An Expressive Speech Corpus of Standard Chinese*

Xia Wang
Nokia Research Center, China
No. 11, Hepingli dongjie, Beijing 100013,
xia.s.wang@nokia.com

Aijun Li
Institute of Linguistics, Chinese Academy of Social Sciences
No. 5, Jianguomennei Dajie, Bejing, 100732
liaj@cass.org.cn

Jianhua Tao
National Laboratory of Pattern Recognition (NLPR), Institute of Automation, Chinese Academy of Sciences, China, Beijing 100080
jhtao@nlpr.ia.ac.cn

Abstract

The paper introduces an Expressive Speech Corpus of Standard Chinese (ESCSC) which is designed for spontaneous speech analysis in human computer. The corpus is characterized by spontaneity and various speaking styles during human communication. Three male and three female professional speakers were main contributors of the corpus, who were recruited from Communication University of China and Drama Institute of Beijing Film Academy. There was about twenty hours' speech from each speaker. The ESCSC corpus contains four sub-corpora, i.e. DIACRIPT of performed dialogues with scripts, DIAREAD of read speech based on the transcription of DIACRIPT, SPONCON of spontaneous dialogues without any script, and EXPDISC of expressive speech of various types of discourses. Part of EXPDISC was recorded with videos. In addition to six professional speakers, fourteen non-professional speakers also contributed to this corpus.

1 Purpose and Principles

Expressive speech is conveyed in speech communication with a variety of different types of information, including emotional elements, paralinguistic elements, various intonation styles and large voice quality variations. Most of us believe that research on expressive speech is able to promote the level of speech technology in real applications. In our work, we aim at collecting speech of different speaking styles for the research of spoken language generation to improve spontaneity and expressiveness of speech synthesis systems, which was supported by the high-tech (863) program "New methods of spoken language generation in natural human machine interaction". The purpose is not only to investigate the characteristics of neutral style speech vs. more natural and expressive styles, spoken language vs. written language, but also to explore paralinguistic phenomena in spoken dialogues, as well as relationship between speech acts and speech representations in spoken dialogues, and relationship between gesture and speech expressions.

To fulfil our research purpose, the following principles were identified before the corpus designing process:

- To collect speech samples with as many as possible speaking styles, e.g. neutral read speech, various expressive styles with scripts, spontaneous natural speech without any scripts;
- To collect speech samples with as many as possible speaking formats, e.g. monologues and dialogues of various scenarios;
- To collect rich content materials that covers daily communications and most common discourses, e.g. weather report, novels and short stories;
- To collect speech samples from as many as possible speakers, with good voice quality and expressiveness;
- To collect reasonable number (big enough) of speech samples from each speaker that fulfils our research needs.

2 Corpus Description

In the realm of human computer interaction, we have to take into account some parallel analysis, such as the spontaneous speech vs. the read speech, dialogues with scripts vs. dialogues without scripts, and expressive speech with various styles. With these requirements, our ESCSC corpus is divided into four sub-corpora.

2.1 DIACRIPT

DIACRIPT is a sub-corpus of performed dialogues based on 968 scripts from 40 topics, consisting of 8720*2 turns. Speakers were asked to do role plays according to but not limited to the scripts. Speakers had the freedom to use their own language in their comfortable way to play the roles. Therefore the sub-corpus was rich of spontaneity, emotions and expressiveness.

2.2 DIAREAD

DIAREAD is a sub-corpus of read speech based on the transcription of DIACRIPT, for the comparison with DIACRIPT. The DIACRIPT sub-corpus was first transcribed orthographically, segmented into turns, then randomized and presented to the speakers. The speakers were asked to read exactly the prompted text to create the DIAREAD sub-corpus.

#	topic	Σ	#	topic	Σ
1	asking for help	16	**21**	interview	80
2	hotel	123	**22**	automobile	9
3	restaurant	43	**23**	daily life	31
4	telephone	23	**24**	sports	23
5	bye	6	**25**	weather	24
6	ordering meals	20	**26**	greet	22
7	ticketing	36	**27**	ask the way	130
8	housing	4	**28**	study	18
9	at work	26	**29**	query	33
10	shopping	42	**30**	music	10
11	international relations	6	**31**	banking	10
12	transportation	31	**32**	Post-office	6
13	finance	11	**33**	gaming	8
14	hospital	14	**34**	entertainment	18
15	technology	19	**35**	baby care	22
16	manner	8	**36**	decoration	12
17	chatting	45	**37**	dress up	7
18	tourism	40	**38**	self introduction	7
19	real estate	33	**39**	car renting	15
20	barber-shop	14	**40**	house renting	21

Table 1: Summary of DIACRIPT: number of dialogues (Σ) for each topic

2.3 SPONCON

SPONCON is a sub-corpus of spontaneous dialogues of two speakers. There are 3 types of stimuli for the discussion: speaker's own choice of topics; picking up topics from a list of 160 topics; watching a video clip of about 5 minutes and then discuss about related topics and own views. The video clips were from TV interview programs like "a date with Lu Yu" by PhoenixTV (http://www.phoenixtv.com) on hot topics. There are 20 pieces of video clips used as stimuli.

All the other sub-corpora were collected from 3 pairs of professional speakers, but this one involved 14 additional non-professional speakers. Some pairs in this sub-corpus were video taped, in order to do a comparative study between gestures and speech representations, as well as relationship between speech acts and speeches. We also designed controlled experiments for the 10-pair speakers in such a way that 10 video clips were watched by one speaker and retold to another in his/her own language, and another 10 video clips were watched by both partners and related topics were discussed afterwards. This part is very natural, containing rich paralinguistic phenomena.

2.4 EXPDISC

EXPDISC is a sub-corpus of expressive speech, in which we selected 38 categories of discourses with a reference to category definition in discourse linguistics, with a total amount of 427 discourses in text. Our focus was daily life related domains like stories, novels, weather reports, etc. There are also various speaking styles like comic dialogue, horror novels, letters, sports commendatory, etc.

#	category	∑	#	category	∑
1	guarantee	8	**20**	modern poetry	26
2	report	27	**21**	Chinese philosophy	4
3	memo	11	**22**	sports commendatory	11
4	menu	15	**23**	weather report	4
5	Restaurant introduction	5	**24**	notice	18
6	manual	3	**25**	modern poem	10
7	prescription	5	**26**	comic dialogue	4
8	biography	5	**27**	short drama	5
9	investigation report	3	**28**	novel	28
10	short message	5	**29**	Univ. billboard	5
11	obituary notice	5	**30**	joke	10
12	classic drama	4	**31**	news casting	5
13	story	72	**32**	letter	31
14	advertisement	26	**33**	announcement	5
15	invitation letter	5	**34**	thesis	5
16	workshop	14	**35**	speech	5
17	travel	5	**36**	anecdote	7
18	city guide	8	**37**	Opera	3
19	ancient poem	10	**38**	wassail song	5

Table 2: Summary of EXPDISC: number of discourses (Σ) for each category

3 Speakers

The key contributors for this corpus were 6 professional speakers, 3 male and 3 female, in which 4 of them were from Communication University of China, majored in news casting, and the rest 2 were from Drama Institute of Beijing Film Academy. Speech collected from each speaker was about 20 hours. They are standard Mandarin speaker with excellent voice quality and expressiveness. They contributed to all of the 4 sub-corpora.

In addition to them, we also recruited 14 non-professional speakers to contribute to SPONCON natural spoken dialogue sub-corpus recording.

4 Recording Environment

The recording was done in a sound-proof chamber as shown in Figure 1. The room on the left is an anechoic chamber for speakers, and the room on the right was a control room for the operator. Microphones used for speakers are in the same type, AKG C420 headset cardioid condenser microphone with frequency range from 20 Hz to 20 K Hz.

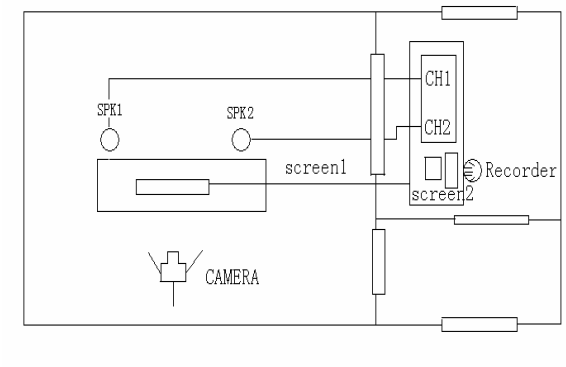

Figure 1: Recording chamber: anechoic room on the left, control room on the right, two synchronized screens

Figure 2: AKG C420 headset cardioid condenser microphone

CoolEditor was used for recording the dialogues and discourses. DIAREAD was recorded by our own recording software CASSRecorder. Video was shot by a SONY digital video recorder.

The corpus is partly done now. What have been finished so far are DIACRIPT about 35 hours, EXPDISC about 20 hours, DIAREAD about 10 hours, SPONCON about 15 hours. The database collection is still on going.

5 Annotation

We applied multiple layer annotation to the corpus, including phonetic transcription, linguistic and paralinguistic transcription, speech acts and discourse markers, etc.

(1) Phonetic transcription [1][2]

Phonetic transcription included orthographic transcription in Chinese characters, syllables, initials and finals, stress, and prosodic boundaries.

(2) Linguistic transcription

Linguistic transcription was to categorize every utterance into one of the 4 functional moods of Chinese, declarative mood, interrogative mood, exclamatory mood or imperative mood.

(3) Paralinguistic and nonlinguistic information

Paralinguistic and nonlinguistic information transcribed affections, emotions and various kinds of spoken language phenomena, according to the specifications defined in [2].

(4) Interactive function annotation

Interactive function annotations include speech act and discourse function annotations. We took a reference of SPAAC for speech act annotation [3]. Table 3 below shows some examples of the speech act categories.

In addition to the speech act annotation, we also annotated functions of discourse markers like “ng”, “a”, as shown in Table 4 [4].

label	superordinate class	broadly characterized as
accept	mainly responding	responding in an active positive way
acknowledge	mainly responding	signaling decoding, understanding
answer	mainly responding	answering a question
answer-self	external to dialogue goals	answer a question asked by oneself
answ（er）Elab（orate）	mainly responding	elaborating the answer to a question
appreciate	mainly responding	expressing appreciation
bye	interpersonal management	saying farewell; closing a dialogue
…	…	…

Table 3: Examples of speech act categories

#	category	label	broadly characterized as
01	feedback	Feed	signaling listening to O, not necessarily understanding what O has said
02	confirm	Conf	expressing understanding of what O has said
03	Think	Thin	non-verbal sounds produced when initiating a new phase of the dialog or in the middle of a dialog when thinking or not sure what comes next
04	initialize	Init	similar to 03 Thin, but focus on initiating a new turn of the dialog and signaling O for own start of a new dialog
05	check	Chec	check with O after own proposal, looking for confirmation or comments
06	repeat	Repe	signaling the need for O to repeat
07	question	Ques	request O to answer and explain regarding to own question
08	self answer	Seas	answer a question asked by oneself
09	deny	Deny	responding negatively, for disagreement; pitch contour often as "falling-rising"
10	self confirm	Seco	confirm one's own utterance
11	suddenly understand	Tumb	suddenly understand or know, similar to 02 Conf, but much stronger in the sense of "all of a sudden"
12	suddenly remember	Sudd	remember something in a sudden; differentiate from 11 Tumb by self-awareness
13	surprise	Surp	expressing surprise to what O said with strong emotion
14	strange	Stra	expressing hardness to understand, similar to 13 Surp and 12 Sudd
15	exclaim	Eecl	expressing emotion or attitude to an event, similar to 13 Surp
16	self repair	Repa	correcting one's own utterance
17	emphasize	Emph	emphasize
18	pending	Pend	an unclassifiable sound, e.g. a sound to fill a silence

Table 4: Function annotation category examples for "ng", "a" type of discourse markers
[O = the other person; S=self]

6 Conclusion

In the paper, we presented our Expressive Speech Corpus of Standard Chinese (ESCSC), which was designed for research on spoken language generation to improve speech synthesis systems. We tried to cover as many as possible different speaking styles and types of discourses, covering as many aspects as possible for comparative studies for spontaneous speech vs. read speech, neutral vs. more natural and expressive styles, spoken language vs. written language, as well as paralinguistic and nonlinguistic phenomena, speech acts, gestures etc., in the context of human computer interaction. The recording is still on going and the annotation will take significant efforts too to annotate segmental, prosodic, linguistic, paralinguistic and nonlinguistic information, as well as function communication information in spoken dialogues. We hope this multiple functional corpus will serve the research needs for spoken language analysis and generation.

* O-COCOSDA, Dec.3-6, 2007, Vietnam

References

[1] Li, Aijun. 2002. Chinese Prosody and Prosodic Labeling of Spontaneous Speech, in *Prosody Speech 2002*, AIX-EN-PROVENCE France.

[2] Li, Aijun, Zu,Yiqing. 2006.Corpus Design and Annotation for Speech Synthesis and Recognition, as chapter11 in *Advances in Chinese Spoken Language Processing*, edited by Chin-Hui Lee, Haizhou Li, Lin-shan Lee, Ren-Hua Wang, Qiang Huo, World Scientific Publishing Co. Pte. Ltd., Singapore, 2006.

[3] Geoffrey Leech and Martin Weisser. 2003. Generic Speech Act Annotation for Task- oriented Dialogues, in *Proceedings of the Corpus Linguistics 2003 Conference*, eds. D. Archer, P. Rayson, A. Wilson and A. McEnery. Lancaster: UCREL Technical Papers

[4] Yin, Zhigang. 2007. Study on "嗯/ng/, 啊/a/" *Type of Discourse Markers in Spontaneous Dialogues of Standard Chines*e, master thesis of Chinese Academy of Social Sciences.