A PCM/VCR SPEECH DATABASE EXCHANGE FORMAT

David S. Pallett

National Bureau of Standards
Institute for Computer Sciences and Technology
Gaithersburg, MD 20899

ABSTRACT

The use of PCM/VCR technology is described for use as a storage and exchange medium for speech databases. In order to provide a limited amount of digital data, use is made of a recorded modem signal for ASCII character string headers associated with the speech tokens. This format can be used to store field recordings of speech material for subsequent digitization, for transfer of speech database material between laboratories and in implementing automated testing of speech recognition devices.

INTRODUCTION

Speech databases are widely used within the speech processing community for research and benchmark test purposes. With the growth of this community, there has been a proliferation of different formats for storage and exchange of digitized and processed speech material. Concurrently, the availability of new audio recording technology has made possible the use of new formats for storage and exchange of the audio source material, recorded speech. The database format described in this paper makes use of PCM/VCR technology to store one to four audio channels along with modem-signal-encoded ASCII character string data. The resulting hybrid format (recorded speech and digital data recorded in the analog audio signal resulting from modem-encoding) is useful for a variety of database exchange purposes.

OTHER MEDIA AND FORMATS FOR DATABASE EXCHANGE

The traditional medium for speech database storage and exchange has been analog magnetic tape. Problems associated with this medium include possible tape saturation and print-through, signal-to-noise degradation with multiple generation copies and problems of mechanical origin such as flutter and wow. It is possible that the use of dynamic compression-expansion noise reduction systems may

introduce subtle distortions in the time domain representation of waveforms. However, there are many speech databases that have been obtained, stored, duplicated and exchanged using analog magnetic tape.

Digitized and processed speech material has many advantages over analog audio derived from magnetic tape. The bulk of speech processing in research is implemented with digitized data. It is convenient to associate data files with the speech data to provide and document such information as: speaker ID, dialect, speech token number, orthographic transcription, signal processing history, test and digitization parameters and segmentation and labeling data. These additional data substantially enlarge the scope of the databases and, unless limited in quantity, are not well suited to recording using an audio format.

There are several digital formats for storage and use of speech database material, and there are no widely accepted standardized formats or agreements on parameters such as sample rate, record length or storage density, (e.g., 1600 vs. 6250 bpi tape drives). The use of differing host computers and operating systems imposes additional constraints. When transferring speech material from one laboratory to another, it may be as cost-effective and efficient to redigitize from high-quality recordings as it is to generate custom software for format conversion and digital copying. This is particularly true when the amount of digital data accompanying the speech material is minimal.

Optical disk technology is under active consideration for speech database storage and access. Presently available compact disk technology, which is particularly well adapted for high-quality two-channel audio, does not appear to be well suited for analog storage of speech database material because of appreciable production costs for the small quantities used within the speech processing community. An alternative may lie in the use of optical data disk technology for storage and exchange of digitized data. Successful

implementation of this technology for speech database exchanges will require the adoption of format and interface standards, the commercial viability of the media, and agreement among speech researchers about acceptable database exchange formats.

PCM/VCR TECHNOLOGY

In recording, this technology uses digital sampling of input audio, encoding of the digital data on a video signal and storage of the video signal on a consumer-quality VCR. In playback, the audio signal is digitally reconstructed from digital data derived from the video signal which is, in turn, obtained from the VCR playback.

More specifically, the two channels of input audio signals are digitally sampled with either 14- or 16-bit resolution at essentially a 44.1 kHz sample rate. A standardized Cyclic Redundancy Check Code is computed and included in the digital data derived from two channels of digitized audio, and the resultant digital data stream is converted into a video signal that is suitable for recording on a variety of consumer-quality video recorders such as Beta and VHS formats. In playback, use is made of the availability of the error correction code to account for the occurrence of drop-outs and other nonlinearities of the video record/playback process. Use is made of buffering, VCR tracking controls, and a fixed clock rate to ensure that flutter and wow (or other frequency modulation effects) are non-existent. Most PCM/VCR systems also implement the error correction capability in a "copy" mode, in which the video signal output provided to a second video recorder has been corrected so as to be an accurate replica of the original video signal. This ensures that negligible degradation of the digitized audio occurs with each successive generation of copy, an attractive attribute for database exchange.

While the PCM/VCR technology is inherently a two-channel recording system, for some database purpose it is desirable to record several audio channels. For speech databases, the additional audio channels can be used for the output of multiple microphones or other sensors, for comments, or for modem-signal-encoded digital data. Because the PCM/VCR data is stored as a video signal, use can be made of the VCR audio record capability. Conventional audio is recorded along the edge track(s) of the VCR tape with quality that is unacceptable for database purposes. The use of Beta Hi-Fi and VHS Hi-Fi audio capability of some VCRs provides an additional two channels of high quality audio which can be used concurrently with the PCM/VCR data that is encoded and stored on the video signal. No provision is made

for signal enhancement or error correction in copying the "Hi-Fi" audio, however, so that this audio data is more susceptible to degradation in copying.

STANDARDIZATION

The PCM/VCR technology relies heavily on standards developed for consumer electronics. Standards for digital audio developed by the EIAJ are implemented within PCM/VCR hardware. These specify the sample rate, quantization and error correction codes so as to ensure compatibility. The video signals are encoded according to NTSC (or PAL/SECAM) video standards as is appropriate.

Users of PCM/VCR technology for the purposes of developing standard research or benchmark speech databases have included the Speech Subcommittee of the JEIDA in Japan [1] and the developers of the United Kingdom's RSRE [2,3] and the French GRECO [4] databases, as well as in the United States.

In France, a special-purpose interface has been developed [4,5] for use in converting the digitized audio data from the EIAJ standard format to other sample rates and formats more suitable for processing and storage in a host computer (e.g., a VAX 11/750). The interface not only allows transfer of data between the host and the VCR, but a marker channel can be provided by sacrificing one of the PCM/VCR audio channels. This marker channel supplies 100 bytes of digital data every 100 milliseconds. Other software systems to permit transfer of the digitized data from the JEIDA digital audio format are reported to be in use in studios for electronic music [6]. The ability to make use of well established digital audio standards and widely available recording hardware and to convert recorded audio to formats more suitable for the signal processing operations performed in speech research is highly attractive. It may eliminate the need for redigitization of audio data, provide alternative low-cost storage media for speech processing work stations and free more costly D/A conversion systems for other uses.

FORMAT DESCRIPTION

A relatively simple and low-cost use of the PCM/VCR technology has been made in order to meet needs for a combination of high-quality recorded audio and a limited amount of digital data. One channel contains the speech material, while the second channel contains modem-encoded (1200 baud) ASCII character string header data typically associated with the speech material to follow on the other channel. Because the modem is relatively slow, these strings are

typically limited in length to 50 to 80 characters so as to provide only the essential data. Use of the PCM/VCR technology ensures that it is possible to make multiple-generation copies with negligible degradation. Transfer of the recorded material (particularly the modem-encoded data) to lower quality analog media such as reel-to-reel or cassette tape is discouraged because of the occasional loss of the ASCII characters. This is believed to be due to the sensitivity of 1200 baud modem link to flutter and wow.

Two widely-used speech databases are presently available on a loan-and-copy basis from the National Bureau of Standards in this format: the TI connected-digit speaker-independent database [7] (produced in this format by ITT/DCD [8]), and the TI isolated word 16-speaker database in the complete 46-word version [9] including the "alpha-set" or complete natural English alphabet (a,b,c,...z). Further information concerning the availability of these databases may be obtained by contacting the author.

Recently, the PCM/VCR technology has also been used by researchers at the Air Force Aerospace Medical Research Laboratory at Wright-Patterson Air Force Base [10] and at Texas Instruments Central Research Laboratory [11] in developing the DARPA Robust Speech Data Base. In this implementation, two channels of PCM/VCR audio are used for two channels of speech data and a third (VHS-or Beta-Hi-Fi) audio channel contains modem-encoded header data derived from prompt strings presented to test subjects. Since the header data precedes the speech data, simple procedures have been derived to largely automate the digitization of the recorded speech material and to associate appropriate headers with the speech files. Similar procedures are suggested for users of the PCM/VCR databases available through NBS for those users who wish to digitize the speech material in their own format.

VCR CONTROL

The use of VCR technology is widespread in the industrial audio-visual and broadcast community. The VHS format appears to be more widely used, and VHS units are available with provisions for remote control of many features. Using one commercially available interface between a VHS recorder and a host microcomputer, NBS has demonstrated control of the recorder to permit access of individual utterances on the PCM/VCR databases. On first playing of the VCR cassette, the tape is labeled with a dubbed audio track containing information on the video frame count, and files are built in the microcomputer relating the video frame count at the start of the

header data for each speech token to the header data itself. On a subsequent request to play a specific utterance, these files are used in directing the VCR to fast-forward or rewind the tape to the specific video frame count corresponding to that utterance. A major drawback of such a system is the considerable access time to locate widely-separated utterances. Nonetheless, it does provide a relatively low-cost means of accessing such a database when training recognizers requiring enrollment tokens in some order other than that found on the database.

RECOGNITION TEST SOFTWARE

In many tests of speech recognizers, a script is generated for the test material, and scoring is implemented through comparison of the script (string) and the recognizer's response (string). A variety of string alignment procedures are used to detect the presence of insertions or deletions. NBS has developed BASIC language programs to implement automatic real-time scoring of speech recognizers and to generate data files for subsequent data analysis. Realtime scoring is based on comparison of information contained in the modem-encoded headers with recognizer response, without reference to scripts or the need for complicated string matching procedures. In effect, the modem-encoded data generates a script in real time and advises the host microcomputer about the next utterance to be input to the recognizer. Data analysis includes conventional performance measures [12], confusion matrices and measures of distributions of scores. A portion of this software will be made available for use with the isolated word database as a contribution toward standardization of procedures for benchmark tests of isolated word recognizers.

REFERENCES

- [1] S. Itahashi, "A Japanese Language Speech Database", Paper 7.5 in Proceedings of ICASSP-86, (Tokyo), April 1986.
- [2] J. Deacon, R.K. Moore et al., "RSRE Speech Database Recordings (1983): Part I, Specification of Vocabulary and Recording Procedure", RSRE Report No. 83010, December 1983.
- [3] M.J. Russell, R.K. Moore et al., "RSRE Speech Database Recordings (1983): Part II, Recordings made for Automatic Speech Recognition Assessment and Research", RSRE Report No. 84008, May 1984.
- [4] R. Descout, (ed.), "Etat D'Avancement de la BD-Sons: Reunion GRECO/Industrie - Vendredi 21 Juin 1985", Centre National de la Recherche Scientifique GRECO

Communication Parlee, June 1985.

[5] Personal Communication with R. Descout, Centre National d'Etudes des Telecommunications and technical brochure of OROS-AI, Chemin des Clos, ZIRST-38240 MEYLAN, France.

[6] Personal Communication with Fred Clarke, Digital Sound Corporation.

[7] R.G. Leonard, "A Database for Speaker-Independent Digit Recognition", Paper 42.11 in Proceedings of ICASSP-84 (San Diego), March 1984.

[8] J. Porter, "The NBS/TI Connected Digit Database in PCM-digital Form", documentation provided to NBS accompanying this database from ITT Defense Communications Division, December 1984.

[9] T.B. Schalk, "The Design and Use of Speech Recognition Data Bases", Proceedings of the Workshop on Standardization for Speech I/O Technology, D.S. Pallett, (ed.), National Bureau of Standards, March 1982.

[10] Personal Communication with Tim Anderson, Air Force Aerospace Medical Research Laboratory.

[11] Personal Communication with Raja Rajasekeran, Texas Instruments.

[12] D.S. Pallett, "Performance Assessment of Automatic Speech Recognizers", Journal of Research of the National Bureau of Standards, Vol. 90, No. 5, September - October 1985.