necessary interactions.

The result is better and more structured training, using fewer resources.

— CONCLUSION —

Speech recognition has delivered on the promise it demonstrated in the 1970s. Understanding its capabilities and limitations is vital to the successful incorporation of speech recognition into training applications. By evaluating the importance of voiced commands, the training phraseology, and the training environment during the initial requirements phase, the user can make an educated decision on using speech technology within a training application.

— *ABOUT THE AUTHORS* —

Robert Rejent is a computer scientist at Logicon in the speech technology area. For the past 4 years, he has integrated speech recognition into training simulators for the Advanced Shipboard ATC Training Systems (ASATS), the Shore Based Radar ATC Training Systems (SATS), and the Tower Operator Training System (TOTS). He holds a master's degree in Industrial and Systems Engineering from the Ohio State University.

Catherine Meyn manages the Speech Technology Section at Logicon. Since 1978, she has worked in the areas of speech recognition/generation, instructor operator stations and performance measurement systems for air traffic control and flight training systems.