# USING SPEECH RECOGNITION IN REALTIME TRAINING SYSTEMS: FINDING THE BALANCE

*Lynne M. Pusanik*
*Robert A. Rejent*
*Tactical and Training Systems Division*
*Logicon, Inc.*
*San Diego, California*

## – ABSTRACT –

Many new technologies are being implemented to enhance training system effectiveness, particularly in phraseology applications such as air traffic control (ATC). One of these technologies, speech recognition, provides both unique challenges to the developer and unique benefits to the user of the system. This paper shows that requirements for good speech recognition and for good overall training are not mutually exclusive. Specific issues will be addressed, including the development of an appropriate training phraseology as well as common concerns of the user. Guidelines for attaining acceptable recognition accuracy will be provided along with some different methods for quantifying accuracy. Examples of air traffic control phraseology are used in this presentation with specific references to the Shore Based Radar ATC Training System (SATS), the Advanced Shipboard ATC Training System (ASATS), and the Tower Operator Training System (TOTS).

## – BACKGROUND –

In the past, training systems which used speech input extensively relied heavily on humans to interpret and implement voice generated commands. This type of training system is still in use for many training applications. For instance, in the ATC training systems, 'bug operators' listen to commands which students speak, and input the corresponding actions on keyboards. Thus, the keyboard input, rather than the voice input, is what actually drives the training system.

With the development of sophisticated speech technologies, training systems can be driven directly by a multitude of complex voice commands. An important issue with the advent of speech technology is the ability to utilize speech in training systems while maintaining or improving the overall quality of the training system. This paper addresses this issue by showing how a speech system, consisting of a large vocabulary (over 200 unique words) and continuous speech, can be developed to improve and enhance a training system.

## – TRAINING SYSTEM – REQUIREMENTS

When developing a training system, the capability of the trainer to simulate a real-world environment is of prime importance to the user and the developer. Due to the relatively recent development of speech technology, the user may doubt the effectiveness of a trainer utilizing speech recognition. In order to be effective, the trainer must realistically represent the environment which it simulates, as well as perform reliably and accurately.

### Realism

The technology of incorporating speech in the simulated environment should be as transparent to the trainee as possible. If some aspect of the voice recognition system causes the trainee to circumvent realistic procedures, then the effectiveness of the trainer is degraded. By its nature, automatic speech recognition technology is less flexible than human recognition; only a subset of human speech is able to be recognized due to technological limitations. Although this presents a restriction to speech patterns allowed in the training environment, this actually enhances the realtime effectiveness of the trainer. By limiting acceptable training phraseology to a specialized vocabulary spoken in the real world, the trainee learns to use standardized phraseology to implement specific procedures. This specialized vocabulary must be able to accurately represent all aspects of the simulated environment, as well as meet the restrictions of the speech system.

### Reliability/Accuracy

The development of a restricted vocabulary can promote the reliability and accuracy of the training system. Total system reliability is a function of recognition accuracy, which is the ability of each phrase in the vocabulary to be recognized consistently. The required accuracy level should reflect the simulated environment. Perfect recognition accuracy is not usually realistic, since real-world conditions are not always perfect. Since accuracy is a function of the phraseology, the design of the phraseology is crucial in developing a reliable training system. The phraseology must be developed with the idea of generating optimal recognition accuracy while satisfying the requirements of the realtime system.

# – SPEECH SYSTEM – REQUIREMENTS

Integrating speech recognition into a training system presents challenges to the developer due to the unique requirements of speech technology. One of these requirements is the development of a very specific training phraseology. In addition, for a speaker-dependent, template-based speech system, voice templates for each vocabulary word must be generated. Additional requirements include developing methods to ensure consistent voice recognition and system response, both in an offline and realtime setting.

## Identifying the Phraseology

The training phraseology should not only satisfy the instructional needs of the user, but must also reside within the constraints of the existing speech technology. The current driving force is the limitation of the system firmware in terms of vocabulary size and syntax complexity. The voice recognition firmware can only handle a finite number of vocabulary words and phrase combinations, with these numbers varying for different manufacturers of speech system firmware. A more subtle restriction is the selection of phrase combinations to ensure reliable phrase recognition. Consideration of these constraints is vital to the development of the training phraseology.

### Firmware Constraints

The speech recognizer integrated within both the SATS and TOTS is the ITT Model 1280 Voice Recognizer Synthesizer (VRS) board. In developing the phraseology, it is important to analyze its structure in terms of the firmware constraints. The structure of a phraseology is expressed in a syntax which defines the relationship of each of its words. For the VRS 1280, the syntax is made up of 3 elements: nodes, templates, and connections. Nodes contain templates, which represent vocabulary words. Nodes connect to each other, thus defining phrases in the vocabulary. For example, the phrase "perform landing checks" from the SATS is implemented in the syntax as shown in Figure 1.

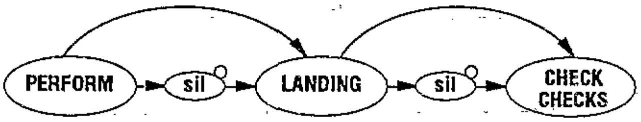

*Figure 1. Syntax Node Structure*

In this diagram, "perform", "landing", "check", and "checks" are all templates contained within separate nodes. These nodes connect such that "perform landing checks" is a valid phrase, whereas "perform checks" is not valid. Note the nodes containing "sil" templates. These indicate self-looping silence nodes, allowing the insertion of pauses between words within a phrase.

The first constraint of the syntax structure is the number of unique templates contained within the syntax. For the VRS 1280 board the limit is 500. This puts a measurable limit on the size of the vocabulary during its development. The number of occurrences of all templates within a syntax (referred to as the number of operational templates for the VRS 1280) is limited to 2000, so a preliminary syntax representation is necessary during development in order to analyze this constraint. Other limitations include the number of nodes and connections, which can also be analyzed from the preliminary syntax diagram. Since these constraints are firmware-dependent, they drive the design of the training phraseology in terms of vocabulary size and number of distinct phrases.

### Phraseology Development

The development of the phraseology begins by identifying all phrases spoken in the training environment. Corresponding system responses should also be clarified. When the user and developer have compiled an exhaustive list of training vocabulary items, the preliminary syntax is analyzed to determine whether the number of unique templates is within the acceptable range for the speech system. If the acceptable limits are not met, then this unrestricted vocabulary is too large for the voice system. An evaluation must be made at this point: Are all elements in the identified phraseology essential to training. Some points to consider in this evaluation are:

- Limit the use of phrases which have no system response: For example, many phrases in the SATS and TOTS are advisory only, and do not effect the performance of the simulation algorithms.
- Eliminate phrases with identical system responses: If 2 phrases perform the same function, it may be possible to remove the phrase used less frequently.
- Eliminate vocabulary items spoken infrequently in the training environment.

If these areas can be addressed without impact to the training effectiveness of the simulator, then a more restricted phraseology is designed by the user and developer. This process of design and evaluation is continued until the phraseology is within the technical constraints.

At this point, the limited phraseology should be analyzed in terms of recognition accuracy, to ensure that the vocabulary can be recognized in a reliable manner. Since speech recognition involves matching some speaker input to a specific list of vocabulary choices, increased recognition results from limiting choices at any point in the syntax. For example:

- Concatenate short words which are often slurred together: Short words when spoken together are generally not spoken very distinctly. The phrase "how do you hear me" is not usually spoken as discrete words, but as one long phrase. By concatenating this phrase into the word group "how-do-you-hear-me", the recognition accuracy is increased by training as a group of words as opposed to distinct, separate words.
- Identify phrases which would never begin a transmission in the training environment: Reducing options at the start of the syntax diagram provides great benefits in terms of recognition accuracy.

- Analyze word competitions: By reviewing the syntax structure, certain word competitions become apparent which may cause recognition difficulty. For example, in the TOTS, the phrases "proceed to runway 28 right" and "proceed on runway 28 right" have different functional requirements, but differ only by the prepositions "to" and "on". After generating a list of word competitions, the developer and user can pinpoint potential recognition problems, and determine whether alternative phraseology could be used which would not impact training effectiveness.

By addressing these points in developing a training phraseology, a restricted vocabulary may be developed which meets the training needs of the user, as well as the technical requirements of the speech technology.

## Generating Templates

For speaker-dependent, template-based speech systems, voice templates are required before a speaker may exercise realtime recognition. A template is a digital representation of the sound of a particular utterance and must exist for each vocabulary item in the phraseology. Trainees must generate templates in an offline mode before entering the realtime simulation.

### Script Development

Before the template generation process can be utilized, the developer must design template generation scripts. Template generation scripts are lists of phrases containing all words in the identified vocabulary, and are used to create speech templates. These scripts are designed to create templates that can be reliably recognized in realtime for all speakers. When designing scripts for a template-based speech recognizer, there are several guidelines to follow in order to generate good templates. Many of these were derived using the VRS 1280, but can be generalized for any template-based speech system:

- Generate templates with a variety of beginning and ending sounds: Each template must occur in the systems' scripts at least 3 times, so at least 2 different surrounding sounds should bracket each template. The template "speed" might be represented in the scripts as follows:

  "take approach speed"

  "take speed fifty"

  "say speed"

This generates a template for "speed" with a variety of word boundaries.

- Keep phrase lengths to an optimal range: Template generation phrases should generally be between 3 and 5 words in length. Short phrases are sometimes not long enough in duration for the recognition system to pattern match. Long phrases may leave the speaker breathless or promote mid-phrase pausing, as well as being difficult to read.
- Create script phrases using the restricted phraseology, where possible. Scripts should be created by analyzing realtime occurrences of each word and designing prompts which represent these occurrences. However, certain phrases are either too long in duration, or do not provide a variety of word boundaries for specific templates. In these cases, shortened or modified phrases can be used to generate acceptable templates.
- Keep script lengths to a minimum: This reduces speaker fatigue as the template generation session continues.

### Enrollment Methodology

Once the offline training scripts and syntaxes have been created, the system is available for template generation. The process of generating an individual's set of templates is referred to as enrollment and/or template training.

Template generation can take many forms but it may be viewed in three parts: template enrollment, template training, and template verification. Generally, enrollment specifies the creation of a unique template, training refers to the progressive modification of an existing template, and verification checks the recognition of a template without any modification.

The generation process may consist of modifying a set of 'seed' templates, or of recording each individual's speech pattern for a particular word, or, as in the case of SATS/ASATS and TOTS systems, of a combination of these two procedures. Seed templates are an average of a particular template across many speakers of the same gender. The generation process is divided into several specific phases: training digits, training carrier words, enrolling vocabulary words, and context training. Each of these phases has the potential for enrollment, training and verification.

In digits and carrier word training, male or female seed templates are used as a starting point. The digits or carrier words are prompted to the user from a script file. During training, the templates are updated by averaging the digital results of the spoken word with the template stored on the voice board. In this way, each template is gradually modified to represent the individual speaker's speech patterns. As training progresses, each template is either promoted to verification, demoted to enrollment or kept at training based on the recognition accuracy of each trainee. The verification mode is used to verify the successful training of each word. If a word is not recognized correctly in verification, then the word is flagged for more training or for enrollment.

The carrier words are words which are used to bracket each vocabulary word when the vocabulary words are being enrolled. In the SATS/ASATS and TOTS systems the carrier words used are: "say", "speak", "again", and "please". "Say" and "speak" are used to precede words, while "again" and "please" follow words. As can be seen, "speak" and "say" have a hard and a fricative ending, respectively. In the same way, "please" and "again" have a hard and a fricative beginning. Carrier word training is needed because if the endpoints of the carrier words are not clear, then the endpoints of the word bracketed by the carrier words will also be unclear.

In enrollment each word which is used in the phraseology (except digits or carrier words) must be spoken with a set of carrier words. By knowing the endpoints of the carrier words, the digitized representation of each vocabulary word may be generated by the VRS 1280. In the SATS/ASATS and TOTS systems, once an initial digitized pattern for each word has been generated, each word is trained by bracketing the word with the opposing carrier words. Depending on the recognition results in the training stage, the word is either promoted to the next step of training, verified, or demoted to enrollment again.

In context training, the words are spoken in realtime phrases, so that the templates are updated in the context in which they are used. This helps to bias words to particular speech patterns. For example, the word "for" when spoken in isolation sounds like the word "four". However, if "for" is used in a sentence: 'What are we having for dinner?'; the word"for" may sound more like "fur" or "fir" than "four". The offline training is aimed at putting words into combinations which will promote recognition of the identified training phraseology.

The time required for each individual to complete the template generation process varies with the individual's clarity and rate of speech and with the size and complexity of the training phraseology. Figure 2 shows individual times from some recent sessions in the Carrier Air Traffic Control Center (CATCC) training environment.

| Subject # --> | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| Digits | 20 | 6 | 10 | 6 | 14 | 10 | 11 | 18 | 12 | 11 |
| Carriers | 5 | 7 | 6 | 8 | 10 | 5 | 7 | 4 | 6 | 6 |
| Enrollment | 97 | 62 | 90 | 62 | 72 | 77 | 100 | 98 | 85 | 67 |
| Context | 52 | 73 | 41 | 40 | 77 | 66 | 43 | 57 | 49 | 58 |
| **TOTAL** | **174** | **148** | **147** | **116** | **173** | **158** | **161** | **177** | **152** | **142** |

*Figure 2. Training Times in Minutes*

## Enrollment Benefits

In addition to preparing the individual's templates for use in the realtime environment, the offline template generation procedure also prepares the individual for the realtime simulator. For an individual who is unfamiliar with speech recognition, the offline environment introduces the user to the system in a series of steps.

One of the most important things an individual learns in template generation is to speak consistently. If words are not spoken consistently in template generation, then the individual cannot progress from one step to the next. In order to complete template generation, consistency is required. Otherwise, templates will repeatedly be demoted forcing re-enrollment or re-training. Consistent speech is not only required by the speech recognition system, it is also necessary for many training applications.

## Vocal Consistency

Vocal consistency is not limited to what is spoken (the standardized phraseology), but also encompasses how something is spoken. In a real-world setting it is not enough for an air traffic controller to speak the correct phraseology in an emergency situation. If the controller panics and speaks too quickly or if the pitch of his or her voice changes significantly, then there is a very good chance that a pilot may not understand what was spoken. This constraint lends itself well to speech recognition technology. If a student is not consistent, he or she will not be recognized. It is interesting to note that the experienced controllers used during tests were extremely consistent vocally and received high recognition scores. In fact, the more experienced a controller was in high traffic field operations, the better his or her recognition was. In environments where verbal instructions are crucial, it is necessary to teach not only what to say, but also how to say the instructions in any situation. Certain guidelines and procedures have been developed to ensure consistency of speech in both offline and realtime environments.

### Offline Speech

Learning to speak consistently begins in the offline environment. In order to ensure that the same information is provided to each user, a list of guidelines has been developed which is presented to each user prior to initiating template generation. These guidelines are also reviewed by the instructor before beginning template generation. The guidelines which have been developed are:

1. ADJUST MICROPHONE POSITION CAREFULLY.
2. USE THE PTT CORRECTLY.
   - Disable the PTT to pause when sneezing, conversing, or taking a break.
3. READ BEFORE YOU SPEAK AND SAY PROMPTED PHRASES ONLY ONCE.
4. SPEAK CONSISTENTLY AND NATURALLY AS THOUGH TO A PERSON STANDING A FOOT AWAY.
   -Don't yell, whisper, or try to add unnatural emphasis.
   -Pronounce words as you would in normal operations.
   -Speak at a normal pace and not too slowly.
   -Connect words within an enrollment or training phrase.

The instructor must learn to listen closely for infractions of these guidelines. Some key areas which the instructor must be aware of include continuity of speech, rate of speech, and 'unnatural' speech.

Continuous speech is very important in training since it defines the endpoints of words. The instructor must listen for users who separate words with pauses. In order to generate optimal templates, all phrases should be spoken as sentences, without breaks or pauses.

The instructor must also listen for users who appear to be speaking very slowly or very quickly. Each user should be speaking at a rate which is normal for him or her and which will be used in the realtime environment. Often, users who have difficulty reading orally speak much slower when reading prompts than they would when speaking to another person. The instructor should encourage these people to read the prompt silently before speaking, and then to speak the prompt. Others will speak rapidly either because they believe that they normally speak rapidly, or because they believe it will be necessary to speak rapidly in the realtime environment. The instructor must listen to these people closely. If the speakers are still speaking clearly, then that rate is not too fast for them; if, however, their speech becomes muddled, then the instructor must advise the trainees to speak slower.

Some speakers may speak awkwardly or may try and emphasize certain words during training. The instructor must encourage a 'natural' speech pattern. Unnatural emphasis on a word will be averaged into the development of the template for that word and should be avoided.

In all cases, the instructor must correct the user as soon as possible. Usually most of these problems can be detected and corrected before the completion of carrier word training. The instructor must remember that high recognition accuracy in realtime is dependent on a successful template generation session.

### Realtime Speech

In realtime, as in offline, there are guidelines for the user to help improve recognition accuracy. A list has been developed which is reviewed by the users prior to commencing a realtime exercise. These guidelines are:

1. ADJUST MICROPHONE POSITION CAREFULLY.
2. USE THE PTT CORRECTLY.

   Disable the PTT to pause when sneezing, conversing, or taking a break.
3. SPEAK NORMALLY.

   -Speak at your natural pace.

   -Speak naturally using your normal inflections.
4. USE CORRECT PHRASEOLOGY.

Two areas of speech are addressed in these guidelines: how something is spoken, and what is spoken. The system response provides feedback to the user in both of these areas. As in offline, how something is spoken includes rate of speech, tonality and pitch of speech, and clarity of speech. The SATS/ASATS and TOTS training environments, utilizing the VRS 1280 voice boards, is fairly flexible in all three areas. However, limits do exist. The VRS 1280 boards will be able to recognize a word if it is spoken at a minimum of half the recorded template rate, to a maximum of twice the recorded template rate. This range allows a wide variation in speech rate. Unfortunately, as rate changes, pitch, tonality and clarity also change. Slight changes in any one of these would not present recognition problems; however, significant changes may produce poor recognition.

The second area of speech which is addressed in the list of guidelines is what is spoken. In fact, the most important item of these guidelines is speaking within the correct training phraseology. Good speech recognition cannot be attained without following this guideline. The nature of speech technology forces the user to speak only in a well-defined, standardized phraseology, just as he or she should speak in the real environment.

In the SATS/ASATS and TOTS systems there are three possible system responses to a spoken command: the correct simulated action and pilot response, an incorrect simulated action and/or pilot response, and a "say again" pilot response. These system responses provide the feedback to the user concerning his or her speech recognition. The assumption is made that a correct action and pilot response represent a correctly recognized phrase. An incorrect action or responseindicates that a misrecognition of the spoken phrase occurred. A pilot response of "say again" indicates an incompletely recognized phrase.

When a user speaks a command not found in the training phraseology, it is possible that the correct system response will occur, but not likely. Most often, a command will be recognized which does not generate the appropriate response or which is incomplete, thus generating a "say again" from the pilot. In either of these cases, the user will see and hear immediately that the action he or she intended is not taking place. The user's first consideration at this point should be to review what was spoken. In this manner, adherence to a standardized phraseology is enforced. If the user uttered the command in accordance with the restricted phraseology, then the user should consider how the command was spoken. Thus, consistent speech patterns are enforced.

Testing has proved it helpful for observers to monitor system responses for users. During a realtime exercise, the observers would primarily make notations regarding the user's speech. However, before the user's can be fully judged on their speech, an objective and quantitative method to evaluate recognition must be developed.

## – EVALUATION –

To evaluate a system which utilizes speech technology, the performance of the trainer must meet or exceed the training system requirements in the areas of realism, reliability, and accuracy. The issue of a realistic training environment is dealt with in the development of the training phraseology. The requirements for reliability and accuracy, however, rely more on the implementation of the phraseology. This can be measured by the recognition performance. Recognition performance can be defined in two ways: a restrictive, word-oriented scoring method, and a system response approach. Both of these methods are useful in the evaluation of an integrated speech trainer. However, the second also provides a measure of training effectiveness. High accuracy does not guarantee effective training, nor does less than perfect recognition preclude effective training. Only by measuring accuracy as it relates to system performance can a judgment of training effectiveness be made.

## Restrictive Scoring

Restrictive scoring is used by the developer to optimize the template generation scripts and realtime syntax. This type of scoring identifies any words or phrases which may pose recognition difficulties. The process begins by performing recognition testing on a pool of subjects. These subjects should be similar in background to prospective trainees for whom the system is being developed. A representative set of phrases are chosen from the phraseology for the test. This set should test all syntax paths, but not take more than an hour to speak, due to speaker fatigue. If the set of phrases is very extensive, then the test procedure should be divided into two sessions. After the speakers have completed the recognition test, the results are analyzed for areas of misrecognition. These problems may dictate changes in the template generation scripts or in the realtime syntax, depending on the nature of the misrecognition. An example which demonstrates this process are the commands:

"mother is in a starboard turn"
"mother is in a port turn".

If the second phrase is commonly misrecognized as "mother is in a starboard turn", then the generation of the template "port" is not optimal. The scripts are analyzed to determine how "port" is generated, and changes made to improve the word boundaries around "port". In addition, both "starboard" and "port" should exist in the same nodes in the template generation syntaxes, since they both compete in realtime. If the second phrase is commonly misrecognized as a variety of other phrases, then the template generation scripts are reviewed in terms of the template "mother". The realtime syntax can also be improved by making the words "in" and "a" optional. This increases recognition accuracy, since speakers normally do not articulate short prepositions and articles. By following this kind of analysis for all areas of misrecognition, the template generation scripts and realtime syntaxes are optimized. Phraseology testing is an iterative process, and this process should be repeated until the system provides reliable training.

## Performance Scoring

Performance scoring is used to evaluate system effectiveness. This scoring method is performed at a higher level than restrictive scoring; its main purpose is evaluating the trainer's ability to simulate a real-world environment. Recognition testing is performed in the same way as in restrictive scoring. However, results are gathered in terms of system performance rather than word misrecognition.

An example of a phrase which fails restrictive scoring but passes performance scoring is the previous example, "mother is in a port turn". In SATS, these two commands are considered advisory commands, so the recognition of "starboard" for "port" does not effect the simulation. Also, deletion of the preposition "in" or the article "a" in the results may fail a strictly word-based analysis, yet the system performance is unaffected. Performance scoring is used to evaluate the overall training system; not its components. The performance score of a training system is indication of the effectiveness of the trainer in simulating the real world.

# - CONCLUSION -

The SATS and TOTS have successfully integrated speech recognition into Air Traffic Control simulators. SATS, for example, has implemented a vocabulary of 250 unique words (1838 operational templates) in its Radar Air Traffic Control Facility (RATCF). This training system has an average system accuracy of 95% per speaker. The TOTS system has a vocabulary of 233 unique words and 1763 operational templates. In preliminary testing this system has an average performance score of 99% per speaker. Performance testing and restrictive testing are continuing for both systems.

The inclusion of speech recognition in these systems results in a realistic trainer, as well as enforcing the use of a standardized phraseology. In addition, good speech habits result in positive feedback, while poor speech habits, such as incoherent or inconsistent speech, result in negative feedback. Thus, the trainer promotes better speech habits from students using the system. By applying the procedures outlined in this paper, any training system which is driven by voiced commands can operate accurately and precisely in a realistic environment.

### *ABOUT THE AUTHORS*

**Lynne Pusanik** works for Logicon, specializing in speech technology. Currently she is involved with speech recognition and voice enrollment/ training for the Advanced Shipboard ATC Training Systems (ASATS), Shore Based Radar ATC Training Systems (SATS) and for the Tower Operator Training System (TOTS). She graduated from Texas A&M University with a Master's degree in bioengineering and a bachelor's degree in mechanical engineering.

**Robert Rejent** is a computer scientist at Logicon in the speech technology area. He has integrated voice recognition into training simulators for the Advanced Shipboard ATC Training Systems (ASATS) and the Shore Based Radar ATC Training Systems (SATS). He holds a master's degree in Industrial and Systems Engineering from Ohio State University.

### *ACKNOWLEDGEMENTS*

The authors would like to acknowledge the Communications Division of ITT Defense for their invaluable support. Special thanks to Jeff Nichols and James Holden for sharing their expertise.