Speech & Expression; the Value of a Longitudinal Corpus

Nick Campbell

ATR Human Information Science Labs
Keihanna Science City, Kyoto, Japan
nick@atr.jp

Abstract

This paper describes a new corpus that has been created for speech technology research. It includes the voices of several speakers recorded over an extended period of time, and illustrates how speaking style and voice quality can change according to differences in interlocutor and utterance content. Whereas most corpora are purpose-designed and collected under controlled circumstances, often resulting in a constrained style of speaking, the ESP corpus (from the JST/CREST Expressive Speech Processing project) was collected to illustrate the wide range of speaking-styles that can occur in ordinary everyday conversational situations. It is limited in the number of speakers and contexts of the dialogues, but provides a unique sample of everyday conversation for the speech researcher.

Introduction

There are now many large corpora of speech available for research uses. Most have been produced for speech recognition training, and tend to contain many voices but only short segments of speech from each. More recently large corpora have also become available for speech synthesis (e.g., the Siemens Synthesis Corpus for German, or CMU_ARCTIC), but with a few notable exceptions (e.g., Call Home, and the Spoken Dutch Corpus), these corpora illustrate prepared or prompted speech rather than the more spontaneous conversational variants used in normal everyday situations. This paper describes a new corpus that provides what may be the first longitudinal data of natural conversational speech.

Recorded over a period of more than three years, using high-quality head-mounted microphones and small portable walkman recording devices, the corpus illustrates the daily spoken interactions of a small number of volunteers as they go about their normal lives. The corpus illustrates not just the different levels of politeness and familiarity of everyday speech, but also the wide variety of paralinguistic information that such interactive speech contains. The data present challenges to current speech processing technology because neither speech synthesis nor speech recognition are yet capable of processing these types of (often non-verbal) interaction-related information.

Manual labelling of the discourse and pragmatic features in the corpus has shown that that the same linguistic utterance can have consistently different paralinguistic interpretations depending on speaking style and phonation settings. In particular, and in contrast to the more traditional purpose-built or domain-specific speech corpora, we find many more examples of non-verbal speech, signalling affect rather than linguistic information. These (often non-verbal) speech sounds are impossible to transcribe unambiguously as text, but have unambiguous meanings to the listener. Many of them are interjections, fillers, and backchannel utterances, which signal to the listener (a) that the discourse is active and interesting (or not), and (b) the state-of-mind, attitude, and emotions of the speaker, and (c) the nature of the relationship between the speaker and the listener at that time. Very few have dictionary entries, and they are often considered as noise.

The JST ESP corpus

The Japan Science & Technology Agency's Core Research in Evolutional Science and Technology (CREST) initiative has recently funded research into Expressive Speech Processing (ESP) as part of the 'Information Processing for an Advanced Media Society' framework. The research is principally based at ATR, in Kyoto, but includes contributions from Kobe University, NAIST, (the Nara Institute of Science and Technology) Chiba University, Keio University, and the ICP at Stendhal University in Grenoble, France.

The goal of the JST/CREST ESP project is speech technology, developing advanced interfaces for spoken-language interactions, but this work is grounded in the production and analysis of a very large corpus of everyday spoken interactions. The project specifically aims to improve speech synthesis technology, extending work performed at ATR on the CHATR speech-waveform concatenation synthesis methodology, with focus on the need to extend synthesis from the relatively simple domain of information provision, or announcements, into the more complex domain of everyday speech. If synthesis is to be used in place of a natural voice in conversational contexts, such as are required for speech translation, customer care, prosthetic devices, and possibly robotics, then a new level of content description will be required.

For a synthesiser capable of reproducing the interactive speech of normal people (i.e., not that of professional announcers or trained performers) in ordinary daily conversational situations, it is necessary first to have a corpus illustrating the range of variation in speaking-styles and voice qualities that are typically encountered in the expression of emotion, affect, and attitude, in addition to those required for the more simple provision of text-based information or propositional content.

Since no such corpus was available, we recruited a small number of volunteer speakers who agreed to wear lightweight recording equipment throughout their daily social interactions. They were paid by the amount of speech they provided, but no other pressures were imposed concerning what and when to record.

The recording media and devices

We experimented with various types of microphone and recording devices, but found that the combination of a Sennheiser HMD410 head-mounted dynamic microphone and a SONY Minidisk recorder was the most effective. Other microphones were tested, including the SONY ECM-77B lavalier microphone which is widely used in television studios around the world, but requires an independent power supply and is, for our purposes, too sensitive to external noise. For television sound recording, this microphone is usually clipped to a lapel, but we found that the large amount of head-movement in conversational situations required the use of an ear-mounted boom to ensure constant mouth-to-mic distances. The Sennheiser adjustable ear-mounted boom proved to be the lightest and most flexible.

It is well known that DAT (digital audio tape) provides a higher quality of recording than Minidisk, which employs ATRAC compression to remove perceptually insignificant parts of the signal, but because of its lightness and portability we used an MD recorder for at least half of our corpus. Tests confirmed that although the compressed signal yields very different cepstral coefficients (making it unsuitable for speech recognition without first retraining the acoustic models), the estimated prosodic variables (pitch, power, duration and phonation-type, or voice-quality) are functionally equivalent to those obtained from DAT recordings (Campbell & Mokhtari 2002).

More recently, we have also been testing the Marantz PMD670 portable solid-state recorder, which has no moving parts and accepts Compact Flash cards, holding from 1GB to 4GB of memory, or 'Microdrive' hard-disks that store up to 72 hours of speech data per gigabyte. At 16kHz, 16-bit sampling rate, a gigabyte of memory holds approximately 9 hours of monaural speech data, which is enough for a day of continuous recording with no need to change tape or disk. The recorder comes with a USB connector that allows simple transfer of the data to main storage. However, typical battery life is only 6 hours, and the device is not pocket-sized (measuring 264 x 55 x 185 mm), so currently it is most suitable for use in fixed locations. In conjunction with an Audio-Technica ATW-R92 receiver and a pair of microphone transmitters (e.g., ATW-T93b), it enables high-quality continuous recording even while speakers are moving around a house or room.

Collection techniques

Three main speech-collection paradigms were employed, after various tests and trials; one being speech-synthesis specific, and the other two more general; one long-term, and the other short-term. The long-term collection was briefly described above; with (mostly) young female adults (few male speakers volunteered) wearing portable recorders throughout the day. The short-term collection was more carefully balanced; with both men and women paid to attend a centre once a week to speak over the telephone to remote partners who were selected according to native-language, age, sex, and familiarity. All telephone conversations were recorded to DAT using head-mounted microphones, and each lasted for half-an-hour. No guidance or constraints were given concerning the content of the conversations (Campbell 2002).

Speakers

The youngest speaker is still only three-months old, the oldest is in her fifties. Student volunteers and their young adult friends were recorded for the synthesis-specific collections, where the objective was to obtain a balance between read speech (which is highly constrained, but relatively easy to label automatically by forced-alignment with the source text) and free conversational speech (which has so far proved extremely difficult to label by automatic methods). By training speaker-specific acoustic models on the read-speech data, we aim to improve recognition on the conversational speech of the same speaker, in spite of its much wider range of variation. This work is still in progress.

The speakers for the short-term collection were selected for us by a recruitment agency, according to criteria of age and sex and first language. Initially, they were strangers, but they came to know each other well as the recordings progressed, over a period of three months. We targeted three types of speech: between strangers, between family members, and between different cultural groups, each with varying degrees of familiarity between speakers, and with same-sex and different-sex pairings. Each member spoke to at least three other members, differing in age, sex, and background. All conversations were in Japanese. We had anticipated that the cross-cultural conversations would be the most difficult, with Japanese talking in their own language to male and female native-speakers of Chinese and English, but it transpired that the 'family' conversations caused most problems – it seems that typical family conversations are usually very short, and for a husband to have to talk with his wife for as long as 30-minutes was unexpectedly difficult (!).

While the above two sub-corpora are already providing interesting material for research into conversational speech and speaking styles, it is the long-term collections that have turned out to be most interesting Ten young female adults and two male adult speakers of Japanese wore recording equipment to capture spoken interactions while going about their daily routines. More than half of the conversations are held over a telephone line (a fact, not a requirement), with only the voice of the volunteer speaker being recorded. All utterances are transcribed, and a large part has been labelled for discourse features.

File formats and sizes

Speech recorded on DAT was sampled at 48kHz, and that on Minidisk at 44.1kHz, or 32kHz using long-play mode. All files were also downsampled to 16kHz at 16-bit precision for consistency of processing and to conserve storage space. We are still experimenting with optimal formats for the speech recorded directly to disk, and are testing the effects of various compression techniques (e.g., raw vs mp3 vs mpeg2). File sizes vary according to the length of the conversation, with the shortest being five seconds, and the longest being forty-five minutes. Many files contain long periods of silence, since listening is an important part of conversation, so each was also cut, using utterance start- and end-time information from the transcriptions, into utterance-sized sub-files with only the speech parts preserved. For general analysis, this resulted in great savings in search time as well as storage space.

All recordings are initially monaural, but the short-term telephone conversations were merged to produce stereo files for ease of listening when transcribing and labelling.

Transcription conventions

Transcriptions facilitate access to the data by providing time information aligned with the text. The Multitrans (Bird et al, 2002) public-domain software is used for the initial transcription, but the data are then transferred to plain-text files (with one utterance per line) and to Excel spreadsheets where the filename, start-time and end-time information is passed to a replay macro so that labellers can listen to the individual utterances when annotating them for speaking-style and speech-act details (see below). In spite of principled objections from the UNIX-preferring research staff, our labelers, who are computer non-professionals, prefer working with the spreadsheet format for annotations. All analyses are performed in UNIX.

We employ a three-pass method for the transcriptions. In the first pass, the transcriber identifies utterance chunks (by pressing on the return-key to enter a marker in the speech waveform and to open a transcription line in the editor) while listening through the whole conversation. In the second pass, text is entered for each chunk and the edges are aligned more accurately. In the third pass, transcriptions are checked, and utterances split more finely if necessary or possible. The definition of an 'utterance' is very difficult and we use a "yen per line" criterion in cases of doubt; which translates to something like "segment as narrowly as possible but don't cut 'in the middle of something'", i.e., roughly corresponding to an intonation-phrase. Many are a syllable or two in length, but the longest can be more than 50 syllables.

Annotation levels

Since the transcriptions must be readable for people as well as for computers, we use Kanji-kana alphabets (the Chinese characters used in normal Japanese writing), with a bracketing system to indicate the actual pronunciations when more than one is possible. Round brackets are used to identify the scope, and square brackets to clarify the content. A similar method is used to indicate noisy sections of speech ([X]), laughter ([W]), lip-noise ([S]), and breaths ([H]). All the recordings are first transcribed manually, then each utterance is labelled by a team of three or four labellers under the following categories: emotion, interlocutor, dialogue-act (see table 1 for details), manner (polite, casual, formal, informal) mood, and voice (breathy, hard, soft, dark, bright, etc).

Analyses performed

We tested the Feeltrace system (Cowie et al, 2000) for labelling emotion, but found that very little of the speech displays any strong emotional content and that most of the speech is just 'mildly positive'. Rather than 'emotion', the corpus is remarkable for variation in the more social dimensions, and the reader is referred to Campbell & Mokhtari (2003) for details of how voice and speaking style change according to the above categories.

Cross-cultural analyses of the interpretation of the non-verbal utterances that form a large part of the corpus, compared the interpretations of American, French, and Korean listeners to the Japanese utterances (Campbell & Erickson 2004), and we are now considering the degree to which they indicate language (or paralanguage) universals that function independently of the linguistic framework.

Samples of the data and the analysis results are accessible under http://feast.his.atr.co.jp, the project web site.

Conclusion

This paper has described the current state of the JST/ATR Expressive Speech corpus, listing speakers, and detailing file formats, sizes, and annotation levels. It also describes the collection techniques, and presents a brief review of some analyses performed on the data. These analyses confirm the unique nature of the corpus (and justify the laborious task of collecting it) by showing that the speakers consistently adapt their speaking styles and even voice quality (laryngeal phonation settings) according to non-linguistic factors such as their relationship with the listener and to the nature of the discourse content. Current speech technology, which is based largely on the analysis of prepared speech, is not yet capable of processing the types of information that we find to vary consistently in this corpus. We encourage the wider collection of such data. The ESP corpus will be made available for research use next year, but because it contains much personal information, confidentiality and non-distribution contracts must be signed before access can be granted.

Acknowledgements

This work is supported partly by a grant from the Japan Science & Technology Agency under CREST Project #131, and partly by aid from the Telecommunications Advancement Organisation of Japan. The author is grateful to Minako Kimura for assistance, and to the management of ATR for their support and encouragement.

References

Bird, S., Maeda. K., Ma, X., Lee, H. J., Randall, B., and Zayat, S., (2002) "TableTrans, MultiTrans, InterTrans and TreeTrans: Diverse Tools Built on the Annotation Graph Toolkit", Proceedings of the Third International Conference on Language Resources and Evaluation.

Campbell, N., (2002) "Recording techniques for capturing natural everyday speech", in Proc Language Resources and Evaluation Conference, Las Palmas, Spain.

Campbell, N. and Mokhtari, P., (2002) "DAT *vs.* Minidisc: Is MD recording quality good enough for prosodic analysis?", 1-P-27, in Proc Acoustical Society of Japan Spring Mtg.

Campbell, N., and Mokhtari, P., (2003) "Voice Quality; the 4th prosodic parameter", in Proc 15th ICPHS, Barcelona, Spain.

Campbell, N., and Erickson, D., (2004) "What do people hear? A study of the perception of non-verbal affective information in conversational speech" in Journal of the Phonetic Society of Japan, V.7,N.4.

Cowie, R., Douglas-Cowie, E., Savvidou, S., McMahon, E., Sawey, M., and Schröder, M., (2000) "FEELTRACE: an instrument for recording perceived emotion in real time", in Proc ISCA w/s on Speech & Emotion, Belfast, N. Ireland.

Table 1. ESP corpus, Dialogue Act Labels

会話のやりとりの形態 Dialogue form	Category	DA	発話末形態 Speech-final pattern
	Q 質問 Question	q 質問 Question	
	確認 Y/N Question	q2 確認 Y/N Question	
	聞き返し	q3 聞き返し Repetition Request	＊q 確認型 Tag question
	C 依頼 Request	c 依頼 Request	
			＊m もちかけ型 Incomplete
	O 意見 Opinion	o 意見 Opinion	
		o2 褒める、感心 Compliment	
質問 Question		o3 希望 Desire	＊d 言い詰まり Disfluency
語り Telling		o4 意思 Will	
		o5 感謝、ねぎらい Thanks	
		o6 謝罪 Apology	＊v 強調 Emphasis
	N 否定的意見 Negative opinion	n 反論 Objection	
r 回答 Response		n2 文句 Complaint	＊q2v 肯定要求型 Request affirmative answer
r 反応 Reaction			
	C 指導的意見 Advisory opinion	c2 指導、忠告 Advice	
		c3 命令 Command	＊qb 倒置（Qのみ）
	P 提案的意見 Suggestion	p 相手に提案 Suggestion	
		p2 自分から申し出 Offer	
		p3 勧誘 Inducement	
	S 説明 Informative	s 説明 Give-information	
		s2 読み上げ Reading	
		s3 自己紹介 Introduce-self	
		s4 話題紹介 Introduce-topic	
		s5 会話終了 Closing	
	G あいさつ Greetings	g あいさつ Greetings	
		r 返事 Response (e.g., r1 = affirmative response, r2 = agreement, etc.,)	聞いている Listening 1 肯定 Affirmative 2 同意 Agree 3 理解 Understand 4 納得 Convinced 5 承諾 Accept 6 興味 Interested 7 理解してるが納得していない Understand but not convinced 8 はっきりした答えが無い Uncertain
		i 相づち Backchannel ii 相づちの相づち Answer to Backchannel (e.g., i2 = agree backchannel, i23 = doubt, etc.,)	9 否定 Negative 11 繰り返し Repeat 12 自己納得(ひと呼吸) Self convinced a 気づき Notice k 考え中 Thinking 21 意外 Unexpected 22 驚き Surprise 23 疑問 Doubt 24 感心 Impressed 25 共感 Sympathy 26 同情 Compassion 31 その他の感情 Other 32 感嘆 Exclamation
	人真似 Mimic h＊ くせ Habit ＊@	a 気づき Notice w 笑い Laugh f フィラー Filler d 言い詰まり Disfluency	x 独り言 Talking-to-self y 自問自答 Asking oneself z 自己確認 Check oneself