Building a Corpus of Natural Speech
- and Tools for the Processing of Expressive Speech -
the JST CREST ESP Project

Nick Campbell

ATR Information Sciences Division, Kyoto, 619-0288, Japan
nick@isd.atr.co.jp

Abstract

This paper presents details and progress of the JST/CREST ESP Project, part of the "Information Technology for Life in an Advanced Media Society" series of research projects. The purpose of the research is to determine the acoustic variables that indicate different attitudes and emotions in speech and to map these to linguistic objects and frameworks so that a model of the para-linguistic structure and content of an utterance can be obtained. The project has now been running for 6 months, and is expected to last for 5 years.

1. Introduction

In December 1999, the JST issued a call for proposals under the CREST Information Processing Technology for an Advanced Media Society, and a submission proposing the study of "Expressive Speech Processing" was accepted the following April. The CREST ESP Project started in July 2000 as joint research between NAIST, Kobe University, and ATR-ISD, with contributions from ICP Grenoble, Keio and Chiba universities, and from Omron's VIT project.

The goal of the ESP Project is to produce a corpus of natural daily speech in order to design applications that are sensitive to the various ways that people can use speaking style and changes in voice quality to signal the intentions underlying each utterance, and to add information to spoken utterances over-and-above that carried by the text of the utterance alone. The corpus will include samples of emotional speech, but will be focused on the attitudinal differences such as politeness, hesitation, friendliness, anger, and social-distance.

The most obvious applications of the resulting technology will be in speech synthesis, but the research will also be covering aspects related to speech-recognition technology with the particular goal of automating the labeling and annotation of the speech databases. The potential of realising the technology in applications for use in real-world speech-interactive situations will also be explored.

2. Progress to date

A kick-off meeting for team leaders was held in Kyoto at the offices of the JST CREST project in September 2000, and the first information-sharing workshop was held in February 2001 for all participants from each team. In the first six months of the project, efforts have been focused on setting up the infrastructure for research, and on establishing small test databases of various speaking-styles and recording-conditions in order to determine the optimal conditions for data collection and analysis.

2.1. Speech Databases

The speech databases that have been collected to date include readings of phonemically-balanced sentences for use in concatenative speech synthesis, samples of emotional speech from television broadcasts, DVD, and video recordings[1], as well as recordings of family conversations and talk between friends. Samples of disfluent speech from autistic and physically handicappped subjects are also being collected and annotated for speaking style characteristics.

The test data forms the basis for a feasibility study into the range of voice qualities and speaking styles that we can expect to cover in the main data collection.

2.2. Main Data Collection

The main data collection will take place starting from the summer of 2001. If we are to include spontaneous emotional speech then the use of broadcast materials is inevitable, and we will negotiate with local radio and television broadcast stations for limited-rights use of their recordings, as well as collecting our own speech data. A large part of the speech data will be recorded in audio-visual conditions.

Preliminary recordings of daily conversational

[1] We are aware that the issue of speaker's rights needs to be addressed as a matter of high priority, and are reluctant to incorporate materials that are covered by copyright, but at the present stage in the research, we have no plans for the public distribution of the initial data that we are examining, and are widening our scope of initial coverage as broadly as possible.

speech using mobile light-weight DAT and Mini-Disk recorders, with pin-mounted microphones of studio quality, resulted in high quality recordings of the main speaker, but suffer from excessive background noise. Clear naturally-spontaneous recordings have been made of one side of a telephone conversation, using a portable phone in conjunction with a broadcast-quality microphone placed close to the speaker's mouth, with the speaker in a sound-treated room.

Because the aim is to produce speech technology able to cover the full range of human speaking styles, it is important to collect that speech representative of normal daily interactions. It is preferable not to use actors but to collect speech in-situ by encouraging volunteers and paid subjects to wear recording devices for long periods of time. This so-called "Pirelli-Calendar" approach is inspired by the fact that photographers once took 1000 rolls of 36-exposure film on location to produce a calendar containing only 12 photographs. Perhaps only by over-collecting data can we guarantee adequate coverage with sufficient naturalness. The resulting corpus will be of interest to many sectors of the speech- and language-processing communities.

3. Tools for Speech Analysis

Since the most widely used software for speech analysis is no longer publically available (Entropic's ESPS and Xwaves having been bought up by Microsoft), we have explored several alternatives and are currently using the Tcl/Tk extension "Snack speech-processing" libraries in conjunction with the "Wavesurfer" software developed and released in the public-domain by the KTH laboratory in Stockholm.

Our prime current needs are for the automatic segmentation and phonemic labeling of the speech signal. For speech segmentation we are testing the "Julius" software released in the public domain by the IPS Project, and the "HTK-3.0" Hidden-Markov modeling software toolkit released by Cambridge University. Software for the analysis of voice quality will be implemented starting from March 2001, but preliminary studies have already started [Marumoto-2000].

For the analysis and annotation of speaking style, we are testing our own implementation of the "Feel-Trace" two-dimensional labeling method proposed by Belfast University under the PHYSTA (Emotion in Speech) Project. This software facilitates labeling of speaking-style along the positive-negative and active-passive dimensions and produces a coloured trace that can be viewed time-aligned to the speech waveform and its transcription.

For pitch-contour labeling, we are testing the MOMEL/IntSint software package from the University of Aix-en-Provence. This eliminates the need for manual prosodic labeling by fitting a quadratic spline to the estimated fundamental-frequency contour and abstracting from the derived target points to produce a series of abstract specifiers of the underlying contour shape. It is not yet known whether the symbols thus derived can be satisfactorily predicted from text, but this investigation is now under way.

4. Synthesis Development

In order to use the ATR-ITL CHATR unit-concatenation speech synthesis software[2], we have developed tools for the design and collection of prosodically-balanced speech corpora for NATR-specific source-unit databases. The automatic creation of a speaker- or speaking-style database still requires a transcription the speech data. However, the transcription of unprompted speech is a step which has not yet been automated, and still requires extensive manual labour.

An improved method of unit-selection, making direct use of the segment feature-labels is being tested. This makes redundant the prediction of prosodic contours for specifying numerical targets and thereby eliminates one source of known error in the unit selection process.

Research on the interaction of voice-quality and speech prosody has already started, and techniques for the automatic tagging of voice quality are being developed. Particular interest is being paid to pressed-voice and breathy-voice, both of which signal information-bearing and meaning-related speaker-attitude differences independently of prosody.

In order to provide a fast speech generation method for people who are unable to type, we have designed a touch-interface and a sockets-based input protocol which uses a NATR synthesiser running as a server on either the local or a remote machine. By clicking icons specifying particular utterances, or components of more complex utterances, the user can quickly produce speech synthesis in the language, voice, and speaking style of choice, using a mouse or pointing device instead of a keyboard.

5. Data & Information Flow

Because of the distributed nature of the ESP project, we have standardised data formats and annotation methods to allow easy interchange of data, information, and tools. Two file servers have been installed for data storage, and transfer of files between sites is primarily by means of web-based upload and download. Home pages for each team are regularly updated to allow quick transfer of ideas between researchers, and the project mailing lists are archived on the same servers. Password-protected pages allow transfer of semi-confidential internal-only information, but the creation of public-information

[2] Now being developed under the name of "NATR", an acronym for Next-generation Advanced Text Rendering.

pages is encouraged. The sites can be accessed from *www.isd.atr.co.jp/esp,* where project descriptions and publically-accessible samples of the speech data can be found.

Recent developments in notebook computers, especially in audio-visual input/output, have made it possible to issue high-powered portable machines to all researchers for general-purpose use while maintaining professional desktop-computing standards. Software packages such as Cygnus and Meadow (Unix and Emacs for Windows respectively) have made it possible to maintain programming environments of research quality while also taking advantage of the peripherals and devices of the newer machines. Our goal is for compatibility between both operating systems and for researchers to switch easily between computing environments to make use of the tools and advantages of both. All the software we develop in the project is expected to run equally well on both UNIX (Linux) and Windows platforms.

6. Current Projects

The following research themes are currently being actively investigated:
Collection of daily-conversation databases – prosodic labeling of Kansai intonation using ToBI for speech synthesis – segmental labeling of Japanese speech databases (using Julius) - feature-based unit-selection – speaker voice modification – segmental labeling of English speech databases (using HTK) – prosody-balance and database recording - analysis of television speech – software for communication aids – autism & speech impairment – development of speech labeling software – NATR speech synthesis databases - NATR visual user-interface – speaking-style labelling – prosodic/syntactic features of fillers – Modeling of prosody and discourse factors

7. Theoretical Issues

Apart from the practical issues of designing speech corpora and software tools for Expressive Speech Processing, we also face several theoretical issues which need to be resolved. The study of para-linguistic information, and of the mapping between linguistic structure and speaker intention as revealed by the spoken message, is still in its infancy. The speech code is a very hard code to crack.

The goal of our research is to develop theories of 'para-language in use.' Previous studies on the communication of intentions, attitudes, and emotions by means of para-language were mainly based on the observation (i) of laboratory speech, which is completely different from spontaneous speech, and (ii) of utterances isolated from the discourse contexts. In contrast, our research focuses on the analysis and modeling (a) of the para-linguistic features of utterances in naturally occurring discourse (b) with consideration of their discourse contexts.

7.1. Classification of Speaking Styles

Although the word "emotion" appears frequently in the context of Expressive Speech, we distinguish three categories; primary emotions, secondary emotions, and attitudes. For speech processing by machine, we are principally concerned with the third category.

- the *primary emotions* enable a human to react to the environment (e.g. fear changes the physiological state to help someone react faster). They can be recognised by an interlocutor and accepted as part of speech communication, but the aim of the "producer" of emotional expression is not for this expression to be perceived.

- the *secondary emotions* are learned with social stimuli, and are expressed in order to be perceived by the interlocutor (amusement for example). These emotions can be conditioned, but their expression, even if they are variant across languages, does not characterise a language community.

- the *attitudes* are also learned. They express the intention of the speaker and are produced by the speaker in order to influence the interlocutor's reaction in the communication act (e.g. doubt, surprise...). The attitudes are completely characteristic of a language community and these conventions must be modeled for translation processing or language generation.

The following are some of the research sub-themes have been proposed for the coming year:

7.1.1. Analysis of acoustic features of fillers

Spontaneous discourse contains lots of fillers like 'anoo' and 'eeto.' These fillers do not merely show delay in speech production process, but are indicative of the speakers' emotions and their attitudes toward listeners. We analyze the prosodic/syntactic features of fillers to develop a model for communication, of intentions, attitudes, and emotions in spontaneous discourse.

7.1.2. Modeling of prosody and discourse

When we move from laboratory speech to spontaneous discourse, we see various external factors affecting speech production, such as discourse plan, topic structure, and information structure. These factors are correlated with prosody, but in previous studies the interaction of prosody and the combined factors is not clear. We will apply a multivariate statistical modeling for this analysis.

7.1.3. Developing communication systems.

In order to simulate the diversity of speaking styles in human-machine communication systems, it is necessary to analyse and simulate the speaker attitudes, so we will carry out perception experiments in Japanese, English, and Chinese in order to determine the significant parts of the prosodic contours which carry the main information. This study will be extended to database analysis, and methods developed for automation. These methods will then be adapted and applied to spontaneously emotional speech in the three languages, The results of the analysis, will form the basis for parameter generation in synthesis for expressive speech.

7.1.4. Social aspects of speech information

The effect of Expressive Speech Processing in practical information-based systems will be evaluated, and feedback from the field trials will form the basis for future research directions. This part of the research will be carried out in cooperation with industry.

8. Future Schedule

Longer-term goals can be categorised under 4 headings; (a) natural-speech database design and collection, (b) statistical modeling and parameterisation, (c) mapping between "speaking style" and "intended meaning", and (d) implementation of prototypes and testing in real-world applications.

(a) Natural-Speech Database Design & Collection
H.13 creating and testing testing initial databases
H.14 collection of (1000-hr) main databases
H.15 addition of subsidiary databases
H.16 packaging and documentation

(b) Statistical Modeling and Parameterisation
H.13 acoustic feature extraction
H.14 stochastic training for prediction/detection
H.15 implementation of synthesis algorithms
H.16 testing and final evaluation

(c) Mapping from "Speaking Style" to "Intention"
H.13 developing "para-linguistic" models
H.14 acoustic to linguistic/para-linguistic mappings
H.15 statistical training for prediction and detection
H.16 interfaces for manual "intention specification"

(d) Protoyping and Testing Real-World Applications
H.13 development of software for database labeling
H.14 development of algorithms for speech synthesis
H.15 testing in closed-world laboratory environment
H.16 testing in real-world applications

9. Conclusion

This paper has described progress during the first year of the JST/CREST ESP Project, concerning the creation of natural-speech databases for the analysis and synthesis of "Expressive Speech", and the development of software tools for parameter-extraction and speech database labeling. The languagse of concern are Japanese, English, and Chinese.

The research is still in its initial stages, but we now have a clear understanding of the types of speech data that will be necessary, and of the software and speech processing tools that are available for analysis and treatment of the data. The testing of applications and prototyping in real-world situations is reserved for future work, and our prime goal now is the collection of a large corpus of natural conversational speech upon which future developments will be based.

Some publications arising from the CREST ESP Project:

[1] Iida, A., Iga, S., Higuchi, F., Campbell, N., Yasumura, M.: "Designing and Developing a Conversation Assistive System with Speech Synthesis and Emotional Speech Corpora," ISCA (International Speech Communication and Association) ITRW on Speech and Emotion, pp.167-172.

[2] Iida, A., Campbell, N., Yasumura, M.: "Corpus-based speech synthesis for ALS patients to assist their communication," Technical report of IEICE WIT00-33. pp37-42.

[3] Nick Campbell "Databases of Emotional Speech" pp 34-38, Proc ISCA w/s on Speech & Emotion, Belfast 2000 Speech Communication Special Issue on Emotion and Speech (forthcoming)

[4] Sylvie Mozziconacci "The expression of emotion considered in the Framework of an intonation model", pp 45-52, Proc ISCA w/s on Speech & Emotion, Belfast 2000

[5] Sylvie Mozziconacci "Expression of emotion and attitude through temporal speech variations", pp XX in Proc Intl Conf on Spoken Language Processing, Beijing

[6] Iida, A., Campbell, N., Iga, S., Higuchi, F., & Yasumura, M., "A speech synthesis system with emotion for assisting communication", Speech Communication Special Issue on Emotion and Speech (forthcoming)

[7] Nick Campbell & Toru Marumoto, "Automatic labeling of voice quality in speech databases for synthesis", pp XX in Proc Intl Conf on Spoken Language Processing, Beijing

[8] Li-chiung Yang. 2000. "The expression and recognition of emotions through prosody", in Proc Intl Conf on Spoken Language Processing, Beijing

[9] Li-chiung Yang. 2000. "Prosody and Topic Structuring in spoken dialog", in Proc Intl Conf on Spoken Language Processing, Beijing

[12] Matsumoto, Emiko, and Toshiyuki Sadanobu 2001 "Rikimi" in Japanese prosody and the degree of foreign students' understanding "rikimi", Proceedings of the Fifth International Symposium on Japanese Studies and Japanese Language Education, The Chinese University of Hong Kong.