EUROSPEECH '97
5th European Conference on Speech Communication and Technology
Rhodes, Greece, September 22-25, 1997

INTEGRATING A RADIO MODEL WITH A SPOKEN LANGUAGE INTERFACE FOR MILITARY SIMULATIONS

E. Richard Anthony, Charles Bowen, Margot T. Peet, Susan Tammaro
The MITRE Corporation
1820 Dolley Madison Drive, McLean, VA 22102-3481 USA
Email: mpeet@mitre.org

ABSTRACT

We incorporated a simulated military radio into a spoken language interface to a distributed simulation environment for military commander training. The resulting architecture bypassed the inherent problem of acoustic mismatch that arises in integrating radio output with a speech recognition front end, while at the same time preserving the realism of speech synthesis output through a military radio. We assessed the utility of formal evaluation methods to benchmark the impact of the radio model on a commercial speech synthesizer.

1. INTRODUCTION

Over the past few years, several research efforts have made considerable progress in developing prototype spoken language interfaces to distributed simulation environments for military commander training [5,6]. Increasingly, the military training community is relying on these distributed simulations to support large-scale military exercises. However, these exercises require large numbers of military or civilian personnel to input orders to the computer simulation and interpret the output. To reduce the overhead costs associated with simulation-based training, the community is working to reduce the number of intermediary personnel required and to make their roles more tactically relevant, that is, making them part of the training audience rather than the technical support team. Spoken language technology holds the promise of enabling commanders to interact directly with virtual troops, thus providing a realistic training experience in a simulation environment.

Previously, we developed a prototype spoken language interface to a simulation battlefield environment, the Army's Modular Semi-Automated Forces (ModSAF) simulation [1,6]. The ModSAF environment simulates platoons and battalions of tanks, as well as other military platforms. Commander trainees practice maneuvers by issuing speech commands into the speech recognizer. The interface converts spoken commands into ModSAF actions and returns virtual entities' confirmations and responses using a speech synthesizer. For example, the trainee issues the command, "First platoon halt," and hears the response, "Order received: first platoon halt".

To avoid "negative training," however, it is important that participants face all of the challenges they would normally face during actual combat. Right now, there is no way to make the exchange of messages between live players and simulated entities seem realistic. Trainees know when their messages are being sent to a simulated entity through the support staff. They are not forced to cope with the distractions of static and background noise over the radio, as they would in the field, nor are they forced to repeat or clarify their reports because the transmission would normally be distorted or partially blocked.

Our goal is to design an architecture that preserves the degradation in communication that the commander experiences in a field training exercise while at the same time relying on commercial off the shelf speech recognition and speech synthesis technology. A number of technical issues must be addressed in developing the architecture.

First, speaker independent speech recognition systems are pre-trained with speech collected in an identical acoustic environment. If microphone input to the recognizer does not match this environment, there is an acoustic mismatch, which negatively impacts recognition. In theory, therefore, it is desirable to first train the system on speech distorted by radio. However, collecting this kind of speech corpus is a costly and time-consuming endeavor, and its usefulness is limited to one kind of radio channel only. Furthermore, commercial systems are essentially black boxes that only the vendors themselves can train.

Second, the issue of tandeming radio communications with speech synthesizers has not been addressed in the research community. Our literature survey showed that most studies of the intelligibility of speech synthesizers focused on telephone environments, and were primarily diagnostic in purpose. We found no experimental protocols for examining the impact of radios or speech coders on synthesis intelligibility.

2. SYSTEM CONFIGURATION

2.1 SINCGARS Radio Model

We integrated a military radio model with our spoken language interface to simulate real communication

10.21437/Eurospeech.1997-317

channels and provide a realistic representation of communication systems within distributed interactive simulations. Our architecture bypasses the acoustic mismatch problem while at the same time preserving the degradation in communications experienced in a realistic training exercise.

The model we deployed, the SINCGARS Radio Model (SRM), emulates the properties of the Army's standard tactical voice/data radio for mounted and dismounted combat units, the SINCGARS radio. The SINCGARS radio operates in the 30 to 88 Mhz range with 25 Khz channels and provides two primary modes, single channel mode and frequency hopping mode. Both modes support digital voice transmission using 16 Kbps CVSD[1] coding and data transmission using 16 Kbps CPFSK[2] coding [7]. Four transmit power levels allow communication from 300 meters to 35 kilometers.

Designed to work within the ModSAF environment, each SRM attaches to a platform, typically a tank, within the simulation. As the platforms move, the SRMs move accordingly. Each SRM, sender and receiver, is configured for single channel operation mode for voice transmission and tuned to the same frequency. The SRMs use the terrain within the ModSAF simulation to determine a communication channel, which models the radio signal propagation from sender to receiver. Depending on the calculated channel, radio communication may or may not be possible. The sending SRM encodes the digital speech samples and sends the data to the receiver SRM. The receiver SRM decodes the speech, determines the channel based on the form of the terrain and the locations of sender and receiver, and adds noise to the decoded speech signal appropriate for the calculated channel. Since the SRM is a digital radio, it calculates a bit error rate (BER), based on the channel, to distort the speech signal. The receiving SRM also determines whether the sender was out of range or too much noise was present and drops the signal. Each SRM is capable of sending and receiving speech signals, but it permits only one radio communication at any time on a transmission frequency.

2.2 Architecture

Our system, shown in Figure 1, submits the trainee's commands directly to the speech recognizer without corruption by the radio pair. The sending SRM also receives the trainee's speech commands which it transmits to the receiving SRM. The controller module

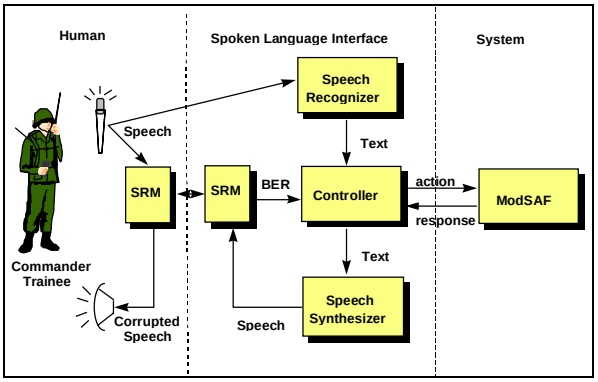

Figure 1

receives the recognized text and uses the bit error rate (BER) from the receiving SRM to determine if the message is intelligible. Low BER values, signifying little distortion, indicate that the message is likely to be understood, and high BER values, signifying high distortion, indicate that the message is likely to be misunderstood [8]. The interface makes a probabilistic decision based on the bit error rate for each transmitted speech message. The controller module forms ModSAF actions from messages that are understood and prompts the trainee to repeat messages that are misunderstood. If the sending SRM is out of range, then the controller simply ignores the speech message. While SRMs are not used to preprocess the trainee's speech signal (for recognition), they are used to distort responses from ModSAF virtual units. Virtual units' responses, generated by the controller, are synthesized and transmitted to the commander trainee's SRM. Thus, the trainee hears a human sounding synthesized voice which sounds like it has been transmitted over a radio.

For voice output, the controller converts ModSAF responses to text, which the synthesizer converts into human sounding speech. The SRM pair transmits the synthesized speech to the trainee. Thus, during a response the trainee will hear a human sounding synthesized voice. But, the SRM pair codes the speech and adds noise so that the speech sounds as if it had been transmitted over a radio. The amount of noise added is appropriate for the type of radio, the location of the sending and receiving units, and the shape of the terrain.

This system successfully simulates the effects of real radio communication and introduces a sense of realism. Distortion, introduced by the radio's speech coder or by the geography of the terrain, makes the radio speech messages difficult to understand. As in field maneuvers, radio communications are not always possible. The trainee cannot issue commands if the trainee's subordinates are too far away or if other users are transmitting on the same frequency.

[1] Continuously Variable Slope Delta (CVSD).

[2] Continuous-phase frequency-shift keying

3. EVALUATING THE IMPACT OF MILITARY RADIO ON SPEECH SYNTHESIS INTELLIGIBILITY

To the end user, the most notable aspect of the interface is the voice output capability, which tandems a high quality commercial speech synthesizer with a digital radio. Listening to the limited set of utterances generated by our scenario, we were struck by the decrease in intelligibility when synthesis was tandemed with the SRM. We wondered whether it would be possible to exploit variables provided by the commercial synthesizer, such as intonation, to enhance intelligibility in this environment. Previous studies have tried to link intonation and intelligibility, but have not been conclusive [2]. Our first step was to establish a benchmark of synthesis intelligibility when it is processed in ways that simulate the effects of the SRM. A literature review found evaluations of speech synthesizers in clean or telephone environments only, with no benchmarks or protocols to assess intelligibility in digital radio environments. We examined a number of experimental protocols, but found none that assessed the impact of radio channels on synthesis intelligibility.

To obtain a benchmark of the impact of the SINCGARS Radio Model on the intelligibility of a commercial speech synthesizer, we chose a recently published protocol originally developed to compare synthesizers across languages. Our goals were twofold. First, we wished to assess whether a protocol developed to compare intelligibility across synthesizers could provide clear results for the intelligibility of a single synthesizer processed through multiple conditions. Second, if this protocol were successful, our goal was to obtain an initial benchmark of the impact of a military radio on speech synthesis.

The test we used, the Semantically Unpredictable Sentences (SUS) test [3], consists of sentences that are syntactically correct but absent of semantic meaning, e.g., "The strong way drank the day." Comparison of results of this test with other tests demonstrate its rigor and level of difficulty [4]. As such, it potentially offers a baseline of synthesizer performance.

3.1 Methods

3.1.1 Stimulus Preparation - Syntactic Structures

Sentences from the 5 syntactic structures suggested in [3] were used, presented in a random order. None of the sentences exceeded eight words in order to avoid limitations of short-term memory. A total of 114 sentences were generated: 21 from two of the syntactic structures, and 24 from each of the other three structures.

3.1.2 Stimulus Preparation - Synthesized Conditions

To obtain a benchmark, we analyzed the processing the speech signal undergoes in the SRM, and broke it into two categories, one of which occurs in a clear line of sight situation between sending and receiving units, and the other which simulates the effect on line of sight when a physical object moves between sending and receiving units. In this way we varied the original test protocol: rather than comparing the intelligibility of separate synthesizers on a single corpus, we compared the intelligibility of three channel conditions on a single synthesizer. We tested three different synthesized conditions: clean (Condition 1), coded (Condition 2) and coded and distorted with a bit error rate (Condition 3). The clean condition was the unaltered high quality commercial speech synthesizer. Condition 2 was coded speech that simulated CVSD coding in a clear line of sight situation between sending and receiving units. Condition 3 was coded and distorted speech that simulated corruption introduced by line of sight effects in the radio model when one of the entities moves behind a physical object and out of the line of sight.

The presentation order of the different synthesizer conditions was randomized. The 114 total sentences were divided equally between the 3 synthesized conditions so that each condition had 38 sentences.

3.1.3 Subjects

A total of 20 subjects with no experience with synthetic speech and no hearing loss participated in this study. All were native speakers of American English, and were between the ages of 20 and 40.

3.1.4 Experimental Procedure

Each of the 3 synthesizer conditions was randomized across conditions and recorded on a single tape. All subjects heard the sentences in the same order. The stimulus tape was presented to subjects during a single test session. This was done to mimic real use of the synthesizer in conjunction with the SINCGARS radio where some transmissions could be distorted in different ways and some could be clean transmissions. Stimulus material was presented to each subject by means of a standard cassette player in a quiet room without interruptions.

As in [3], subjects were first asked to complete a training session in which they listened to a tape of 12 sentences generated by a human voice and the same 12 sentences generated by the three different synthesizer conditions. The experimental task was for the subjects to listen to the sentences and write down what they heard as accurately as possible. If only part of the sentence was intelligible, subjects were requested to

	Clean: Condition 1	Coded: Condition 2	Coded& Distorted: Condition 3
% Correct Sentences	33.55%	12.89%	6.97%
Total Sentences Correct	255	98	53
Mean/sd Correct	m=12.75 sd=6.11	m=4.9 sd=2.731	m=2.65 sd=2.084

Table 1

leave dashes for the words which they could not understand.

3.2 Results

As suggested in [3], only sentences which were entirely correct were scored as correct. In order for a sentence to be scored as correct, all words had to be in their correct position in the sentence. For all 20 subjects, the percentage of correct sentences in each synthesizer category was calculated. Table 1 shows these results. As might be expected, the clean condition resulted in the highest percent of correct answers with the coded condition having the next best performance and the coded and distorted condition demonstrating the fewest correct answers. This was a statistically significant difference, $F(2,28) = 105$, alpha $< .01$.

3.3 Discussion

When integrating the output of a speech synthesizer with a military radio model, we became aware of the extent of the degradation in intelligibility. Although there have been several attempts to correlate prosody with intelligibility, e.g. [2], results have not been conclusive. As a preliminary step in determining whether prosody impacts intelligibility of speech synthesis in a military radio channel, we obtained a benchmark of system intelligibility in a clean environment and in environments that simulate the effects of a military radio. We first evaluated whether a methodology used to compare different synthesizers could also be used to compare different synthesizer conditions (e.g., clean, coded, and coded and distorted). Our results using the SUS test looked promising, and it appears to be a sensitive measure of differences in conditions. We plan to use this methodology for future studies. We have already begun to look at the impact of intonation on synthesis intelligibility in a radio model environment.

4. ACKNOWLEDGEMENT

This work was supported by the U.S. Army Simulation, Training and Instrumentation Command (STRICOM), Contract No. DAAB07-96-C-E601, Project 86840.

5. REFERENCES

[1] J. Beetem, R. Macmillan, J. Pace, M. Peet and M. Salisbury, "*Evaluating the Feasibility of Voice Input-Output in a Simulation Environment*," Proc. 1st International Symposium on Command and Control Research Technology, National Defense University, , pp. 295-299, 1995.

[2] C. Benoit, "An intelligibility test using semantically unpredictable sentences: Towards the quantification of linguistic complexity," *Speech Communication*, Vol. 9, No. 4, pp. 293-304, 1990.

[3] C. Benoit, M. Grice and V. Hazan, "The SUS test: A method for the assessment of text-to-speech synthesis intelligibility using Semantically Unpredictable Sentences", *Speech Communication*, Vol. 18, No. 4, pp. 381-392, 1996.

[4] V. Hazan and B. Shi, "*Individual Variability in the Perception of Synthetic Speech*," Proc. 3rd EUROSPEECH '95 pp. 1849-1852, Berlin, 1995.

[5] R. Moore., J. Dowding, H. Bratt, J.M. Gawron, Y. Gorfu and A. Cheyer, "CommandTalk: A Spoken Language Interface for Battlefield Simulations," Ms., 1996.

[6] M. Peet, E.R. Anthony, S. LuperFoy, M. Salisbury,"Lessons Learned: Voice Input-Output Testbed for Distributed Interactive Simulations," MITRE Working Note WN950000249, 1995.

[7] J.G. Proakis, "Digital Communications," McGraw Hill, Inc., New York.1989.

[8] C. P. Smith, "An Evaluation of the Speech Intelligibility and Voice Quality of the Electrovox Narrowband Digital Voice Communications Terminal," The MITRE Corporation, pp. 53-60, 1981.