tap2talk : an interactive interface for large speech corpora

Nick Campbell

ATR Human Information Science, JST/CREST
"Keihanna Science City", Kyoto 619-0223
nick@atr.co.jp

Abstract

This paper describes a software interface for the interactive display of speech segments from a very large speech corpus. The interface also has applications as an interactive speech synthesiser, given that the corpus probably contains many examples of much that the user typically needs to say in routine everyday conversations.

1.Introduction

The JST/CREST Expressive Speech Processing Project [1] has now collected more than 500 hours of unconstrained spontaneous speech from a range of subjects using two collection paradigms. The first is completely uncontrolled for content, with volunteers telephoning each other at weekly intervals to talk freely for half-an-hour per session over a period of ten weeks [2]. The second employs volunteers who record their daily spoken interactions for extended periods throughout each day [3]. This paper describes an interface for viewing either type of corpus, but is especially concerned with an extension of interactive access for the second.

The goal of the ESP Project is two-fold; a) to provide a knowledge-base for research into speech and emotion, or "expressiveness", and b) to provide a source database for expressive speech synthesis. Our original expectation was of a synthesiser capable of "emotional speech", but experience with the corpora suggests that the variety of speaking-styles encountered in daily conversational interactions requires more than just a control for "emotion" alone.

The problem this paper addresses is that of accessing speech data from within such a large corpups, both from the standpoint of linguistic (and paralinguistic) analysis, and of speech synthesis. Samples must be extracted selectively by use of filters or keys, in much the same way as using a search engine to browse the internet. The filters are required to reduce the amount of information, but we propose that they can also be used as input for a conversational speech synthesiser.

2. Text & Prosody

Speech is defined by both its text and its prosody, so the speech-database access/display module is initially text-based, with filters for parts-of-speech, lexical words, and tables of user-defined labels (such as emotion-type, addressee, speech-act, etc) that might facilitate research into the corpora. The results are displayed as clickable points in a two-dimensional window to allow finer prosodic selection. See Figure 1 for an example. The prosodic functions on the X and Y axes of the display can be selected from a menu offering f0 (mean, range., max, min, etc), duration (linear or log) and speaking rate, power (as f0), and voice-quality (from AQ and normalised AQ) [4].

A further selection filter (MORA) allows limits on the length of an utterance; for example, after selecting a lexical item (the 3-mora Japanese expression "honma"is selected in Figure 1) either the exact number of mora (3) or a maximum number of mora can be selected (0 is a wildcard setting allowing any length utterance that includes the target item). By specifying more morae than the length of the target item, we can include longer phrases in the selection (such as "honma-ni", "a-honma!", or "a-honma-ni?"). The text display (selected from the options menu) adds labels to points as a further filter (e.g., the addressee in Fig.1).

Clicking on a point in the display plays the audio sample that is selected. Moving a point on the screen changes its prosodic coordinates in the database. Audio output is achieved using Snack library routines [5] that are available for all main computer architectures. The interface is programmed in tcl/tk. Source code is available for research use at [6] but will need rewriting for use on non-ESP corpora and data sets. The ESP corpora are not yet publically available.

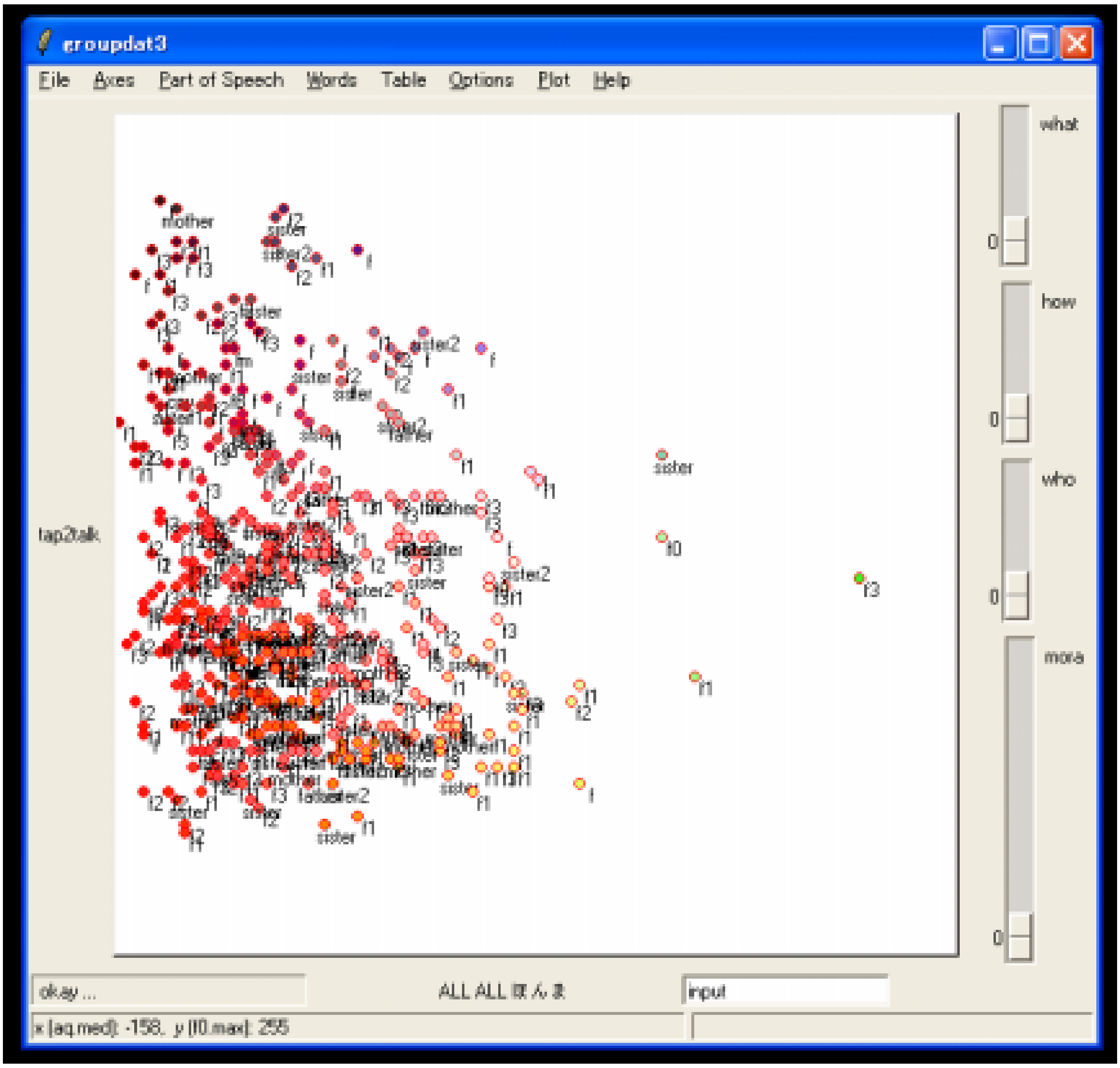
3. Selection filters

Three types of data are loaded into the interface on startup. An index of the speech waveforms, a transcription of each utterance, and a set of labels. These labels comprise acoustic information extracted for each utterance [4] as well as perceptual labels annotated by human listeners [7]. Being fixed-format, there is no hard-coding of the types, and the menus are generated automatically for the data as it is loaded.

4. tap2talk

Further to the work reported in [7], we are now testing this interface with a speaking-style-based selection procedure (see right-hand edge of Figure 1) for access by combinations of What, How, and Who. "What" selects the utterance type, or speech-act, "How" the speaking-style or emotion, and "Who" the type of addressee. The work is experimental, and extends the keyboard-free "tapagochi" approach for fast message generation that was presented in [8].

We estimate from an analysis of the corpora collected to date that this three-point selection of an utterance and a speaking style will suffice for the synthesis (or replay) of 80% of conversational speech, and are working to integrate a concatenative synthesiser to provide the remaining 20% of utterance content as a hybrid system for audio messaging.



okay ...

input

x [aq.med]: -158, y [f0.max]: 255

Figure 1. The tap2talk interface. *Points represent utterances in the speech corpus. Labels indicate (in this example) the interlocutor. X and Y axes represent variation in two prosodic dimensions (voice-quality and pitch range here). Menus (top, right) and free-input (bottom right) can be used for selection. Clicking a point plays the speech waveform.*

5.Conclusion

This paper has presented a design of an interface for access to very large speech corpora for the study of spontaneous speech (with software examples) and has suggested that the same framework can be used for the interactive synthesis of conversational speech (e.g., by the speaking-impaired, or for business and entertainment applications), if used in conjunction with a concatenative (CHATR-type) engine for the occassional insertion of proper nouns into otherwise pattern-conforming phrases. The system depends on the availability of many transcribed and annotated speech samples, and the automation of such transcription/annotation is necessary before wider use can be considered. Synthesis samples from the hybrid system will be available at [6] shortly.

Acknowledgements

This work is supported partly by a grant from the Japan Science & Technology Corporation and partly by aid from the Telecommunications Advancement Organisation of Japan.

References

[1] The JST/CREST Expressive Speech Processing project, introductory web pages at: www.isd.atr.co.jp/esp

[2] Campbell, W. N., “The Recording of Emotional speech; JST/CREST database research”, in Proc LREC 2002.

[3] Campbell, W. N., “MD vs DAT”, in Proc ASJ, 3/2002.

[4] Mokhtari, P, & Campbell, W. N., "Automatic detection of acoustic centres of reliability for tagging paralinguistic information in expressive speech", in Proc LREC 2002.

[5] The Snack Sound Toolkit – a tcl/tl audio library programming package: this free software can be downloaded from www.speech.kth.se/snack (and is highly recommended!)

[6] http://feast.his.atr.co.jp (under the ESP project web pages)

[7] Campbell, W. N., “Labelling natural conversational speech data”, 1-10-22, 273-4 in Proc ASJ Fall meeting, 2002.

[8] Campbell, W. N., “What types of input will we need for expressive speech synthesis?”, in Proc IEEE Speech Synthesis Workshop, Santa Barbara, 2002.

[1]