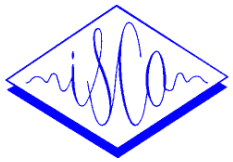

ISCA & IEEE Workshop on Spontaneous
Speech Processing and Recognition
Tokyo Institute of Technology, Tokyo, Japan
April 13-16, 2003

PRONUNCIATION VARIANTS MODELING IN KOREAN SPONTANEOUS SPEECH RECOGNITION

Kyong-Nim Lee and Minhwa Chung

Dep. of Computer Science, Sogang University
1 Shinsu-Dong, Mapo-Gu, Seoul, Korea
{knlee, mchung}@sogang.ac.kr

ABSTRACT

Pronunciation variants in spontaneous speech tend to be more variable in planned speech. Spontaneous speech has significant sources of variations as well as serious phonological variations, which make recognition extremely difficult. In this paper, we analyzed the auditory transcriptions of the dialogue for spontaneous speech recognition, and then classified the characteristics of conversational speech. To deal with these characteristics, we first used the special garbage model, the silence model and the filled pause model for the improvement the acoustic model; second, we optimized the multiple alternative pronunciations using the pruning method. Finally, for reflecting on freely the phonological variation, we enhanced the pronunciation lexicon by adding alternative pronunciation based on the frequently used phonological variants. Experimental results showed that modeling of garbage, silence, and filled pause reduce word error rate by a relatively 4.9%, while pruning the lexicon and adding the alternative pronunciation reduced word error rate by relatively 0.8%.

1. INTRODUCTION

Spontaneous speech is notably different from planned speech in several ways. In particular, the former added a variety of phenomena to speech recognition tasks: false starts, human noises, environment noise, new words, and alternative pronunciations [4][5]. In this paper, the characteristics of spontaneous speech and its recognition system were dealt with and analyzed pronunciation modeling as a means of improving training phonetic transcriptions for acoustic model and pronunciation lexicon of spontaneous speech, particularly lexicon pruning are examined.

In the previous research on Korean spontaneous speech recognition, [2] implemented as a part of an automatic translation system and [8] obtained the baseline system for conversational speech recognition by reflecting dialogue characteristics.

In this paper, we analyzed the auditory transcriptions of the real conversational speech, and then classify the characteristics of conversational spontaneous speech in view of the speech recognition. The pronunciation variants modeling were divided into two broad categories. In dealing with the conversational speech characteristics, we used the garbage, silence, duration, and filled pause model for improving the acoustic model. In pronunciation lexicon modeling, we applied the lexicon pruning method for improving performance and extend the lexicon by reflecting free the phonological variation. In addition, we also enhanced the pronunciation lexicon by adding alternative pronunciation for most frequent phonological variations.

2. SPONTANEOUS SPEECH RECOGNITION

2.1 Databases and Task

Spontaneous Korean corpus is developed by our laboratory under contracts of ETRI for part of the project C-star in 1999. The speech data used in this study where from 50 native Koreans. The data were recorded via a headset microphone to digital audiotapes, and digitalized at 16 KHz and 16-bit sampling. Within the total recording time (8.75 hours), the recording of real speech sounds lasted for 7.5 hours.

In the course of our experiment, it was our principle to exclude speeches that were completely free (for example, switchboard task) and use only the situation scenario. From 25 groups, each of which conducted five dialogues about 15 scenarios of hypothetical condition for travel schedule. The dialogues where segmented into utterance and the text size in total were 68,947 words (=Korean morphemes).

For language model training, all transcribed text were used. The training-set for acoustic model training consisted of 100 dialogues and test-set for evaluation had 25 dialogues specified in Table 1.

Table 1. Size of database for train and test

	Training	Test	Total
Dialogue	100 (20 groups)	25 (5 groups)	125 (25 groups)
Utterance	4,437	1,054	5,491

2.2 Transcription Rules

To do this, we listened to a direct voice and transcribe the text including the noise, filled pause, repetition or repair, and pronunciation variants. We used to the bar symbol "/" to separate the auditory transcription based on the actual pronunciation written on the left side, and the orthographic transcription presented on the right side in order to generate a language model and analyze the dialogue phenomena. This transcription model also included both the informal and ungrammatical speech found in the actual recording.

2.3 Analysis of Spontaneous Speech Characteristics

In Korean language, spontaneous speech contains ungrammatical as well as serious phonological variations compared with planned speech. Typical examples of spontaneous speech errors are phonological contraction or deletion, vowel alternation according to vowel harmony, attaching final auxiliary particle at the end of sentence, combination case and auxiliary particle, frequent deletion of particle, sentence fragment and insertion of filled pauses, etc [3][5].

In this study, we only analyzed the main factor of decreasing the speech recognition performance and not considered all kinds of spontaneous characteristics. We dealt with the transcriptions of the conversation, and then classified the characteristics in the speech recognition aspect. Table 2 summarizes the characteristics of the analysis in case of disfluencies[1].

Table 2. Classification of Korean spontaneous speech characteristics about disfluencies [8]

Classification		Example
Disfluencies	Noise	Human garbage, pause, Environment noise
	Filled pause	ye/ ne/ jeo/ eo/ mwo/
	Repetition / Repair	mat/ matsseumnikka yeyak/ yeyakasyeotsseumnida

Based on the data gathered, 1,710 filled pauses occurred among the 36,142 spacing units (see Table 3 for the distribution of noise). Mostly, the acceptance and affirmation expressions 예(ye) and 네(ne) covered 72.1% of the total occurrence of filled pauses. Next table 3 shows the distribution of noise in total 7,916 occurrences. Table 4 shows the examples of pronunciation variants of word compared to their canonical or standard.

Table 3. Distribution of Garbage/Noise Model

Symbol	Occurrence	Distribution (%)
Lip sound (ls)	3,536	44.7
Environment noise (N)	1,903	24.0
Breath (h)	1,702	21.5
Total	7,916	100

Table 4. Examples of having various pronunciation variants per one word (( ) is represented meaning of English)

Standard	Non-standard pronunciation variants
Geureomyeon (Then)	geum, geureom, geureum, geureumeun, geureomyeoneun
Geurigo (And)	geurigu, geureugo, geurogu
Eotteoke (How)	eotteuke, eoteokke, eoteukge, eoteuke, eoteoge, eoteoke, eoteuge

It is important to note that in Korean, spontaneous speech has many phonological variations that could not be treated with rule generation. Specially there is final auxiliary particles 요(yo) mispronounced 여(yeo). Other variations in Korean phonology include sound contractions or deletions, frequently shifting of light vowel into dark vowel and free patterns of phonological expression.

3. PRONUNCIATION VARIANTS MODELING

3.1 Acoustic Model

The performance of an automatic speech recognition system greatly depends on the acoustic modeling quality. For a stable acoustic modeling, we perform the adaptation method simultaneously using the maximum likelihood linear regression (MLLR) and the maximum a-posteriori (MAP). We used 4,437 utterances for this adaptation. Pre-constructed acoustic models for dictation system have been designed for common use on 43,000 sentences that covered various phonological environments and balanced sub-word units in this case; we used triphone models. The adaptation training data are insufficient to consider the noise, silence and filled pauses. Hence, we used the constructed models [6] for human garbage and noise modeling. In the case of silence model, we separately trained the silence model in the sentence using the spontaneous speech data.

The speech signal was sampled at 16 KHz and segmented into 25 ms frames with each frame advancing every 10 ms. Each frame was parameterized by 39-D feature vector that consists of 12 MFCC parameters and their differential coefficients of the 1st and 2nd order, together with power, and their corresponding time derivatives. Each of the sets is modeled by tied-state triphone CHMM model which is a left-to-right model of five states. Each state, consisting of six mixtures, had been trained and tested using the Baum-Welch and Viterbi algorithms.

3.2 Pronunciation Lexicon

We used the multiple pronunciation lexicons, which have the possible alternative of standard pronunciation in one morpheme. Also, we designed the pronunciation lexicon without out-of-vocabularies, but with all fragment word and filled pauses. The total entry sizes are 1,126 morphemes in the 68,882 occurrences. Since Korean is an agglutinative language, spacing unit is not similar to English. It is *eojeol*, which is a spacing unit of Korean orthography that combines substantial and formal morphemes. Korean pronunciation variations depend on morphological categories and morpheme boundary information as well as phonemic contexts. Thus, morpheme-based pronunciation lexicon is required to cope with pronunciation variations, especially in cross-morpheme, including within-morpheme.

3.3 Lexicon Pruning Method

There are a wide variety of approaches, including selecting the most frequently observed variants in a corpus. In this work, we follow the general approach in order to decrease the confusion of words and reduce the lexicon size. Three pruning criteria for determining the number of pronunciations for each word in the lexicon were utilized [4]. We also used the analysis of hand transcriptions by enhancing the pronunciation lexicon using the frequently used pronunciations in order to reflect the various pronunciation variants.

1) Frequency-based pruning: list of pronunciations for each word ordered by count criterion in the corpus.
2) Probability-based pruning: pronunciations are pruned with prior probabilities less than a parameter α .
3) Recognition-based pruning: pronunciations are obtained through forced alignment using the phone recognition.

4. EXPERIMENTAL RESULTS

4.1 Garbage Modeling

We made one garbage/noise model and one silence/pause model. In our experiments, our baseline word error rate (WER) is 14.92%; we obtained WER reduction such as 0.39% (relative 2.6%) for garbage model and 0.73% (relative 4.9%) for silence model.

4.2 Disfluencies Modeling

Previous work [3] indicated that filled pauses contain information for the prediction of neighboring words. Similarly, we found out WER reduced to a rate of 0.44% (relative 3.1%) WER. However, we decided the baseline performance using language model excluding disfluencies probability. We also considered the duration model in case of filled pauses; frequently used response word 예 (ye) and 네(ne). Generally, the duration time was longer than other phonetic model. Therefore, we compensated on the duration time adding the same phone into the pronunciations. *e.g.* '네 N EY EY'. We obtained the 0.29% (relative 2.1%) improvement of WER.

4.3 Pronunciation Lexicon Modeling

The results in figure 1 show the performance of the designed pronunciation lexicon. We implemented the pronunciation variants modeling in order to enhance the lexicon. To build the pronunciation lexicon, we used these four methods; 1) frequency-based ranking 2) probability-based pruning 3) recognition-based ranking 4) stochastic lexicon with assigned probability value in each word.

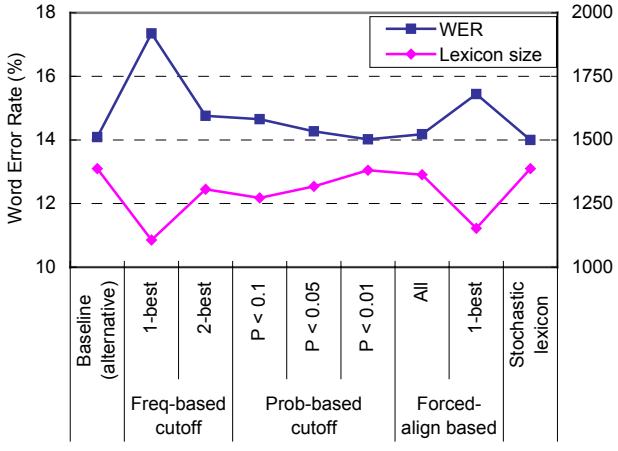

Figure 1. Word error rates (%) and size of alternative pronunciation lexicon following several lexicon-building methods. The baseline has all possible alternative pronunciation.

In our experiment, the canonical lexicon containing all possible standard alternative pronunciation has 1,387 entries (=average 1.2 per word) and 14.09% WER. Frequency-based ranking for 1-best single pronunciation lexicon decrease lexicon size but deteriorated the recognition performance. When choosing the entries according to distribution threshold, the lexicon size decrease. The best performance was 14.02% WER using 0.01 threshold. In addition, recognition-based lexicon using forced alignment showed better performance than frequency-based with aspect to 1-best lexicon.

Meanwhile, we also performed the pronunciation lexicon modeling reflecting the free phonological variants. Based on our experiment, it was difficult to generate rule and to manage the changeable pronunciation of inter-speakers. As many researchers have observed, simply adding several alternative pronunciations to the dictionary increases the confusable words to the extent that the gains from having them are often more than nullified [4][7]. We also gather the similar result, shown in Figure 2. We particularly treated this problem using lexicon pruning method as already explained in 3.3.

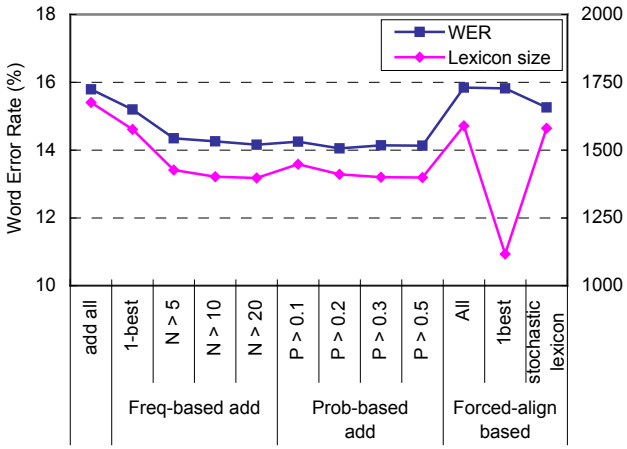

Figure 2. Recognition performance after adding the pronunciation variants. We added alternative pronunciation into canonical lexicon based on several methods (see 3.3).

Table 5. Recognition performance of best result in case of each pronunciation lexicon

Lexicon	Description	WER(%)	Size
Standard Pronunciation Lexicon	(a) All possible alternative baseform	14.09	1,387
	(b) Best performance using pruning	14.02	1,381
Expended Pronunciation Lexicon	(a) + all hand-labeled alternative variants	15.79	1,676
	(a) + pruned variants	14.05	1,411
	(b) + pruned variants	13.98	1,401

As shown in the Table 5, the best performance in the standard pronunciation lexicon is 14.02% WER. However, in case of the use of probability-based pruning (0.02 threshold) in expanded pronunciation lexicon, the WER resulted in 13.98%. This means that adding phonological variants to the standard lexicon helps in decreasing the WER.

5. CONCLUSION

Given the inevitable Disfluencies in conversational spontaneous speech, we found the imperative of designing suitable pronunciation lexicon reflecting the phonological variants. This pronunciation lexicon supplements the dictionary-based standard pronunciation lexicon and offers systematic expanded pronunciation variants.

What can be considered for future study is the construct of alternative pronunciation variants using the confusion matrix or rule-based generation, which can be applied to certain phonological analyses of language except the hand-labeled pronunciation.

6. ACKNOWLEDGEMENT

This research was supported by Korea Science and Engineering Foundation (PN M10107000015-01A2200-01030). Also, the authors are grateful to HCI Lab, SAIT for providing the read speech DB and ETRI for providing the conversational speech DB used in the experiments.

7. REFERENCES

[1] E. Shriberg, "Preliminaries to a Theory of Speech Disfluencies," *Ph. D. thesis*, Univ. of California, Berkeley, 1994.

[2] H.-S. Lee, J. Park, O.-W. Kwon, "An implementation of Korean Spontaneous Speech Recognition System," *Proc. of the Workshop on Speech Communication and Signal Processing of Korea*, pp. 145-147, 1996.

[3] M. Siu, M. Ostendorf, "Modeling Disfluencies in Conversational Speech," *Proc. of ICSLP*, pp. 386-389, 1996.

[4] M. Tsai, F. Chou, L. Lee, "Improved Pronunciation Modeling by Property Integrating Better Approaches for Baseform Generation, Ranking and Pruning," *Pronunciation Modeling and Lexicon Adaptation for Spoken Language Workshop*, 2002.

[5] M. Saraclar, "Pronunciation Modeling for Conversational Speech Recognition," *Ph.D. thesis*, Johns Hopkins Univ., 2000.

[6] K.-N. Lee, M. Chung, "Human Garbage Modeling for Processing Human Noise in Korean Dictation System," *Proc. of the Conf. on Acoustical Society of Korea*, pp. 323-326, 2001.

[7] W. J. Byrne, et al, "Pronunciation Modeling Using a Hand-labelled Corpus for Conversational Speech Recognition," *Proc. of ICASSP*, pp.313-316, 1998.

[8] Y.-H. Park, M. Chung, "Analysis of Korean Spontaneous Speech Characteristics for Spoken Dialogue Recognition," *The Journal of the Acoustical Society of Korea*, 21(3), pp.330-338, 2002.