ULTRA-LOW BANDWIDTH SPEECH COMMUNICATIONS VIA PHONEMIC QUANTIZATION

Lockwood Reed
US Army CECOM
Fort Monmouth, NJ 07703

Abstract

Arguably, the key element of IT/C4ISR (Information Technology/Command, Control, Communications, Computers, Intelligence, Surveillance and Reconnaissance) is communications, and key to communications is available channel bandwidth. For a military force to survive and be effective it must communicate efficiently and with minimum exposure. The available communications spectrum is finite. The growing requirements for high-speed data communication for Future Combat Systems and the Objective Force Warrior are putting further pressure on the dwindling communications bandwidth.

This paper proposes a novel speech encoding technique in which speech is analyzed and classified into constituent components known as phonemes. Any language can be broken down into a finite set of phonemes: English having 64, with approximately 56 addressing about 98 percent of the common English lexicon. Currently one of the most efficient encoding algorithms for speech is Mixed-Excitation Linear Prediction (MELP). Bit rates as low as 2400 bps are possible yielding acceptably intelligible speech. This paper will demonstrate that encoding by Phonemic Quantization (PQ) can yield intelligible real-time speech with data rates as low as 50 bps. Therefore a single 2400 bps MELP channel could in theory support 48 PQ encoded channels. In addition, do to the low data rate requirements of PQ, various covert transmission techniques can be employed, such as spread spectrum or burst transmission, which can reduce communications exposure and improve survivability.

1. INTRODUCTION

To date, the most popular approach to reduced bandwidth speech encoding is Mixed Excitation Linear Prediction (MELP). MELP can yield encoded rates as low as 2,400 bps. By exploiting the capability of state-of-the-art speech recognition technology to partition speech into to a bounded set of sub-word classifications, known as phonemes, ultra-low bandwidth speech encoding, on the order of 50 bps should be feasible. In addition the encoded speech can be simultaneously re-synthesized as speech or directly transcribed into textual messages.

1.1 TECHNOLOGY

The evaluation architecture is comprised of the speech recognizer, word-to-phoneme converter, transmitter simulator, receiver simulator, and phonemic synthesizer. For this project several large vocabulary, phonemic based speech recognition technologies were evaluated. Particular attention was directed toward performance with lexicons focused on tactical messages (such as 'call for fire'). The ability of the recognition technology to provide it's output in phonemes rather then whole words was desired, unfortunately most phoneme based speech recognizers do not deal with individual phonemes: most are based on "tri-phone" models. For reasons of cost and schedule a decision was made to utilize the word output of a recognizer, and through a simple translation program, convert the word output into a string of ASCII characters representing the phonemic components of each word.

Upon selection of a suitable speech recognition technology, a search and evaluation of available phonemic speech synthesis technologies was conducted. Particular attention was directed to the selection of the synthesis technology with the highest intelligibility. Various combinations of 64 phonemes can represent the entire English lexicon. Each phoneme can, in turn, be represented by 7 binary bits. Each phoneme code was then applied to a phonemic speech synthesizer for re-synthesis. Given that the average individual speaks at a rate of two phonemes per second, normal speech rates could be encoded at 14 bps. Allowing for higher speaking rates and error detection and correction coding, it was hypothesized that the encoding of real-time speech at 50 bps was feasible.

The transmitter and receiver simulators consisted of two applications connected via network socket. Each

application had the capability to monitor network traffic and record activity.

Therefore the architecture consisted of the speech recognizer, word-to-phoneme converter and transmitter network application residing on one computer, connected via network to a second computer hosting the receiver network application and phonemic synthesizer (see Figure 1.0).

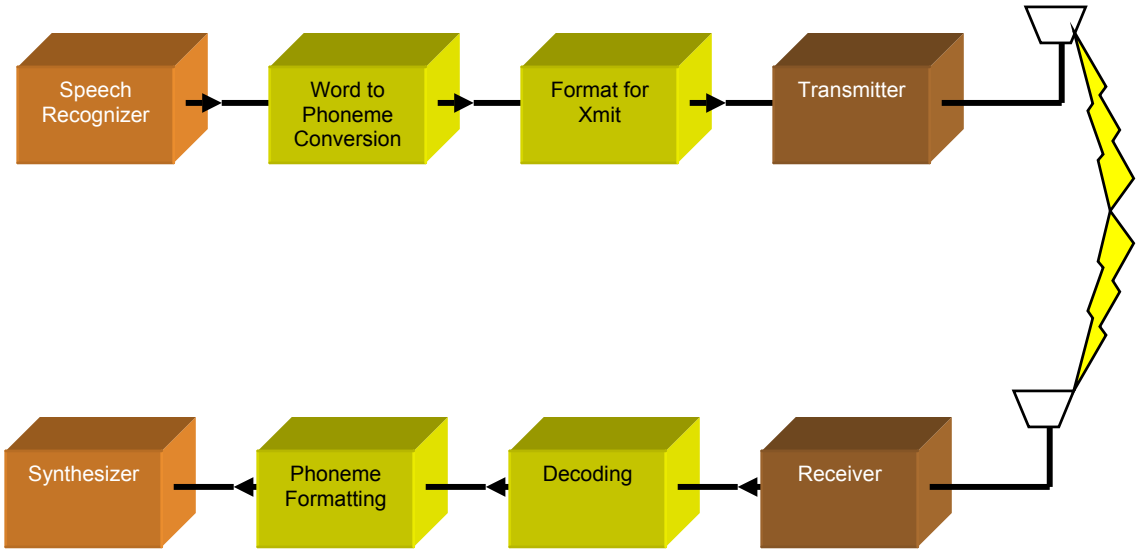

FIGURE 1.0

2. RESULTS

To achieve a real-time output, for a corpus of 50 phrases, required an average bit-rate of approximately 79 bits/sec. The higher bit rate is a result of transmitting the phoneme codes in ASCII format. If encoded in binary, the data rate would be approximately 20 bps. The results of the experiment demonstrated sub-50bps speech compression and validated the performance of phonemic encoding as a viable approach to low bit rate speech encoding.

3. CONCLUSION

The practicality of the phonemic encoding approach is constrained by the accuracy of the recognition engine. This technology would benefit greatly by the revolutionary speech recognition technology contemplated by the Advanced Cognitive Interactive Speech Technology (ACIST) research program. The ACIST would take a new approach to the architecture of the recognition engine, by exploiting recent developments in cognitive research, resulting in the robust recognition of quasi-grammatical speech.