A VOICE TO DATA CONVERTOR FOR USE IN A HOSTILE TACTICAL MILITARY ENVIRONMENT

T. A. Moore and R. A. King

Royal Military College of Science (Cranfield)
Shrivenham, Swindon, Wilts, ENGLAND

Abstract

This paper describes a portable Time Encoded Speech (TES) based Isolated Word Recognition (IWR) Direct Voice Input (DVI) system designed as a demonstrator for investigating very-low-bit-rate recognition/synthesis communications for possible use in a hostile tactical military arena. To cope with the special demands of the military environment, the system has been designed to be operated almost entirely by voice. Dedicated speaker "archetypes" (templates) are stored on plug-in credit-card size memory cards, retained by the individual user. The equipment has no visual display, all feedback being presented aurally via a speech synthesiser. Provision is made for a display as an aid for training, demonstration and diagnostic purposes only.

1 Introduction

Pressures to reduce the bandwidth required for reliable tactical military digital speech communications are continual. Lower bit rates reduce frequency allocation problems and/or allow reliable communication to be maintained under more hostile electronic conditions. Alternatively, low bit rate protocols may provide low probability of intercept (LPI) via, for instance, the use of spread spectrum modulation schemes. In some operational scenarios where the military user is under task-related stress, such systems may offer an effective man-machine interface. Recently the use of recognition/synthesis techniques for achieving limited vocabulary very low bit rate voice communications has received some attention. By this means a message is source encoded using word recognition techniques for the user equipment interface, and reconstructed at the receiver by means of a voice synthesis system.

Such techniques may provide adequate (synthetic) voice communication at bit rates in the order of 100 bps [1], at the expense of flexibility in vocabulary and the implicit speaker verification associated with collaborative users operating conventional voice systems. In some limited applications the use of such systems may be especially advantageous, for example data entry into battlefield ADP systems where the users hands and eyes are essentially occupied.

The prototype unit described in outline in this paper has been developed to permit the investigation of a number of direct voice entry roles. It is intended for exposure to an acoustically hostile environment and has been designed, with the forward area tactical environment in mind, to use low to medium complexity components with a minimum of power requirements. The recognition algorithm used has been demonstrated to perform well in severe acoustic environments [2] and utilises Time Encoded Speech (TES) techniques.

2 Time Encoded Speech

TES is a form of speech waveform encoding. The speech is divided into time intervals (epochs) between successive zero crossings. For each epoch of the waveform, the code consists of a single digital descriptor derived from two parameters, its quantised time duration and its shape. Shape is usually described in terms of the number of extrema occurring within the epoch. Isolated word recognition systems have been developed using TES coding techniques [3] and have demonstrated robust performance in the presence of severe acoustic background noise [2]. The technique is based upon the formation of "archetypes" as reference templates for each word produced by merging "A-Matrices" for several different utterances of each word. The A-Matrix is a frequency distribution histogram recording the second order epoch occurrence statistics over the entire word. Word comparisons are carried out between the A-Matrix created by each input utterance and the archetypal A-Matrices of each word which have been created during training.

3 The Voice to Data Convertor

3.1 Operational Philosophy

The system is designed around a message preparation task, and has a pre-defined vocabulary and grammar. For the military role, the rigid syntax demanded is not a prohibitive imposition on the user, as it might be in a civilian context, since the services are accustomed to, and indeed welcome, well-defined operating procedures.

The user vocabulary is to be divided into subsets in descending order of operational importance such that...
The transmitting VDC operates in the messaging terminal role, whilst the receiving VDC receives the encoded message via its serial port, decodes the content and speaks the original message.

The other serial port allows the connection of a standard VDU, and displays information useful for demonstrations and training (for example the options available as input utterances in the current vocabulary node).

Control of the unit is entirely by voice, with the exception of some functions selected by front panel switches. These select basic system options which must be chosen before the voice system can be used; characterisation options for the audio front-end; transmit/receive operation; train/use pre-stored archetypes. Control of input and output audio levels is provided for.

3.3 Design of Software

The software design has adopted a layered approach (Fig 2). Fundamental device dependent routines form the first layer, which is necessary for any software written for the system. Routines for the formation and management of TES A-Matrices and archetypes form the next layer, and are necessary for any TES application. Above these layers is the code for the particular application. All software is written in 'C', with the exception of a few low level routines.

3.3.1 Confirmation, Error Correction and Timeout

The user interface protocols have been designed to ensure that error-free messages can be constructed efficiently and without the use of any visual feedback. In order to do this two principles were adhered to:-

- Under usual operating conditions, misrecognition is expected to be rare. Thus the user should not need to confirm each word input, rather the machine should assume it correct unless told otherwise.
- Errors which do occur must be corrected by an "Nth Choice" procedure (ie the user can reject the machine's best guess, to be prompted with the "second" best, and so on until the correct word has been selected).

To implement these conditions, the best guess at each word input is assumed correct initially, and the recogniser echoes the word it thinks was said, calculates the next grammar state and requests the next input utterance. Certain of the actions it would take consequent on the recognised word are held over, however, until the utterance is confirmed. Confirmation is achieved by one of two methods. If the utterance was recognised correctly, the user is presented with the next vocabulary node and immediately enters his next word. This is taken as implicit confirmation of the previous utterance, and the system takes those actions held over from the previous input and continues. If, however, the utterance was misrecognised, the user simply waits for a period of a few seconds, for the device to "timeout". At this juncture, a timer interrupt routine prompts the user.
for a “yes/no” response to the question “Was there an error?”. If the answer is “no”, confirmation is assumed and timeout disabled. If “yes”, the machine then enters its “Nth Choice” mode, giving its next best guess and the user responding with yes/no until the correct word is selected. This procedure is represented in Fig 3.

3.3.2 Archetype Updating

In order to maintain recognition performance over time, the unit has the facility to modify archetype data after each correctly recognised utterance. This process ensures that variations in speakers’ pronunciations of words, whether due to injury or other factors, will not require complete retraining of the system. It also reduces the requirement for initial training to be rigorously carried out, since the adaption algorithm will rapidly improve the system performance even if the initial archetype data is formed from fewer utterances of each word than usual.

4 Conclusion

A prototype Voice to Data Convertor has been designed to provide a low cost, low power voice input/output messaging terminal suitable for use by a small population of military personnel - ie highly trained operators with a thorough knowledge of military operational priorities and system syntax. This constraint on the population of users admits a system which, potentially, is extremely small, economical in power, lightweight and allows hands and eyes free data entry. The system is based on Time Encoded Speech techniques, the utility of which in conditions of severe acoustic background noise has been indicated, and limited trials have shown promising performance with whispered speech input. The VDC, as a demonstrator system, may be used to illustrate the advantages and disadvantages involved in the use of voice recognition equipment in the tactical military arena.

Acknowledgements

The authors would like to thank The Director of the Royal Signals and Radar Establishment and the Principal of the Royal Military College of Science (Cranfield) for permission to publish this paper. This work is supported by RSRE via research agreement 2004/64.

References


Figure 2: The Software Architecture for the VDC

Figure 3: The Recognition Loop Flow Diagram