REAL TIME NOISE TOLERANT SPEECH RECOGNITION

R. Bradley Cope and David Kotick
Naval Air Warfare Center
Orlando FL

ABSTRACT

Navy Air Traffic Control (ATC) trainers use simulations of airspace traffic to improve skills and to provide realistic, safe training environments. Real-time continuous speech recognition is used to interface directly with ATC training devices, thus eliminating the need for human pseudo-pilots or sim-pilots. With the expertise gained in constructing ATC training devices utilizing speech recognition technology, the Navy has applied speech recognition to a real time Navy Carrier Air Traffic Control (CATC) application called ISIS (Integrated Shipboard Information System). ISIS is located in three CATC areas, Air Operations, Primary Flight Control, and Carrier Controlled Approach (CCA), all with varying noise levels. Operationally, the ISIS speech recognition system must maintain near perfect recognition accuracy while being operated in a high noise environment.

In general, speech recognition syntax is constructed with a balance between the phraseology and discernment between similar sounding words. Furthermore, syntax development for a particular users/group phraseology demands an intimate understanding of the users/group functional requirements. Beyond what is required for basic recognition, the ISIS recognizer uses an additional noise rejection algorithm which models one's speech, compares the speech input with the model, and competes the result with real-time templates whereby sounds other than the intended speech can be detected. Using this noise rejection model in a different manner, the ISIS speech recognizer provides an out-of -phraseology capability that filters incorrect words and phrases. Exhaustive testing of a particular syntax with and without a noisy background is necessary to establish a desirable noise rejection threshold. Finally, this report discusses the hardware, software, syntax development, noise studies, out of phraseology algorithm, results of the fielded system, and implementation of the ISIS speaker dependent speech recognition system.

ABOUT THE AUTHORS

Brad Cope is an Electronics Engineer with the Simulation & Models Division of the Naval Air Warfare Center Training Systems Division. His responsibilities are primarily the applied research, coding , and integration of ISIS speech recognition. He holds a BSEE from Temple University (5/83) and a Masters of Engineering in Engineering Science from Penn State University (5/91). He has worked at NAWCTSD since September 1993. Brad previously worked applied research programs at the Naval Air Development Center in Warminster PA.

David Kotick is an Electronics Engineer with the Simulation & Models Division of the Naval Air Warfare Center Training Systems Division. His principal responsibilities include combat system embedded training design and the integration, evaluation and design of microcomputer based board-level products. He has received both a BSE and MSE in Electrical Engineering from the University of Central Florida. He has worked in the area of real-time multi-processing at NAWCTSD since 1983. Recent work includes speech system syntax tool development, speech post-processing algorithms and word-scoring methods in simulation systems incorporating speech technology.

REAL TIME NOISE TOLERANT SPEECH RECOGNITION

R. Bradley Cope and David Kotick
Naval Air Warfare Center
Orlando FL

INTRODUCTION

Speech recognition has been implemented into the Integrated Shipboard Information System (ISIS) to meet a fleet requirement for speech input. This is a natural mode of entry for this application because carrier air traffic control information is currently passed via sound powered phone and the OJ314 intercommunication system. Aircraft information such as fuel states and altitudes is flight critical. Speech recognition improves the overall accuracy of this information by reducing the number of times the information is passed before it is used for making critical decisions.

The most significant difficulty facing the implementation of speech recognition into carrier Air Traffic Control (ATC) areas was the severe noise environment. The speech technologies group at NAWCTSD has a combined 31 years of experience in speech recognition research and development. Experience with speech recognizers in Air Traffic Control Trainers led to the selection of the ITT speech recognition system because of its robustness to noise, high reliability, and familiarity for a proof-of-concept system. Inherently the ITT system offered the lowest risk to the program because of its widespread use in Naval air traffic control trainers. Importantly, the ITT template generation system (TGS) is common to the training school thereby simplifying operational use.

The ISIS system is a joint effort between NAWCTSD Orlando FL and NAWCAD Lakehurst NJ. NAWCTSD developed the speech recognition system, software and the syntax design for the ATC phraseology. NAWCAD developed the status board display system and database application. Figure 1 shows the ISIS concept.

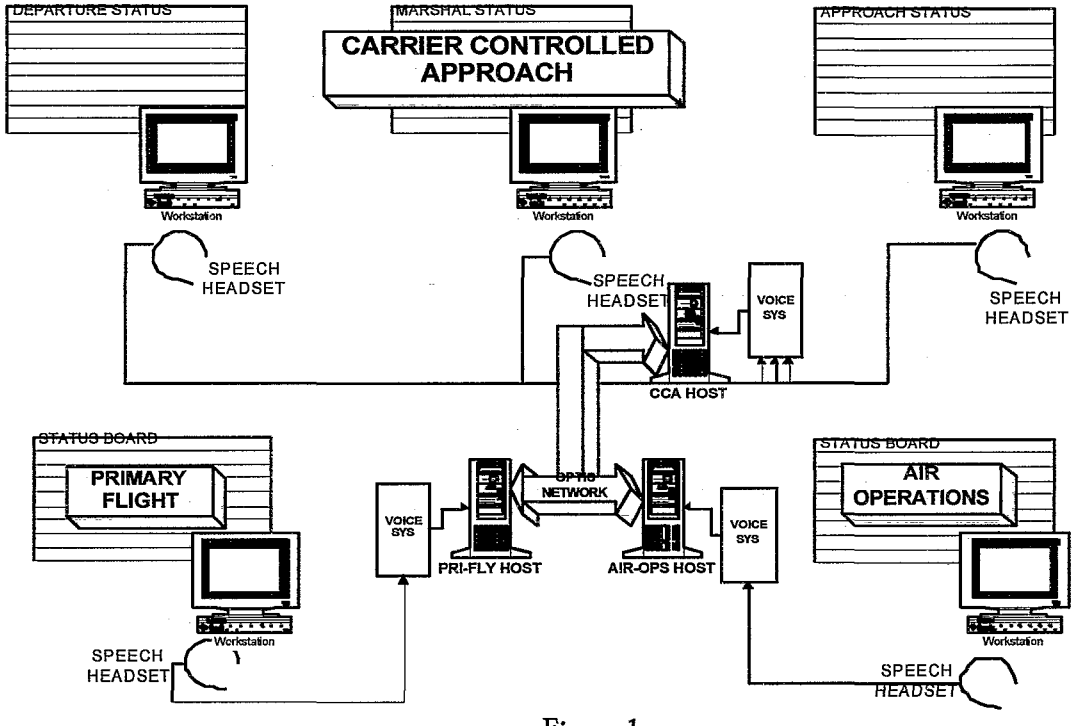

Figure 1

DISCUSSION

The ISIS speech recognition system was developed based on the following requirements:
1) Speaker Dependent, 2) Continuous, 3) Optimal vocabulary at 200 words, and very importantly 4) Noise tolerance. These requirements were necessary to achieve a real-time noise-tolerant system whereby a Naval Air Traffic Control Status board writer could freely vocalize critical air traffic control (ATC) information to be simultaneously displayed in three flight control areas (see figure 1). These areas are Carrier Controlled Approach, Primary Flight Control(the tower), and Air Operations. Additionally, ISIS air traffic control information is displayed in all of the pilot's ready rooms and the landing signal officer's platform.

Speaker dependency was chosen to meet performance requirements, specifically, the recognizer had to recognize in perceived real time and complete its output within 300 milliseconds from the end of a valid phrase. Speaker dependent systems require an enrollment session to capture a particular user's speech characteristics. Enrollment for the ISIS application is relatively short and does not need to be repeated once a good set of characteristics has been obtained.

Continuous speech recognition allows the user to speak freely within the established phraseology for this application. The alternative is non-continuous speech where a pause must be inserted between words. Continuous speech systems typically have a smaller vocabulary, but do allow the user to speak quite rapidly without any pause between words. The ISIS vocabulary is 118 words and is sufficiently small enough to qualify use of the continuous form of speech recognition. Intuitively, the ability of the recognizer to recognize quickly and accurately is proportional to the size of the applied vocabulary. Additionally, the complexity of the syntax (the manner in which the vocabulary is assembled) is related to the accuracy of the system. This is called perplexity. Overall, the ISIS system has a small vocabulary and a moderate to large perplexity, both contributing to the accuracy of the recognizer.

Integrating The Real-Time Speech System

Integration of the speech recognizer with the ISIS host computer was accomplished by way of serial communication. A protocol was established to encode valid ISIS phraseology into a serial packets from the recognizer to the ISIS host. This approach worked well for the proof of concept system because the two systems were independently developed. The protocol served as the interface specification used to communicate modified system requirements and/or syntax changes between the two development teams. A production version of the speech recognition system is in the works and it is expected to be fully integrated with the ISIS host. This system will provide more capabilities than the proof-of-concept system.

The ISIS speaker dependent speech recognizer operates in real-time. The response from the recognizer is nearly instantaneous and eliminates the possibility of data latency from the time the operator completes his command to the occurrence of the data in the appropriate status board field. Further, three recognizers operate simultaneously in a single host and in real time in the carrier controlled approach area. The speech recognition system uses an event driven scheme where the host process places individual recognizer responses into a queue. Each of the recognizers queue their respective independent tasks when a valid command has been completed. The host process then parses the results and packetizes the information based on the interface protocol. All of this occurs in less than 150 milliseconds on average, a factor of two better than the system requirement (see figure 2).

The recognizer hardware consists of a PC full width recognizer board. All recognition processing is performed locally using a digital signal processor (DSP), a Motorola 68000 CPU, and a dynamic time warp (DTW) chip. The DSP performs a fast fourier transform on the incoming time domain speech input. The 68000 subsequently performs a melcepstral transform before passing recognized reference templates to the custom DTW chip. The DTW attempts to match the incoming templates with the set of stored syntax templates. Each word is scored prior to output to the post processor, which is contained in the host PC. Much of this paper is focused on the post processing developed by NAWCTSD for this application.

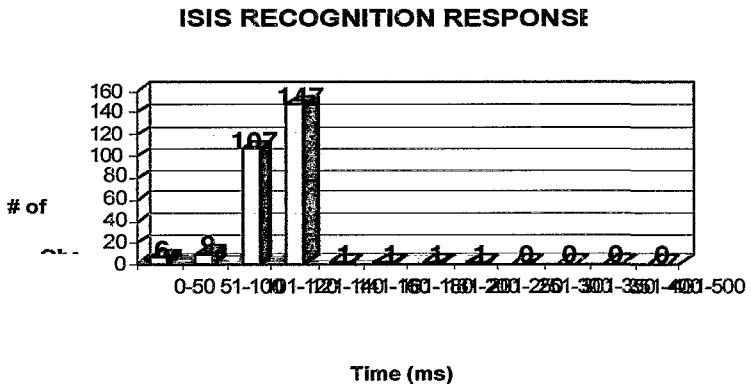

Figure 2

NOISE TOLERANT SPEECH RECOGNITION

Noise Study

Extensive shipboard noise testing was performed during Carrier Qualifications (CQ's) on the CVN Carl Vinson prior to development of the ISIS speech recognition system [1]. The carrier air traffic control area has approximately 64 dB of background noise and 96 dB transient noise from the aircraft arresting systems. Specifically, the EA6-B traps at 96 dB, the retraction of the arresting gear generates 78 dB, attention whistle at 74 dB, and closing of the compartment door is 71 dB. Although loud to the human ear, the trap noise arresting gear) does not interfere with speech recognition because the energy is primarily in frequency bands lower than the speech recognition lowest quantized band(see figure 3) [2]. A series of bells are sounded to indicate the current hour and announcements. Some of the bell frequency contents happens to be in band with natural speech. There is a spike of energy at just below 2 Khz (see figure 4) which can interfere with speech recognition if the PTT is depressed during the bell. Generally, ambient noise in CCA and Air Operation is primarily generated by ATC personnel. Primary Flight Control experiences noise from the surveillance radars as well as conversation from ATC personnel. Considerable ambient white noise raises the noise floor of the recognizer and is adjusted by initiating a noise calibration and using an automatically adjusted automatic gain control.

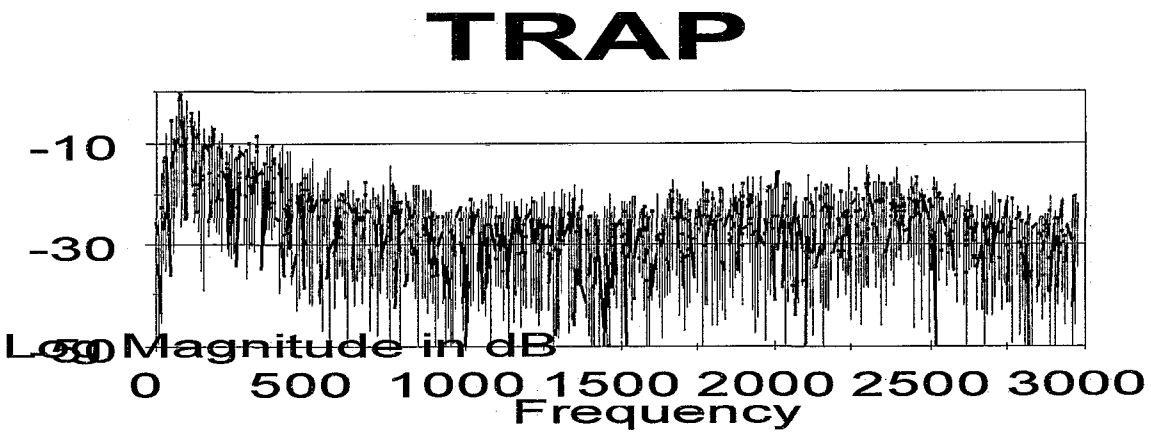

Figure 3

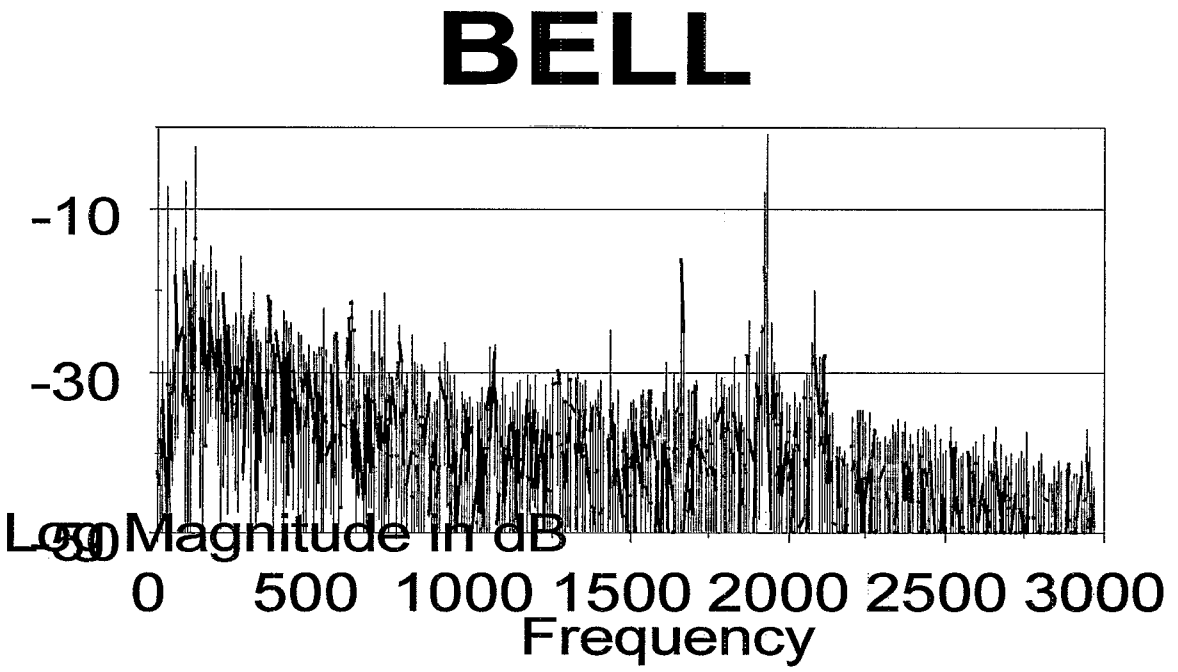

Figure 4

Implementing Noise Tolerant Speech Recognition

The recognizer uses seven noise canceling solutions: automatic gain control, software transient filter, software log-likelihood filter, differential noise canceling microphone, optimal syntax development, noise calibration, and the push-to-talk for enabling the recognizer only during speech input [3]. Combined, these noise canceling solutions offer sufficient noise rejection needed to accomplish speech recognition for this air traffic control application.

Automatic gain control (AGC) adjusts the amplitude of the voice input to an optimal level. Specifically, the input level at the end of a phrase is compared to the previous phase and subsequently adjusted as needed in 2 dB increments. Adjustments are not performed during the speech input to prevent inadvertent compensation of naturally varying voice levels. Effectively, automatic gain control compensates for varying ambient noise levels by maintaining a constant optimal input level relative to the surrounding noise.

A built-in feature of the ITT recognizer is an optional software transient filter that averages the energy of an unwanted transient noise over some predetermined sampling period. Theoretically, this evenly biases the range of frequency components to yield a discernible input response free from the influence of the transient. This option is implemented in the ISIS system.

The log-likelihood noise filter uses the recognizer word score for each word to eliminate poorly recognized speech. The recognizer provides two scores for each recognized word, a word score (W) and a rejection score (R). Scores are filtered by a task oriented application program. Specifically, the difference of the scores is limited to a threshold. The threshold is set to optimize correct and incorrect rejection. Basically, the rejection score indicates how well recognized speech compares to your voice and the word score is how well recognized speech compares to a stored syntactically correct template of your voice. The difference of the two scores is compared to the threshold. Ideally, unwanted noise will equally offset both scores effectively separating noise from human speech.

The ISIS speech recognition system employs a dual element noise canceling microphone. Sound that enters both microphones is canceled by diametrically combining the signals. Only sound that independently enters the primary microphone is forwarded to the recognizer. The amplitude of the incident speech to the primary microphone is

significantly larger than any reflected input to the secondary microphone. Ideally there wouldn't be reflected input. ISIS operators are seated in front of a computer workstation with a smooth monitor surface thereby reflecting some speech input the secondary microphone. Research in this area continues leading to improvements for a production ISIS speech recognition system.

Syntactical optimization means providing a minimal set of valid syntax paths to satisfy system requirements. More paths than needed increases perplexity and can decrease recognition accuracy. Further, design of training scripts to develop reference patterns (called templates) requires familiarity with the intricacies of template training. For example, certain beginning and ending phonemes have a strong dependency on their surrounding words. Optimal implementation of the vocabulary for training depends on use of the phraseology in context while considering sensitive boundary phonemes. An effective training script significantly improves overall recognition accuracy thereby reducing susceptibility to ambient noise.

Noise calibration is an operator initiated process that samples speech input relative to ambient noise and adjusts stored speech templates to compensate for the added noise. Templates are created in a quiet enrollment session and are not adjusted until a calibration has been initiated. Experience shows that calibration significantly improves recognition accuracy in a high noise environment.

Lastly, the push-to-talk switch is engaged during active speech input and therefore disengaged when not in use. Use of the PTT significantly reduces the probability of unintentional speech entering the system and potentially being recognized as valid input.

Out of Phraseology Study

Out of Phraseology (OOP) attempts to eliminate forced recognition of illegal phrases and establishes confidence levels for recognized phrases. Effectively, the OOP rejection algorithm discards phrases that have a low confidence level.

Filler templates are generic small samples of speech. Sounds other than speech do not compare well with generic speech filler templates and are therefore scored poorly. Additionally, the filler templates will generally score well with any spoken word, including incorrect or out-of-phraseology words. This has led the NAWCTSD Speech Technologies Group to investigate possible improvements for words spoken out-of-phraseology. Further, OOP speech would naturally have low word scores thus providing a discernible difference between correct and incorrect words. This difference is then compared to a threshold parameter to determine if the word was indeed part of the established phraseology.

An optimal OOP algorithm is obtained through analysis of acquired test data from operators using a particular syntax. Testing was performed at the carrier air traffic training school in Memphis using the ISIS syntax and the ATC personnel from the CVN George Washington. The OOP algorithm balances acceptance of mis-recognized words and false rejection of valid words. Currently, a difference algorithm (W-R) is implemented for ISIS (see figure 5). The OOP study [4] shows that a linear algorithm is sufficient for the ISIS syntax and further testing would be beneficial in fine tuning tolerable false rejection of valid speech input.

Analysis of the Memphis data shows that the ISIS speech recognition system has an overall raw word recognition accuracy of 94.7%. This is found by totaling all spoken words and separating the correctly recognized from the mis-recognized words.

CORRECTLY RECOGNIZED - 8450 correct words
MIS-RECOGNIZED - 471
OOPS - 68 incorrect words

RAW WORD RECOGNITION ACCURACY RATE
100-((471/8450+471)*100)= 94.7%

WORDS SPOKEN OOP
((68/8450)*100) = 1.5%

Forced recognition could occur in the absence of a rejection algorithm thus forcing recognition of words spoken out-of-phraseology. The W-R rejection algorithm significantly reduces forced recognition of

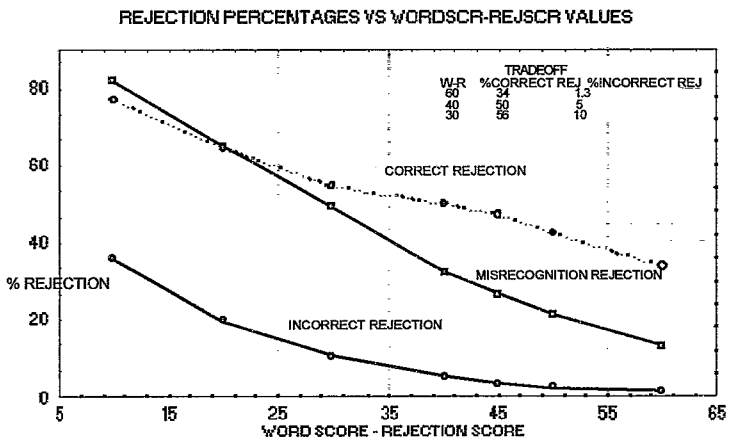

Figure 5

OOP words. Further, optimization of the rejection threshold from a W-R =60 to a W-R =40 increases correct rejection of OOPs from 34% to 50%. This means an increase of 16% words spoken out-of-phraseology would be rejected. The penalty would be an increased incorrect rejection from 1.3% to 5% (see figure 5). Testing of an adjusted threshold would be necessary to determine if this increase in incorrect rejection is tolerable by the user.

CONCLUSIONS

The ISIS speech recognition system operates in real-time with a 118 word vocabulary and a moderate to large perplexity.

The ISIS speech recognition system employs seven noise canceling solutions: automatic gain control, software transient filter, software log-likelihood filter, differential noise canceling microphone, optimal syntax development, noise calibration, and the push-to-talk for enabling the recognizer during speech input. Combined, these noise canceling solutions offer sufficient noise rejection needed to accomplish speech recognition for this carrier based air traffic control application.

The host recognition process parses recognition results and packetizes the information based on the interface protocol in less than 150 milliseconds on average, a factor of two better than the system requirement

A series of bells are sounded on board ship to indicate the current hour and anouncements. Some of the bell frequency contents happens to be in band with natural speech. There is a spike of energy at just below 2 Khz which can interfere with speech recognition if the PTT is depressed during the bell.

Ninety-six dB of arresting gear noise does not interfere with speech recognition because the energy is out of band with the speech system.

Analysis of the Memphis data shows that the ISIS speech recognition system has an overall word recognition accuracy of 94.7%.

Optimization of the rejection threshold from a W-R =60 to a W-R =40 increases correct rejection of OOPs from 34% to 50%. This means an increase of 16% words spoken out-of-phraseology would be rejected.

RECOMMENDATIONS

Investigate the implementation of a 1.7 Khz-2.0 Khz notch filter to remove the bell sound from the recognizer input.

Investigate use of a displaced noise canceling microphone device to eliminate adverse audible noise.

Fully integrate the speech recognition system with the ISIS host to provide a more synergistic user interface.

The OOP study shows that a linear W-R algorithm is sufficient for the ISIS syntax and further testing would be beneficial in fine tuning tolerable false rejection of valid speech input.

REFERENCES

1. Williams, Kotick, Giambarberee (1993) *"ISIS Shipboard Speech Recognition Testing"*, NAWCTSD

2. Richie, Sam (1995) *"ISIS Shipborad Noise Analysis Report"* , NAWCTSD

3. ITT, Defense & Electronics (1994), *" VRS-1290 Hardware Interface and Application Library"*, ITT Aerospace/Communications

4. Smith, Dana (1995) *"Test Report for ISIS Speech Data Collection"*, NAWCTSD