Collecting Really Spontaneous Speech

Nick Campbell

ATR Human Information Sciences
JST/CREST Expressive Speech Processing

nick@atr.co.jp

Abstract

This paper reports on recent developments funded by the Japan Science and Technology Agency, for the creation and analysis of very large databases of emotional and attitudinally-marked speech for the support of research into concatenative methods for producing synthesised speech which is capable of expressing the range of speech prosody required to emulate interpersonal human spoken interactions. We discuss the problems of ensuring high spontaneity in the speech while at the same time producing data that is of high enough quality to allow signal analysis by automatic processing techniques.

1. Introduction

There is a growing interest in the study of speech prosody, not just in order to further scientific and linguistic knowledge, but also to facilitate advances in speech technology. With the increasing use of internet-connected portable telephones, more applications will make use of speech to enable people who are not yet computer-literate to access the growing number of internet-based information and entertainment services through voice-based dialogue-type interactions.

In order to facilitate the study of speech prosody, the research community is collecting corpora designed to be representative of the varieties of speech found in a wide range of everyday situations. This speech should not be limited to the received pronunciation of radio announcers and trained professionals, but should also include examples of local dialects and, especially, the speech of ordinary people, expressing various attitudes and emotions, in a variety of interactive situations.

Only by recording really spontaneous vocal interactions can we build a representative knowledge-base of the types of non-verbal information typically signalled by the prosody of interactive speech. Yet at the same time, it is also necessary to define a clear model of the categories of prosodic variation that signal linguistic information, in order to provide a framework against which the paralinguistic variations of spontaneous speech prosody can be understood.

2. The Observer's Paradox

It is well known that, in many situations, the presence of an observer can have an influence on that which is being observed. Spontaneous speech is no exception. The presence of a microphone (or worse, of a recording engineer) can severely hamper the spontaneity of the speech being produced. Similarly, the need for a balanced scientific design frequently places unnatural requirements on the speech corpus, which render the content less than spontaneous. We can find many examples of such contrived-speech in the literature and in the corpora that are currently being used for scientific and technical analysis.

The problem addressed in this paper concerns the design of corpora of spontaneous speech, the methods that can be used in their collection, and the techniques that can be used in their analysis. We suggest an analogy with another recording medium, photography, and recall that in 1970 a team of photographers took 1000 rolls of 36-exposure film on location to an island in the Pacific in order to produce a calendar, which was expected to contain only twelve (glamour) images. Extending the analogy; if we can record an almost infinite amount of speech, and develop techniques for processing it, in order to extract only the significant or interesting portions for further analysis, then we should be able to gather a corpus which is both truly representative and of sufficient coverage to allow us to define the full range of prosodic variation and its use in human communication.

If the goals were limited to the linguistic uses of prosody, then it would be sufficient to design a corpus which is based on read texts or prepared utterances. These minimise the speaker's personal involvement, and their attention is focused instead on presenting the form of the text, by restructuring the two-dimensional graphical information, through a media-transform, into a one-dimensional vocal representation. The role of prosody in this case is simply to represent the syntactic and semantic relationships among the component words through the use of timing and pauses, and tones and accents. Interrogative and declarative forms are determined not from any desire of the speaker to provide or request information, but instead from the punctuation.

Given and new information and focus information are limited to what can be inferred from the text alone, since the speaker is usually not the originator of the message, but just its interpreter.

Task-based speech collection requires more speaker-involvement, but this is reduced to a linguistic minimum since the speaker is usually not motivated from internal desires, but by the need to perform as requested. By use of task-based techniques, we can elicit prosody that signals not just the bare linguistic framework but also pragmatic function, since, in a dialogue situation, the listener is as involved as the speaker, and a request for information (for example) must be signaled as such in order to obtain a reply without scripting of the speech.

However, in neither of the above scenarios is the speaker acting spontaneously. Even if the speech itself is, by nature of being unscripted, spontaneous, the situation is contrived, and the speaker is cooperating rather than operating. In order to collect a corpus for the analysis of speech prosody in paralinguistic functional mode, we need to devise ways of observer-free recording.

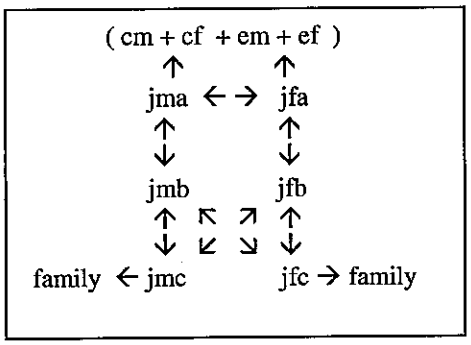
2.1. Unobtrusive recording

Fortunately, the popularity of walkman-type devices provides us with cheap and efficient recorders. The SONY minidisk, for example, weighs less than 135 grammes, with battery included, and measures only 8 x 7.5 x 2 cm. It can record continuously on monaural setting for 160 minutes on a 7 cm disc at 44.1 kHz, and has a wow and flutter of close to zero. These devices are easy to carry unobtrusively on the body, and can be used in conjunction with a high-quality ear-mounted close-talking microphone to provide studio-quality recordings of speech in a variety of contexts [1].

The JST/CREST Expressive Speech Processing [2] project has started collecting spontaneous speech from paid volunteers in order to produce a corpus of "street-speech", to complement the "laboratory speech" of previous studies. It currently employs two classes of informant, for short-time and long-time recordings. The former consists of a group of about ten people who telephone each other at fixed weekly intervals for half-an-hour of free unprompted conversation per session; the latter consists of individuals who have agreed to wear lightweight recording devices while going about their daily work and social interaction.

The telephone recordings of the first group are balanced for familiar and unfamiliar pairs, and for age, sex, and nationality of the speakers. The conversations are all held in Japanese, but the respondents include non-native-language speakers. Only the data from the Japanese native-language speakers is currently intended to be used for analysis, but all recordings will be preserved. The pairing arrangements are as shown in Table 1. Recordings are being made over a period of ten weeks, so that the speaking-style correlates of changes in familiarity between the speakers can also be studied.

Table 1. *Speaker combinations for regular weekly telephone conversations (the first letter j/c/e indicates the native language of the speaker: Japanese, Chinese, English respectively; the second letter the sex of the speaker, m:male, f:female, and the third letter the group identity. Each speaker interacts with at least 3 other speakers, but Group A experiences cross-cultural difficulties, while Group C interacts with family members. Group B provides the baseline comparison.*



For the long-time recordings of daily social interaction, we have currently collected 50-hours of speech from each informant, and this is now being transcribed as a first step towards automatic segmentation and feature extraction. Because the speakers quickly become accustomed to wearing the lightweight recording devices, their speech is highly natural and typical of normal daily conversations. The recordings are always made in familiar surroundings and with known interlocutors, and are of course completely unscripted.

The use of close-talking ear-mounted microphones ensures that only the voice of the target speaker is captured well – that of the interlocutor is often barely perceptible – so the confidentiality of the discourse can be assured. Because their interlocutors have not signed release agreements, only one side of each conversation can be analysed, but since our goal is the analysis of prosody, not of conversation, then this data can be considered satisfactory. By having the speakers transcribe their own conversations, we also allow them the right to remove any portions of the recordings that they consider to be too personal to be made public, though, to date, few such deletions have been requested.

2.2. Unscripted speech

Our informants have all expressed concern that the data they are providing must be too repetitive to be of any

use for analysis. On the contrary, although the text of the discourse may be limited, the prosodic variation that has been revealed is remarkable. There are indeed many repetitions, at the text level, but they are each recorded in different contexts, and each utterance can be considered unique in that it reveals a different relationship with the interlocutor. Much of the speech is non-verbal – laughs, grunts, and simple one-word utterances are common. The speech is also proving very difficult to transcribe using standard orthography. Yet this is the way that ordinary people speak in everyday situations. It is becoming clear to us that our linguistic concepts of the potential of the language (Chomsky's "competence") are strongly influenced by text-based traditions, and that corpora recorded after careful design may not be at all representative of this "street-phonology". The data that are now being collected require novel methods of transcription and novel methods of description. A new "grammar of spoken language" is required, and it will almost certainly be very different from that which dictates the way we can form written sentences in a paragraph on a page to be read. The grammar of spoken language cannot be written without an encoding of the prosody of the speech, and of the pragmatic function of each utterance in an interactive discourse.

3. Analysis of the data

In addition to the transcription difficulties noted above, we are also concerned that the ease of recording using the new media may come at a cost to speech signal analysis. The ATRAC (perceptual-masking-based) compression [3] 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.

Our findings are not yet complete, but we are currently satisfied that every technique we previously used in the prosodic analysis of speech can be used equivalently with speech recorded on Minidisc. However, there are definite numerical differences in the signals, compared to those recorded under similar conditions using the heavier DAT recorders, and the following section reports our findings to date. Further results will be reported in [4].

3.1. DAT or Minidisc?

To measure the difference between recording quality on DAT (Digital Audio Tape) and Minidisc, we used a single condenser microphone (*Sony C-355*) 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 [5]. Both Wavesurfer and Entropic's ESPS software [6] were used for pitch-estimation, spectral display, and formant analysis. The ESPS software was used for an 18-pole LPC analysis in order to produce a representation of the glottal signal by inverse filtering, and a synthesised waveform was produced from the analysis coefficients.

Table 2. *Comparison of prosodic (top: fundamental frequency) and spectral parameters (bottom: formants and their bandwidths) derived from signals recorded simultaneously on 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)

In all cases, the visible signals (or audible in the case of resynthesis) were perceptually equivalent (compare Figures 1 and 2), but the values were not identical. Table 2 shows F0 and formant statistics. Since the start points of the different waveforms were aligned manually and the processing used 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 spectral similarities, revealing identical peaks, but with slightly less energy in the troughs.

3.2. Speaking-style analysis

The remaining work is for the (semi) automatic analysis of this recorded speech in order to produce a relatively small subset that will be of most use for further prosodic research.

Previous work (e.g., [7]) has shown that tension of the glottis (or breathiness of the voice) correlates well with differences in speaking style. We have developed algorithms to label unknown speech with this feature and

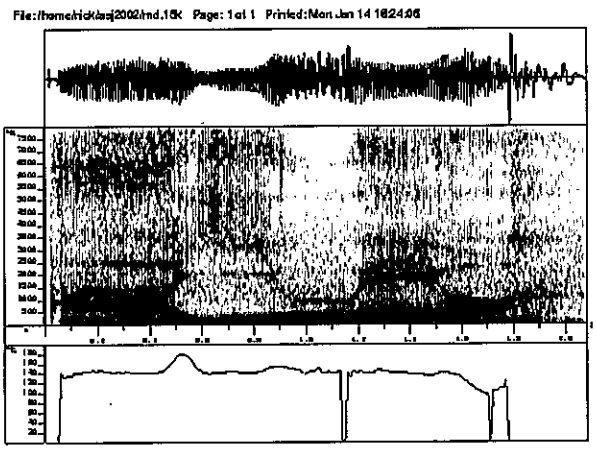
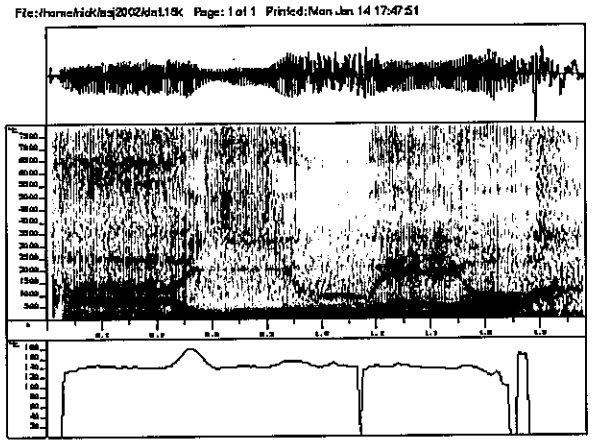


Figure 1. Spectrogram and f0 of the 5-vowel sequence (from the male speaker) recorded using a Minidisc.



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

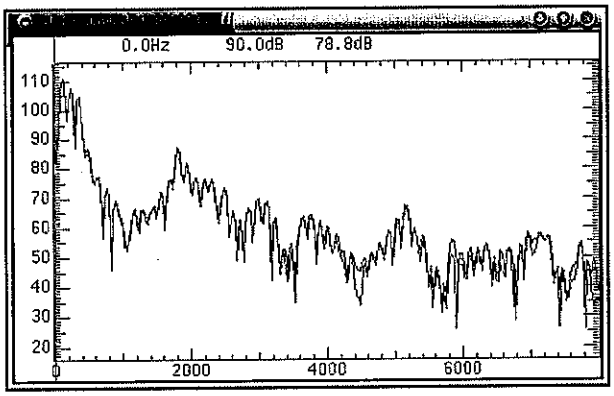


Figure 3. Spectral slices of vowel /a/ (both DAT & MD).

can now identify individual sections of the speech according to breathiness of the phonation, in combination with the more traditional prosodic features such as duration, amplitude, and fundamental frequency.

Manual transcription of all the speech will still be necessary, for indexing and for reference, but from that starting point we can estimate the phonetic sequences and determine the prosodic characteristics of each portion of the speech by automatic processing.

It remains as future work to classify the different speaking styles according these parameters and to select portions from the larger corpus to retain in the final production corpus, and for more detailed analysis.

4. Using spontaneous speech prosody

We expect that voice-interfaces which make use of the prosody of spontaneous speech will be considered by the general user as more friendly and personal. We plan to use the above corpora in concatenative speech synthesis systems, so that the variety of speaking styles and the increased ability to express attitude and emotion will enable a greater flexibility in the way that the machine talks to the customer. For example, if a regular user of an information-providing service can be recognized by the system, then its spoken output can be adapted to a style that is appropriate for the relationship and the interaction. Initial prototypes are now being tested.

5. Conclusions

This paper has argued that a corpus for prosodic research should include samples of really spontaneous speech, and has discussed some of the reasons why the collection of such speech can be very difficult. It has proposed methods and presented examples of both collection techniques and analysis methods, showing that convenient devices exist and can be used at little cost for high-quality data collection. Since the linguistic uses of speech prosody represent just a small part of the overall role of prosody in spoken interaction, we as a community should consider the collection and analysis of more representative types of data as an essential foundation for future research.

6. References

[1] Campbell, N., "The Recording of Emotional speech; JST/CREST database research", in Proc LREC 2002.

[2] JST/CEST ESP Project web page : www.isd.atr.co.jp/esp

[3] ATRAC compression: www.minidisc.org/aes_atrac.html

[4] Campbell, N., "DAT vs MD – is MD recording quality enough for prosodic analysis?", 1-P-27, Proc ASJ Spring Meeting, 2002.

[5] Wavesurfer: see http://www.speech.kth.se/wavesurfer

[6] Entropic ESPS Signal Processing Software – no longer available after being bought by Microsoft.

[7] Mokhtari, P., Iida. A., Campbell, N. "Some articulatory correlates of emotion variability in speech: a preliminary study on spoken Japanese vowels", *pp431-436 in Proc ICSP 2001, Seoul, Korea.*