Recording techniques for capturing natural every-day speech

Nick Campbell

ATR Human Information Science Laboratories
Kyoto 619-022, Japan
nick@atr.co.jp

Abstract

This paper describes techniques for the collection of natural spontaneous speech from daily conversational interactions for a large corpus that is currently being produced by the Japan Science and Technology Agency. This corpus will form the basis for further development of tools and software for the improvement of concatenative speech synthesis and for the development of spoken-language interfaces for information-providing devices that will be sensitive not only to the content of an utterance, but also to the manner in which it is spoken, so as to be able to detect speaker emotions and attitudes.

1. Introduction

There is now a growing interest in modelling the characteristics of emotional speech and attitudinally-marked speech, but as the majority of presentations and the few notable exceptions at the recent ISCA Speech & Emotion tutorial and research workshop [1] showed, there are still very few corpora of real spontaneous speech being used in current research [2].

This is may be due to an 'observer's paradox' effect, in that as soon as a person realizes that there is a microphone in front of them, they switch into a public or self-conscious mode of behaviour in which the expression of emotion is considered taboo, or at least potentially embarrassing.

As a result, we have little knowledge of how people actually perform spoken language in the wild, or of how they express their feelings and attitudes through variations in speaking style. We have subjective experience, but little data, and the majority of analyses of emotional speech, for example, are still based on the intuitive performances of actors or of subjects simulating emotional speech in the reading of otherwise balanced texts.

This paper reports on our experiences in trying to produce a corpus that includes representative examples of spontaneously emotional speech and on the techniques that we developed in order to overcome the observer effects on the naturalness of the speech..

2. Expressive Speech Processing

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. To improve speech synthesis so that it can emulate the variety of ways that people use tone-of-voice and speaking style to convey non-verbal information, and to develop speech recognition modules that are able to distinguish between different intentions displayed on the same underlying text, in order to detect how the speaker is relating to the spoken discourse to facilitate interaction with information-providing devices.

Although the project started out aiming to study emotional speech, it soon became obvious from the initial data that we collected that the expression of speaker state and attitude is by no means linked to emotion alone, and that there is a wide range of speaking styles which are used to show commitment of the speaker to the utterance and to indicate the relationship with the listener [3].

2.1. Project goals

This research is being carried out in conjunction with several university labs and includes modules for discourse analysis, speech processing, linguistic structure, interface design, and system integration. Their common link is a large corpus of expressive speech, which forms the core of the analysis and development. The focus is on improving concatenative speech synthesis, both in quality and in range of speaking styles.

The goal 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. If these devices are to be used in domestic as well as business environments, then they will likely be exposed to 'private-mode' speech.

Our schedule for speech data collection was originally as follows: 10 hours in the first year, 100 in the second, and 1000 in the third. We assumed that each year's data would replace that of the previous year as we learnt from our mistakes. In fact much of the early data can still be made use of, although we now have a better understanding both of the collection techniques and of the types of speech that we wish to collect.

Most of our early experiments concerned choice of microphone type, and its placement with respect to the speaker, along with choice of device for recording the data, with lightness and wearability being one of the main considerations, since we need to capture speech in a wide variety of contexts.

2.2. Speech data collection

In order to capture natural speech, it is important to impose as few constraints on the speaker as possible.

When previously recording speech corpora as a source of units for concatenative speech synthesis, we prepared balanced texts that ensured even coverage of all phone combinations in most prosodic environments, but the resulting sentences, being generated by a greedy algorithm, were lexically dense and phonetically complicated for the reader to produce. The resulting stress in the voice remained throughout the speech synthesis process, and the results were less than satisfactory to listen to.

Developments in the synthesis corpus recording led to the use of longer texts such as novels or short stories, which had simpler and sequentially-related sentences but which, when read for long enough, provided similar prosodic and phonemic balance to the previously-used sentence lists. The stories, however, had the advantage of producing much more relaxed and natural-sounding speech.

The resulting corpora were both natural-sounding and phonetically/prosodically balanced, but they exhibited the characteristics of only one fixed speaking style. If the source text was sad, for example, then the whole corpus would be read in a sad voice, and any synthesis produced using that speech would also sound sad. Synthesis of e.g., a weather forecast using a sad voice can introduce a level of interpretation of the text that is quite different from what was intended.

By extension, if we are to collect natural-sounding speech, then we need to widen the task-specification still further to include all the varieties of speaking style encountered in daily interactive communication. When we consider the number and types of interlocutor typically encountered in a day's interaction, then that becomes a very wide range indeed. Our current goal is therefore to produce recordings of daily spoken activity, from morning to night, including interactions with family members, friends, strangers, traders, and casual acquaintances.

2.3. ESP Recording

Volunteers wear light head-mounted studio-quality microphones, suspended from the ears and largely hidden by the hair. They take time to become accustomed to these and learn to adapt their behaviour so that they do not accidentally knock the microphone or drag the cable to create unwanted mechanical noise.

Some volunteers use radio transmitters, some connect directly to DAT recorders, and some to a portable minidisk recorder. Although the quality of the DAT recording is high, the recorders are still bulky, and the radio microphones can suffer from range problems.

3. Comparing DAT & MD

Although DAT recorders are now very small, they are not yet pocket-sized, and are still too heavy to wear comfortably on the body during everyday activities. Portable Minidisc Walkman technology is considerably smaller and lighter, but makes use signal compression to reduce the amount of data to be stored on disc.

The ATRAC perceptual-masking compression [4] used in the Sony Minidisc recorders may render the recorded speech unsuitable for conventional signal processing techniques. We therefore carried out tests to determine the extent to which traditional methods of e.g. voice pitch estimation, formant-tracking, spectral analysis, and cepstral encoding may be degraded as a result of using speech data which has undergone perceptual-masking for compression of the recorded signal [5].

3.1. Microphone atttachment

The Sony ECM-77B studio-quality miniature lavalier microphones are provided with a 3-pin XLR plug about 15 cm in length, which contains a type-3 1.5V battery and transformer circuitry to power the microphone. By substituting this with a small 1.5V SR44 camera battery and holder, the power unit can be made small enough to fit into the extension battery pack of the SONY MZR900 MD Walkman. Fitting the microphone into a Sennheiser NB2 adjustable ear-mount provides a comfortable, light, portable, and unobtrusive recording unit.

To measure the difference between recording quality on DAT (Digital Audio Tape) and Minidisc, we used the above microphone arrangement to record a 5 vowel sequence (a-i-u-e-o) from a male and a female speaker, taking the signal to a DAT recorder (*Sony DAT TCD-100*) and a Minidisc recorder (*Sony MZ-R900*) simultaneously.

We also recorded a 1kHz-10kHz sweep tone and a 200Hz-800Hz chirp tone with a sinusoidal waveform produced by an NF Electronic Instruments DF-194A variable phase digital function synthesiser. The recording levels of the two devices were adjusted to an approximately equivalent setting using these tones. The signals were transferred directly to computer disc using optical fibre via a Canopus MD-Port, and down-sampled to 16kHz 16-bit using Wavesurfer software [6]. Both Wavesurfer and Entropic's ESPS software [7] were used for pitch-estimation, spectral display, and formant analysis. We used the HTK-3.0 program [8] to calculate the Mel-scaled cepstral coefficients..

3.2. Comparison of the recordings

In all cases, the visible signals were perceptually equivalent but not exactly identical (compare Figures 1 and 2), nor were the numerical values identical. Table 1 shows F0 and formant statistics. Since the start points of the different waveforms were aligned manually and processed using identical (default) settings of the software, we could expect the values to be identical, except for small differences in signal power arising from differences in the recording levels of the two devices. Figure 3 shows very close spectral similarities, revealing identical peaks, but with slightly less energy in the troughs for the MD data.

Table 1. Comparison of prosodic (top: fundamental frequency) and spectral parameters (bottom: formants and their bandwidths) derived from signals recorded simultaneously from the same microphone to both DAT and Minidisc (MD) recorders.

	Male		female	
	Mean f0	Sd	Mean f0	sd
DAT	98.69	9.77	171.06	34.7
MD	98.73	9.68	169.31	38.6

	F1	F2	F3	F4
DAT	701 (371)	1615 (390)	2726 (451)	3771 (403)
MD	678 (365)	1603 (380)	2683 (455)	3750 (402)
	B1	B2	B3	B4
DAT	336 (273)	389 (217)	451 (253)	493 (224)
MD	336 (268)	392 (244)	439 (237)	478 (237)

We can see from Table 1 that although the values for the voice fundamental frequency are not identical for the data recorded simultaneously to both DAT and MD, they do fall within approximately the same range. Similarly for the formants and their bandwidths. Thus, if the purpose of this comparison were to prove that there is no difference between the two media, then we could conclude immediately that this is not the case. However, our goal is to determine whether speech data from the two media can be treated equivalently, and it appears that the differences, mainly in the range of a few Hz, can be considered as both being different but similarly accurate estimates of the underlying prosodic processes.

Similarly, Figure 3, which plots two spectral slices overlapping, shows that the differences in the spectra are limited to occasional valleys, and that the structure of the spectral peaks can be considered almost identical. These spectral sections were taken from the same point in the vowel /a/, at a steady part of the vowel center. figure 4 shows a pair of spectral slices (as above, one from the DAT recording, and one from the MD) but here taken from a more active part of the speech signal. We notice that there is a greater difference between the two spectra, particularly in the area around 6kHz, and perhaps a slight difference in the peaks at around this point, but this example represents the biggest level of difference that we found in the test data.

Thus, if our goal is simply to use the prosodic or spectral information derived from the speech signals, we can freely use data recorded on either medium. However, as we shall see below, the numerical differences become significant in the case of cepstral transformed data.

Table 2. Mean-squared differences in cepstral distance between pairs of vowels, for MD and DAT speech data. Top: same media, different vowel; bottom: same vowel, different media.

	/i/ - /a/	/i/ - /u/	/a/ - /u/	mean
MD-MD	0.0381	0.0297	0.0267	0.0315
DAT-DAT	0.0422	0.0268	0.0261	0.0317
	/i/ - /i/	/a/ - /a/	/u/ - /u/	
MD-DAT	0.0207	0.0186	0.0060	0.0151

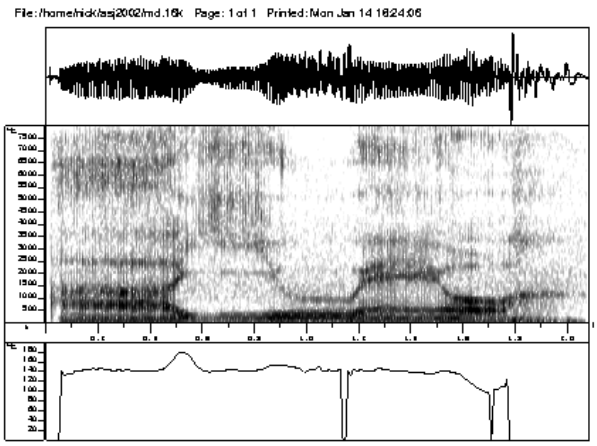
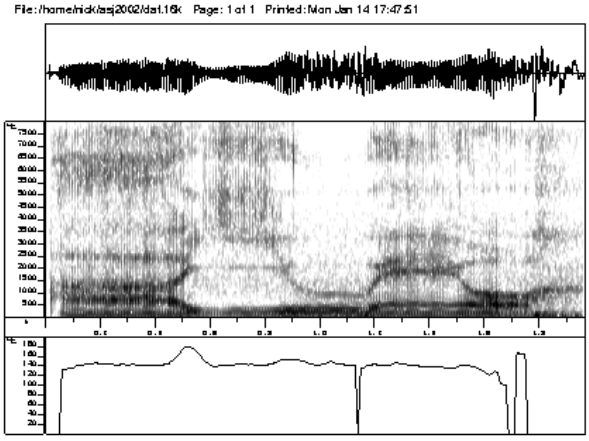


Figure 1. Spectrogram and f0 of the 5-vowel sequence (from the male speaker) recorded using a Minidisc.



.Figure 2. Spectrogram and f0 of the 5-vowel sequence (from the male speaker) recorded on DAT.

3.3. Cepstral differences

Table 2 shows differences measured from cepstra calculated by HTK's HCode program from the same DAT and MD-recorded speech data, using default parameters to produce 16 Mel-scaled cepstra per 10 msec of speech. The cepstral distance is frequently used in speech recognition to measure similarity between speech signals, e.g., for automatic speech segmentation or phonemic alignment.

Since the cepstral distance measure in itself can be difficult to interpret, we established a baseline cepstral distance for each medium by taking measures between pairs of vowels; /a/ - /i/, /a/ - /u/, and /i/ - /u/, and then averaging this distance to produce a value which might represent the typical inter-vowel distance. This (possibly unitless) distance is about 0.03 in the case of both DAT and MD.We then measured cepstral distances between same-vowel cepstral sequences, but varying the coding; DAT vs. MD (see bottom row of table 2). The calculated distances were different depending on the type of vowel, greatest for /i/ and least for /u/, but they averaged 0.015. Since this is half the inter-vowel distance, there is probably cause for some concern.

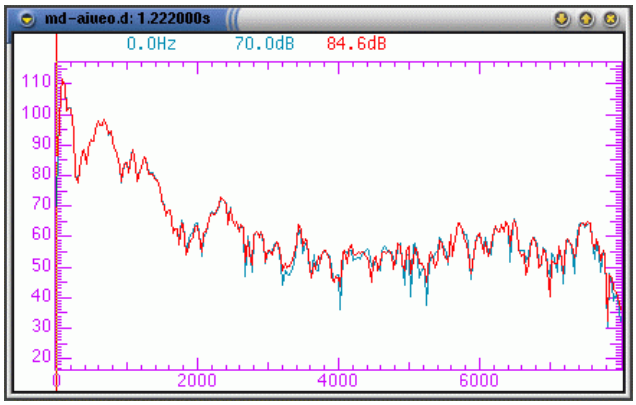


Figure 3. Spectral section through the vowel /a/. The plot shows both DAT and MD data overlapped.

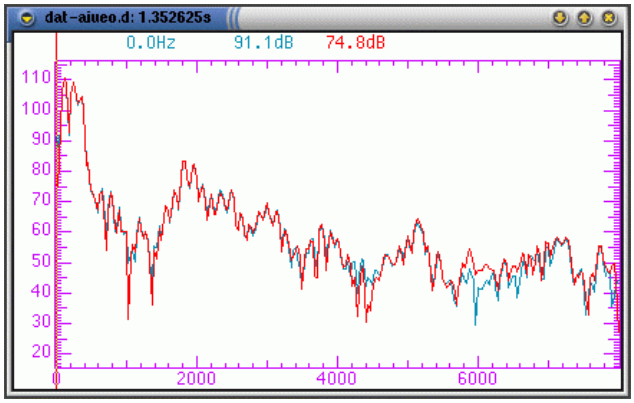


Figure 4. Spectral section through the diphthong /ai/. The plot shows both DAT and MD data overlapped.

3.4. Discussion

There are undeniably differences in the speech signal between DAT and MD recordings, but the recorded speech of both sounds identical 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.

Although the numerical data of our comparison reveal differences, presumably arising as a result of the perceptually-masking-based compression used in the MiniDidc, the derived estimates of formants, fundamental frequency, and glottal parameters reveal only small differences and the two recording media can be considered equivalent for the purposes of prosodic analysis.

4. Microphone placement

In many cases the signal-to-noise ration is more important for speech signal processing than compression, so we consider that as long as the microphone placement is correct, and the type of microphone used is sufficiently powered, then we will be able to collect data which will not only serve our own purposes for speech synthesis but also benefit the wider community of speech researchers.

We experimented between head-mounting and lapel-mounting for the lavalier microphones, but concluded that although lapel-mounting is more widely used in studio recordings, there is too much head movement in daily-interactive speech and the mic-to-mouth distance changes considerably unless head-mounting is used.

We experimented with two forms of ear-mounted setting – using a boom (e.g., Sennnnheiser NB2) or attaching directly to the ear as an earring. The latter method is much less obtrusive (and quite comfortable) but results in a drop in high-frequency speech signal and in spite of the inconvenience, we now prefer the ear-mounted boom arrangement. With careful placement to avoid nasal airflow and oral plosive air-bursts, it can produce a high-quality constant-level signal that allows faithful recording of the subject's speech.

We are still undecided between dynamic and condenser microphones however, as each has both good and bad points, but by reducing the size of the power supply, it has become much easier to make more use of the latter.

5. Conclusion

This paper has described our techniques for collecting spontaneous speech data and presented the results of a comparison between DAT and MD recorded speech. We found that although there are differences as a result of the signal compression, the MD speech data can be considered equivalent to that collected on DAT tape for use in large-scale recordings. We are using Minidisc recorders with high-quality head-mounted microphones to collect continuous spontaneous conversational speech from a range of volunteer subjects throughout Japan.

6. Acknowledgements

Part of this work was sponsored by the JST/CREST under Project Number 131. The author is grateful to an anonymous reviewer for helpful comments. A shorter version of this paper was presented at the Spring Meeting of the Acoustical Society of Japan in March this year.

7. References

[1] ISCA Tutorial and Research Workshop on Speech and Emotion, Belfast 1999 : www.isca-speech.org/archive.
[2] Douglas-Cowie E, Campbell, N., Cowie, R., Roach, P. "Emotional speech: towards a new generation of speech databases", in Speech Communication Special Issue on Speech & Emotion, forthcoming.
[3] Campbell, N., "Collecting really spontaneous speech", iProc Mombusho w/s on Speech Prosody, Tokyo, 2002.
[4] ATRAC : www.minidisc.org/aes_atrac.htm
[5] Campbell, N & Mokhtari, P., "DAT vs. Minidisc - Is MD recording quality good enough for prosodic analysis?", Proc ASJ Spring Meeting 2002, 1-P-27
[6] Wavesurfer: see http://www.speech.kth.se/wavesurfer
[7] Entropic ESPS Signal Processing Software – no longer available after being bought by Microsoft
[8] HTK Speech Recognition Toolkit, web pages and download site at htk.eng.cam.ac.uk