Recording and storing of speech data

Nick Campbell

JST/CREST Expressive Speech Processing Project
ATR Human Information Science Laboratories
Kyoto 619-022, Japan
nick@atr.co.jp

Abstract

This position-paper accompanies the paper entitled "Recording techniques for capturing natural every-day speech", which is published in the LREC-2002 proceedings, and puts forward our current thinking for the recording and storage of large amounts of every-day conversational speech. There has been considerable discussion recently regarding the optimal choice of media for recording, and the problems of data compression. This paper reports the stance taken on these matters by the JST/CREST ESP Project. It should be read as an opinion-piece, rather than a report of scientific findings.

1. Introduction

The Japan Science & Technology Agency recently provided funding for a five-year project to produce speech technology interfaces for an "Advanced Media Society", under the auspices of CREST Core Research for Evolutional Science & Technology. The goal of this research is to provide the knowledge, software, tools, and databases for the development of spoken-language interfaces that are people-friendly. One of the sub-goals of the project is to collect 1000 hours of spontaneous interactive speech in the first three years and to spend the remaining two years on prototype development and system evaluation. To date we have collected more than 250 hours of speech and plan to collect the remaining 750 hours during the coming year.

It is essential that the speech be of high signal quality, so that automatic techniques for segmentation and annotation may be applied, and that it is at the same time representative of the full range of spoken behaviour that information-processing devices are likely to encounter in the near future. Since we need to capture speech in a wide variety of contexts, our early experiments concerned choice of microphone type and placement, along with choice of device for recording the speech data, with lightness and wearability being one of our main considerations.

Volunteers wear light, head-mounted, studio-quality microphones to record their every-day conversations to DAT recorders or portable minidisk recorders, sometimes via radio transmitters. Although the quality of the DAT recording is high, the recorders are relatively bulky, and the radio signal can suffer from range problems. Minidisc Walkman technology is considerably smaller and lighter than DAT, but makes use signal compression to reduce the amount of data to be stored on disc.

2. Speech data compression

The accompanying LREC paper [1] reports on tests to determine the extent to which traditional methods of e.g. voice pitch estimation, formant-tracking, and spectral analysis may be degraded as a result of using speech data which has undergone perceptual-masking for compression of the recorded signal. We compared speech signals recorded simultaneously from the same microphone to DAT and MD devices.

2.1. DAT vs MD – recording quality

In all cases, although the visible signals were perceptually equivalent, they were not identical. However, the differences in their spectra were limited to occasional valleys, and the structure of the spectral peaks can be considered almost identical. We noticed a greater difference between the two spectra in the area around 5.5kHz, and a noticeable difference in the peak heights at that frequency. We concluded that while there are undeniably differences in the speech signal between DAT and MD recordings, the derived estimates of formants, fundamental frequency, and glottal parameters reveal only small differences, and we maintain that the two recording media can be considered as equivalent for the purposes of prosodic analysis. Informal listening tests, switching between the two sources, confirmed that the recorded speech of both media sounds the same to the ear even when played over high quality headphones. The difference in recording quality between the two media becomes obvious when listening to music, but in the frequency range of human speech it can be considered imperceptible.

2.2. DAT vs MD – compression

The default sampling rate that we use for our DAT speech recordings is 48kHz, although we usually downsample the resulting data to 16kHz for analysis and storage. The MD recorder is switchable between 44.1kHz and 32kHz (LP mode) though it can also record at 48kHz from digital input.

There may be no gain in data storage space from the use of MD recordings. This is because although the signal undergoes considerable compression internally, it is always decompressed for playback, even when using optical fibre cables for "direct" digital output. The laws preventing unauthorised copying of music have resulted in hardware that prevents easy access to compressed speech data as well.

Since we have not yet found a simple way to extract the raw compressed data directly from the minidisk (though for interesting insights, we recommend [2] to the adventurous reader), the main advantage that we see for the use this recording medium is the lightness of the recorders and size and availability of the storage media.

The tonal separation technology introduced with the ATRAC3 encoding is particularly effective for audio signals like speech, in which the energy is concentrated in a relatively small number of frequency components, and ensures a high signal-to-noise ratio. While it may be less effective for rapid acoustic transitions (from percussion instruments, for example) it appears to preserve the plosive sounds of speech without noticeable problems.

Software is available both to compress linear PCM files to ATRAC3 format (e.g., Goldwave [3]), and vice-versa (e.g., [4]), for an ATRAC3 encoded file to be replayed on the computer, but since disc space is now so cheaply available, we feel little need to advocate its use for the archiving of speech data.

2.3. MD vs. MP3 – compression

There has recently been considerable discussion of the differences between MD and MP3 (see for example [5,6]). The advent of cheap and ultra-light MP3-based stick recorders and data encoders may be seen as an advantage if they are to be distributed to a large number of people for recordings in the field, but whereas the MP3 encoding potentially offers considerably more compression of the data than MD recorders (1:11 for MP3), we find the differences in quality to be unacceptable for our purposes.

The MP3 devices were designed for the portable music market, and the stick-recorders for business meetings and memos. They are optimized for use with light ear-speakers, and are not designed to be interfaced with high-fidelity audio components. There have been reports of considerable distortion (especially loss of sharpness and of bass) when this has been tried.

However, MP3 compression may have an advantage in media-streaming, over poor-quality or low-bandwidth lines. It has good potential for wide distribution of sample speech data, such as from a web-page, in situations where the listener is more concerned with listening to the content of the speech than with an analysis of its characteristics.

3. Recommendations

To those who are considering recording speech data in the field, we offer the following recommendations. They are based on our limited experience, and are intended as a basis for discussion, rather than as firm or fixed guidelines.

3.1. A compromise - quality vs. convenience

There is a tradeoff between quality and convenience. DAT recorders offer the highest quality and are light and portable enough to be carried easily, but they are not yet light enough to be carried in the pocket or worn un-noticeably on the body. MD recorders are lighter, smaller, and considerably cheaper. The discs are more widely available than DAT tapes, thanks to their popularity for music recording. The advantages of the random-access playback, track-marking, track division, and re-joining are considerable.

Whenever possible, we much prefer the use of DAT for recordings, but more than half of our recordings to date have been made using the more convenient MD. Similarly, we prefer head-mounted microphones, because they offer much clearer and consistent reproduction of the speaker's voice, but there are occasions when a far-field microphone must be used. This results in at least 4 levels of data quality. More are introduced by the use of portable radio-frequency transmitters remotely linked to the DAT recorders. Yet to date, we have not noticed a drop in recording quality that has been serious enough to prevent acoustic-prosodic analysis of the speech signal. Formant and pitch extraction is unaffected, and even voice-quality estimation from the derived glottal waveform appears to be effective. Much more variation has been found to result from poorly set recording levels or from bad microphone placements than from the difference due to recording medium.

It is, however, essential to differentiate between the different recording combinations, and we take care to note the settings and hardware combinations for each recording.

3.2. Archiving & distribution

There must be redundancy in archiving. Since the original recording media should be considered subject to decomposition or degradation over time, we make backups to DVD from the reconstituted audio signal for off-line storage, and then copies to disc (after downsampling to 16kHz) for interactive use.

The Memory Stick Walkman uses the same ATRAC3 signal compression as the MiniDisc, and the small chewing-gum sized stick currently holds up to 128 megabytes of data. While it is not capable of recording, we find it a useful device for data transfer and offline listening. Up to three copies of the original data can be made without resorting to breaking the copy-protection.

4. Future work

One advantage of MP3 encoding that we have not yet fully explored is its potential for random-access or streaming audio. By use of this technology, the listener can listen to the speech while it is being downloaded, without waiting for the whole file to be received before playing can begin. This offers clear advantages for the distribution of data over the internet. Conventional ftp file transfer can take several hours, and the use of postal services several days, before the recipient is able to listen to the data. Streaming makes the speech data available almost instantaneously, and there is no shortage of available software to facilitate this service.

There are not yet clear standards for the recording of very-large natural-speech corpora, nor are there protocols for the distribution and storage of the resulting data. This short paper has presented our limited experiences in the hope of encouraging active discussion, experimentation, and debate about the methods and resources we should use.

5. Acknowledgements

Part of this work was sponsored by the JST/CREST Project Number 131. The author is grateful to Dr. Parham Mokhtari for ongoing discussions and advice, and to Peter Wittenburg for his encouragement..

6. References

[1] "Recording techniques for capturing natural every-day speech", Nick Campbell, in Proc LREC 2002.
[2] http://www.minidisc.org/part_hacking.html
[3] http://www.Goldwave.com
[4] http://www.minidisc.org/atrac3.zip
[5] http://www.minidisc.org/atrac_vs_mp3.html
[6] http://www.minidisc.org/hifichoice_april2000.html