Robust Speech Recognition over Narrowband Communications Channels in Tactical Battlefield (U) Environments

Bryan George[2], Fred Goodman[2], Burhan Necioglu[2], Frank Ruscil, George M. Shuttic, Peter Wilkes, Shiho Fujii, Roger Yarosh

The MITRE Corporation
12 Christopher Way
Eatontown, NJ 07724
(732-389-7858)
gshuttic@mitre.org

MITRE Corporation(2)
11493 Sunset Hills Rd.
Reston, Va. 20190

(U) ABSTRACT

(U) Currently, low echelon military commanders have no ability to directly access the remote Situation Awareness (SA) and Command and Control (C^2) information contained in Battlefield Tactical Operations Center (TOC) databases. Previous attempts, using Speech Recognition(SR) over a Narrowband Communications channel using a coded speech signal have been unsuccessful. However, our research and testing has shown that, due to recent advances in the areas of narrowband Digital Speech Processing and SR, it is now possible to perform robust recognition of a useful vocabulary over narrowband tactical communications channels.

(U) This paper describes an implemented and tested system which was designed to provide secure, bandwidth efficient, hands-free, automatic remote database access to commanders in the field. This system uses a COTS-based Digital Speech Coder with an advanced Noise Preprocessor to communicate over a narrowband IP-based tactical network into a database server. The server employs COTS-based Speech Recognition(SR) system components and a SQL-generator program to provide direct access into a TOC-type database. Queries made by the user are transmitted to the database over a coded Narrowband Communications Channel and the answers are spoken back to the user.

(U) The experimental results of our implemented wireless, Voice over IP (VoIP) system show that the end-to-end system performance is quite good even in the presence of ambient tactical noises.

(U) INTRODUCTION

(U) During Tactical Battlefield Operations, Situational Awareness (SA) and Command and Control(C2) data are being acquired, processed, and stored at an ever-increasing rate. Despite the importance of this information in the success of battlefield actions, disseminating it to the Battlefield Commanders over standard military communications networks within a useful interval of time, is a very difficult and ongoing problem.

(U) Two technical areas which contribute to the difficulty of this task are: one, the communications bandwidth that is required to transmit the information to the users and two, the mechanism that is used to get the data from the database of interest

(U) At the higher Echelon levels (Divisions, Corps, EAC), where the Communications Channels have large ATM (Asynchronous Transfer Mode) –size bandwidth, even Video TeleConferencing (VTC) can be transmitted and shared among Network users. However, at the lower tactical levels, this

amount of communications bandwidth is not available .

(U) In particular, at the level of the Lower Tactical Internet (Platoons, Squads) , the Low-level Echelon Commanders receive only a small portion of the Battlefield Information which is stored in the TOC databases. Additionally, they are not able to directly request information from the Tactical Operations Center (TOC) databases, which contains information of importance to them.

(U) Furthermore, at the lower Tactical Internet levels, successfully sending data over a narrow-band communications channel requires using a compression algorithm on the data before sending it over the channel. And most standard compression algorithms do not perform well in noisy military environments.

(U) The second major technical area of concern is how to access the information of a TOC database and return the information to the user. Database accesses are usually performed by having a person type Database Standard Query Language (SQL) statements into a keyboard at a computer. Therefore, even if the communications bandwidth was available, the question of how to automatically access the database information still remains.

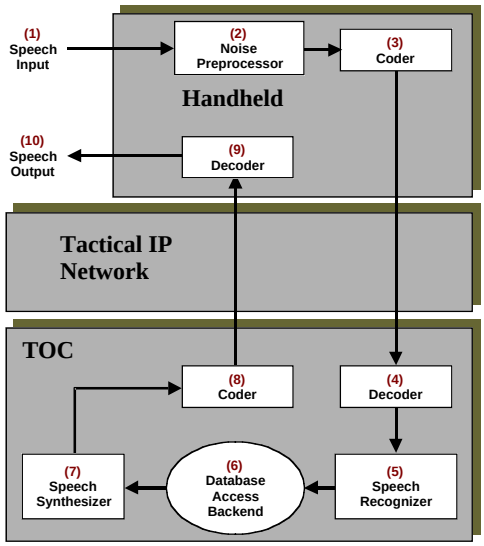
(U) Speech Recognition (SR) has been suggested as a new technology, which could be used in the battlefield. However, until recently, Commercial-Off-The-Shelf (COTS) SR products have not been successful when used outside of their restrictive low-noise office environments.

(U)We believe that, due to several recent advances in the areas of both speech compression and SR, it is now possible to use these two technologies together effectively in a tactical battlefield environment. What we have developed is a system for allowing voice-driven remote access of database information by battlefield commanders over narrow bandwidth communications channels.

(U) METHODOLOGY

(U) Figure 1. provides a high level view of our demonstration system

UNCLASSIFIED
System Overview



UNCLASSIFIED

Figure 1.

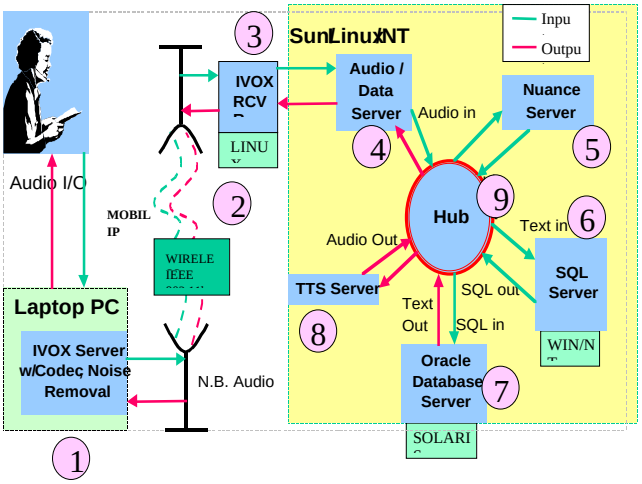
(U) The system can be described as having three parts: one, the handheld PDA/radio which is used to process the input and output speech signals; two, the communications network over which the compressed speech is transmitted; and three, the TOC computer and database components which are responsible to getting data from the database, and returning it to the user. The individual system components are designed to work in one of the system areas.

(U) The Department of Defense (DoD) has recently adopted the Mixed Excitation Linear Prediction (MELP) as its new standard Digital Speech Processing Coding algorithm for 2400 bps, replacing the previous standard, Linear Prediction Coding (LPC-10e) algorithm. The Noise Pre-Processor used is the Harsh Environment Noise Pre-Processor (HENPP) filter, which was developed by ATT from NSA. Testing has shown that the HENPP

provides significant improvement when compared to using the MELP coder alone[1,2,3,4].
(U) Although the project began using only the MELP Speech Codec to compress the input speech signal, we decided to test other codecs, including commercial types, to determine their bandwidth vs quality performance with respect to the SR components of the system. To this end we additionally tested other algorithms including CodeBook Excited Linear Prediction (CELP), the Military's standard at 4.8 Kbps, and the Commercial algorithms; G723.1 (5.2,6.3 Kbps), G729 (9Kbps), and the GSM algorithm (13Kbps).
(U) Referring back to Figure 1, the input query is spoken by the user, and the ambient noise level is reduced by the HENPP before the speech signal is processed by the transmitter portion of the MELP Speech Codec (numbered 1, 2, 3 in the picture).
(U) Because of our belief that the majority of the future communications of the military will be digital, the communications channel is assumed to be a standard IP-based network. Therefore, the coded speech output parameter bit stream is packetized into IP packets and routed over the network to its destination.
(U) The TOC area in Figure 1 is where most of the computationally intensive processing steps occur. The packet data are removed from the bitstream, decoded by the receiver portion of the Speech Codec, and used to generate a synthesized wav file(number 4). The coded wav file is sent to the SR which then generates a token string consisting of its best guesses for the words of the original spoken speech query.
(U) This token string is then passed to an SQL generation program which converts the query token string into its corresponding database SQL string(number 5). This SQL string is input to the Database, and a data string is output containing the answer to the query.
(U) The output data string is then sent to a Text-to-Speech (TTS) module which converts the text string into a speech wav file. The speech wav file is then processed by the transmitter portion of the Speech Codec and the bitstream is packetized and sent out over the IP network to the user.

(U) The speech data is removed from the packets by the user, and the parameter bits are used to construct a synthesized speech signal of the answer to the original query. Finally, this wav file, which contains the answer to the user's query, is output through the user's headset speakers (number 8, 9).
(U) A detailed description of our system is shown in Figure 2. The arrows show the complete path of the query from the user and the answer back to the user.

UNCLASSIFIED
System Description with IVOX VoIP



. UNCLASSIFIED

Figure 2

(U) The system was designed to require minimal assumptions concerning the types of Operating Systems which are needed for each of the system components.
(U) The real-time, wireless, VoIP demonstration system uses IVOX, to perform the necessary narrowband Digital Speech compression using several types of Speech codecs, including MELP, CELP, G723.1, G729, GSM, CVSD, LPC-10e, and PCM. This is shown in Figure 2, areas 1, 2, and 3. A laptop PC (region 1) was designed as the "thin client" system, running Windows 2000.
The Speech codec system, shown in region 3, used RedHat LINUX 7.0 as the Operating System.

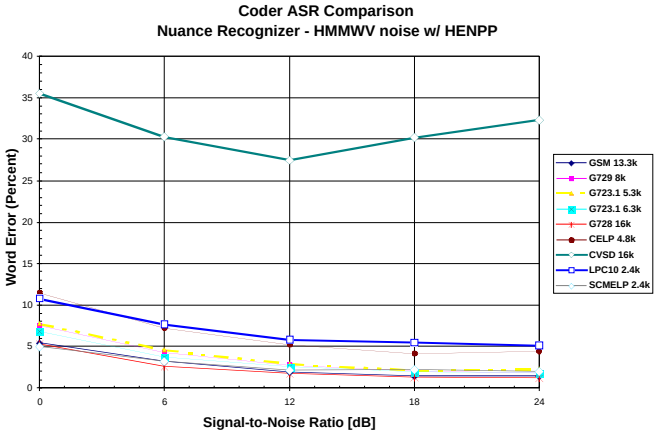
(U) The software running in the TOC area, numbered 4-9 ran under the Solaris Operating System, except for the EASYASK SQL generator (number 6) which ran on a WINDOWS 98 platform.
(U) The utility of our system to the Tactical Battlefield Commander is defined by the information which has been stored in the database. Initially, we used an MS-ACCESS database to do preliminary testing of the overall system. For the demonstration system, we obtained a more realistic ORACLE database which had been designed by the MITRE Corporation for the recording of Incident Reports and Situational Awareness (SA) information in Bosnia, where it is currently being used.
(U) Due to the classified nature of the actual populated database, for the purpose of demonstration it was necessary to generate consistent, but artificial data, including incident reports, types of incidents, times of occurrence, locations, and weather reports with which to populate it.
(U) A speech grammar was generated for the test system which allowed the user to ask questions of the database like the following: What time is it?, What's the current Threatcon level?, What's the weather forecast for the next 'n' hours?, How many incidents have occurred in the last 'n' hours?, What was the type of the last reported incident?, and How far away are we from the last reported incident?.

(U) In order to make the system easy to use and as transparent as possible to the user, variations in the grammar of the queries were allowed, e.g. What is the time?, What's the time?, Time?, and Time Check? are all treated by the system as requests for the Current Time.
(U) In order to coordinate the necessary components of the TOC software and exchange the necessary data, we leveraged the work of the ongoing DARPA sponsored COMMUNICATOR project. The COMMUNICATOR project, is a several year effort whose goal is to design an Open Source Software interface for use by Researchers of the Speech Community. Our use of this software provides the demonstration system with a standard, general, interface for all of the software component modules of the TOC, even allowing them to be run on different computer platforms.
(U) Although we initially evaluated several SRs, most of the commercial SRs were designed to be "dictation systems", using large vocabularies. And, in noisy environments they suffered in their ability to robustly discriminate between spoken words. Because, of the relatively small number of data types in the database, we needed only a small vocabulary. Unfortunately, most of the dictation system SRs did not provide the capability of constraining the vocabulary set over which recognition is attempted. As a result, only two appeared to be viable candidates for use in the system, the NUANCE Commercial SR system and the Open Source ISIP system.
(U) For the TTS component, we used the Open Source Festival system. There were several different voice speaker profiles which were available for Festival and we chose the profile which provided us with the best intelligibility results after the TTS output wav file had been processed through the MELP and G723.1 Speech Codecs.
(U) Two different SQL generators were tested; EASYASK, and one which was available from NUANCE. For the purpose of the demonstration project, we chose to use the one from NUANCE.

(U) RESULTS

(U) For the purpose of our research, we looked at the Speech Coder/SR word error rates for several noise levels(dB) of different noise types. Figure 3 shows our results for HMMWV background noise using the HENPP Noise Pre-Processor with the MELP algorithm (SCMELP).

UNCLASSIFIED



UNCLASSIFIED

Figure 3

Note that one of the Digital Speech Codecs performs very poorly across all background noise levels, while all of the other coders follow each other in a much closer fashion as the input noise level is changed. This Codec is the Continuously Variable Slope Delta-Modulation (CVSD) algorithm which is used in the SINGARS radio in the Lower Tactical Internet.

(U) CVSD is a waveform coder and as such, distorts the input speech signal and dramatically reduces the ability of the SR is correctly determine the spoken speech. The reason for this problem is that the SR has been trained on standard voice speech data, i.e. data that has not been passed through CVSD.

(U) Because the commercial speech processors which we used did not allow us to "re-train" their recognizers on different types of input signals we were forced to remove CVSD from consideration as a possible candidate for the demonstration system's Speech Codec.

(U) CONCLUSION

(U) What we have shown is that for a small vocabulary sized system, it is possible to use narrow bandwidth coded speech to obtain information from a remote database in a reliable fashion, even in the presence of tactical battlefield noise.

(U) Although the initial design of this system addressed the concerns of the Lower Echelon Tactical Commander, the system does have potential usefulness in two other areas. In SUO (Small Unit Operations) or MOUT (Military Operations in Urban Terrain), the individual members of the units can be aided to operate more autonomously and effectively by allowing each of them to obtain SA information directly from a central unit database.

(U) Secondly, the Military is becoming increasingly responsible for providing support and assistance in MOOTW (Military Operations Other Than Warfare) operations; particularly providing protection and support for non-military NGO (Non-Governmental Organization) personnel (e.g. UNICEF, Red Cross, Doctors without Borders). This system, when used by the NGO personnel, would provide some level of support for them, and some measure of safety to them without draining the limited resources of the military units.

(U) References

[1] Accardi, Anthony J., and Cox, Richard V., "A Modular Approach to Speech Enhancement with an Application to Speech Coding", ICASSP-99, Phoenix, AZ

[2] Collura, John S., "Speech Enhancement and Coding in Harsh Acoustic Noise Environments", IEEE Speech Coding Workshop 99

[3] Collura, John S., Brandt, Diane F., and Rahikka, Douglas J., " The 1.2 Kbps/2.4Kbps MELP Speech Coding Suite with Integrated Noise Pre-Processing", MILCOM-99

[4] Malah, David, Cox, Richard V. and Accardi, Anthony J., " Tracking Speech-Presence Uncertainty to Improve Speech Enhancement in Non-stationary Noise Environments", ICASSP-99, Phoenix, AZ

[5] Dellomo, Michael, Hoyt, JoAnn, and Shuttic, George M., "Design and implementation of a real-time mediumband speech coding system: critical point coding with time harmonic domain scaling: Volume I. (U)" MTR84W00188-01

[6] E. B. George and M. J. T. Smith, "Speech analysis/synthesis and modification using an analysis-by-synthesis/overlap-add sinusoidal model," IEEE Transactions on Speech and Audio Processing, September 1997.