CONSTRUCTING MULTI-LEVEL SPEECH DATABASE FOR SPONTANEOUS SPEECH PROCESSING

*Minsoo Hahn, **Sanghun Kim, **Jung-Chul Lee and ***Yong-Ju Lee

* Audio Information Processing Section
** Spoken Language Processing Section
Electronics and Telecommunication Research Institute
P.O. BOX 106 Yu-Seong Post Office, Taejeon, KOREA
Tel: 82-42-828-6234 Fax: 82-42-828-6231
e-mail: mshahn@audio.etri.re.kr
*** Dept. of Computer Eng.,WonKwang Univ., Chonbuk, Korea

ABSTRACT

This paper describes a new database, called muti-level speech database, for spontaneous speech processing. We designed the database to cover textual and acoustic variations from declarative speech to spontaneous speech. The database is composed of 5 categories which are, in the order of decreasing spontaneity, spontaneous speech, interview, simulated interview, declarative speech with context, and declarative speech without context. We collected total 112 sets from 23 subjects(male: 19, female: 4). Then the database was firstly transcribed using 15 transcription symbols according to our own transcription rules. Secondly, prosodic information will be added. The goal of this research is a comparative textual and prosodic analysis at each level, quantification of spontaneity of diversified speech database for dialogue speech synthesis and recognition. From the preliminary analysis of transcribed texts, the spontaneous speech has more corrections, repetitions, and pauses than the others as expected. In addition, average number of sentences per turn of spontaneous speech is greater than the others. From the above results, we can quantify the spontaneity of speech database.

1. INTRODUCTION

According to the advances of speech processing technologies, the spontaneous speech processing becomes one of the important topics. In particular, a speech synthesizer has to produce at least spontaneous speech-like output if it is adopted in spontaneous speech translation system.

The final goal of this work is to develop a dialogue type speech synthesizer with proper spontaneity by improving our present text-to-speech system[1]. However the current text-to-speech systems mainly handle the declarative speech only, thus the knowledges acquired from the declarative speech processing cannot be directly applied to the dialogue speech synthesis system. And that is why many researchers started their studies on dialogue speech processing[2][3][4]. This paper can be considered as our first step for the dialogue speech processing.

This paper is organized as follows. The motivation of collecting multi-level speech database will be presented in section 2. In section 3, we will explain criteria to classify the multi-level speech. In section 4 and section 5, Collecting strategy and specification of multi-level speech database will be described respectively. Finally, we will present the preliminary results on multi-level speech database.

2. MOTIVATIONS

It is likely that various speech types will be synthesized before long[5]. The dialogue speech synthesis is more valuable to develop than the other speech type[6]. However, since only a few researches on dialogue speech processing have been performed, there are many problems to solve for implementing the dialogue speech synthesis system based on the current techniques.

First of all, the lack of knowledges about the characteristics of spontaneous speech makes more difficult to develop the natural dialogue type speech synthesis system. As you know, dialogue speech has various linguistic phenomena, i.e. interjections, anaphora, repetitions, and repairs. And also it has more dynamically prosodic changes than declarative speech. In speech recognition,

these factors deteriorate the performance of recognition system. In speech synthesis, the algorithm for parsing the syntactic structure of ill-formed texts will be more complex and also the prosodic pattern modeling will be more difficult.

To obtain the knowledges of spontaneous speech, we built up a new database called multi-level speech database. The goal of this research is a relatively textual and prosodic analysis at each level, quantification of spontaneity of diversified speech database for dialogue speech synthesis and recognition. The proposed multi-level speech database will be a great help to our studies of dialogue speech processing such as designing of appropriate synthesis units covering acoustic variations in dialogue speech, dialogue type prosodic modeling, and dialogue type text processing.

3. CLASSIFICATION CRITERIA

The proposed multi-level speech database can be classified into 5 different categories. They are

- natural spontaneous(dialogue) speech
- interview
- simulative interview
- declarative speech with context
- declarative speech without context

To classify the various speech database into 5 levels, we chose speaker's voluntariness and spontaneity as criteria.

As mentioned, one factor is a voluntariness. It means a speaker's free intention or willingness when talking with each other. In case of declarative speech, the degree of voluntariness becomes almost zero because he has to read the predetermined texts. Accordingly, words used by speaker are under control by the manager. On the contrary, words used by speaker cannot be constrained in spontaneous speech. The above mentioned two levels represent extreme cases in the sense of voluntariness among our multi-level speech database. The other levels show the degrees of voluntariness between them.

The other criterion is a spontaneity. Spontaneity is related with unplanned speaking. When talking about something, one tends to speak with preselected words and utterances in general. But in case of spontaneous speech, it is likely that there frequently occur impromptu words and utterances. For that reason, the number of occurrences of ill-formed words, repetition, interjection and disfluency dramatically increase as the degree of spontaneity goes higher.

The classified speech databases using above two criteria are presented in Table 1.

Level	Voluntariness	Spontaneity
DIAG	very high	very high
INV	mid	high
SIMV	very low	mid
DSWC	very low	low
DSWOC	very low	very low

Table 1: Classified multi-level speech database where DIAG: Dialogue, INV: Interview, SINV: Simulative Interview, DSWC: Declarative Speech with Context, DSWOC: Declarative Speech without Context

4. DATABASE COLLECTION

We selected the travel domain as our data collection target domain, and recorded the database with a DAT recorder. To collect less constrained spontaneous speech data, we applied a top-down approach rather than bottom-up one. Namely, the speakers were totally unaware of the recording when the natural spontaneous speech data were collected. The other four categories of speech data were collected on the same theme based on the transcription of spontaneous speech data. And to obtain more useful speech data, we considered the environmental conditions and performed a rehearsal. Each database was collected by the following procedure.

- dialogue speech
  - Speaker should lead the conversation
  - Dialogue manager should only play the role of encouraging the speaker to talk freely on the same topic
- interview
  - Interview scenario is rewritten using the transcription of dialogue
  - Speaker should answer to the question of dialogue manager
  - Speaker should not be interrupted by dialogue manager
- simulative interview
  - Speaker should be fully aware of interview scenario.
  - Speaker should answer to the question of dialogue manager
  - Speaker should be constrained by dialogue manager

- declarative speech with context
  - Declarative scenario is rewritten using the transcription of dialogue.
  - Each utterance is transformed to well-formed sentence.
  - Speaker should be fully aware of the transformed declarative scenario
  - Speaker reads the declarative scenario
- declarative speech without context
  - The sentence order of declarative scenario is mixed
  - Speaker reads the mixed declarative scenario

5. SPECIFICATION OF DATABASE

We collected total 112 sets from 23 subjects(male: 19, female: 4). Total number of sentences(Q:Question, A:Answer) in each level are 793(Q), 2523(A) in dialogue speech, 323(Q), 554(A) in interview, 303(Q), 486(A) in simulative interview, and 1080 in declarative speech with context and without context, respectively.

The database was firstly transcribed using 15 transcription symbols according to our transcription rules, i.e., speaker's noise, environmental noise, repetition, interjection, fall-rise intonation, silence, foreign expression, etc. Table 2 shows a transcription example for the interview type speech. To minimize transcription errors, the transcribed texts were carefully corrected three times by an expert.

Q: 특별한 이유라도- 있습니까{↗}{Q}
A: /ls/할아버지께서- 일본에 살고 계십니다{P}
Q: /ls//ah/ 일본 어디-에 살고 계십니까{↗}{Q}
A: 오사까 근처라고만 알고 있습니다P
Q: /h#//mm-/ 그럼 한%1% 번도 뵌적이 없습니까{Q}
A: 일%1%년에*S* 두%2%어번 나오셨-습니다{P}
Q: /er//h#/그럼 일본에 사시게 된겁니까{↗}{Q}
A: 일제시대때- 징용당하셨습니다{P}
Q: 예 고생이 많으셨겠군요{↘}{P}
A: 예{P}

Table 2: Transcription example for interview

Secondly, for the prosodic labeling of the speech database, the recorded speech was digitized in 16KHz, 16bit resolution, and then segmented and labeled using an autolabeler for the durational modeling. For the intonation modeling, we extract pitch using the ESPS/Xwaves+ speech signal processing tool. The prosodic information was attached using the modified Korean ToBI system.

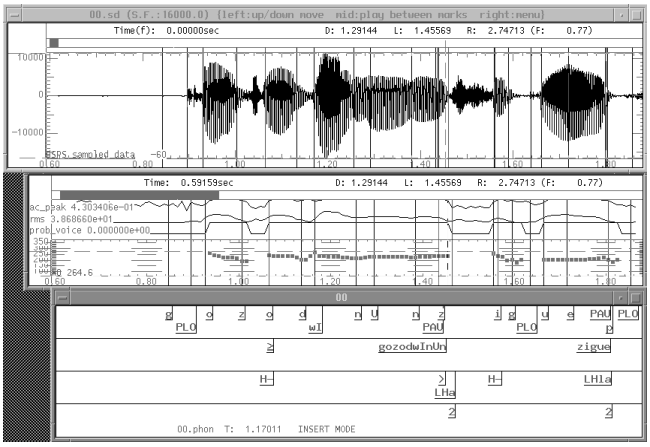

Figure 1: One example for prosodic labeling

The more detailed prosodic labeling is in progress. For example, as our very next step, we will add a discourse act for investigating discourse structure.

6. PRELIMINARY RESULTS

To investigate the characteristics of multi-level speech database, we statistically analyzed the transcription texts as a first phase. The text data used for statistical analysis was 112 sets. The results are shown in Table 3.

As expected, the dialogue speech has more corrections, repetitions, and pauses than the others. In addition, the average number of sentences per turn in dialogue speech is also greater than those in other types. Even though some other parameters have to be further exploited, we hope the above features can also be useful in quantifying the degree of spontaneity in Korean.

7. DISCUSSION

Although the main motivation for this kind of database is to develop a dialogue type speech synthesizer, still we can not deny the usefulness of our database in spontaneous speech recognition. In our future study, we will try to get more meaningful statistical prosodic analysis results from the comparative study among the different categories' data. And the results from this study will be utilized to implement the dialogue type Korean speech synthesizer by adding spontaneous prosody patterns to our current TD-PSOLA(Time-Domain Pitch Synchronous Overlap and Add) based text-to-speech system[7].

References

[1] Sanghun Kim, Jung-Chul Lee. Youngjik Lee, "Korean Text-to-Speech System Using Time-Domain Pitch Synchronous Overlap and Add Method", in *Proc. fifth Australian Int. Conf. on Speech Science and Technology,* Perth, Australia, pp. 587-592, Dec., 1993.

[2] G. Bruce, "On the analysis of prosody in interaction," in *Int. Workshop on Computational Modeling of Prosody for Spontaneous Speech Processing,* pp.2.35-2.38, Apr., 1995.

[3] Nakajima, Tsukada, "Prosodic Features of Utterances in Task-Oriented Dialogues," in *Int. Workshop on Computational Modeling of Prosody for Spontaneous Speech Processing,* pp.2.39-2.44, Apr., 1995.

[4] Bertenstam J., et al,"The Waxholm Application Database", in *Proc. EUROSPEECH'95,* Vol. 1, pp. 833-836, Sep., 1995.

[5] N. Campbell, "Mapping from read speech to real speech," in *Int. Workshop on Computational Modeling of Prosody for Spontaneous Speech Processing,* pp.3.20-3.25, Apr., 1995.

[6] G. Bruce, et al, "Speech Synthesis In Spoken Dialogue Research," in *Proc. EUROSPEECH'95,* pp.1169-1172, Sep., 1995.

[7] Jung-Chul Lee, Sanghun Kim and Minsoo Hahn, "Intonation processing for Korean TTS Conversion Using Stylization Method," in *Proc. ICSPAT'95,* pp.1943-1946, Oct., 1995.

	DIAG	INV	SINV	DSWC	DSWOC
Average duration(second)	317	77	64	70	98
Average number of turns/speaker	37.4	14.1	13.7		
Average number of answer sentences/turn	2.50	1.39	1.43		
Total number of sentences (Q:Question, A:Answer)	793(Q), 2523(A)	323(Q), 554(A)	303(Q), 486(A)	540	540
Average number of correction/sentence	0.40	0.09	0.06		
Average number of repetition/sentence	0.05	0.02	0.02		
Average number of pause/sentence	1.19	0.80	0.34		

Table 3: Preliminary results on textual analysis