THE JOURNAL OF THE ACOUSTICAL SOCIETY OF KOREA Vol.21, No.1E 2002. 3. pp. 3~11

Korean Broadcast News Transcription Using Morpheme-based Recognition Units

Oh-Wook Kwon*, Alex Waibel**
*Brain Science Research Center, KAIST, Korea
**Interactive Systems Laboratories, University of Karlsruhe, Germany

(Received January 1 2001; revised Feburuary 7 2002; accepted March 13 2002)

Abstract

Broadcast news transcription is one of the hardest tasks in speech recognition because broadcast speech signals have much variability in speech quality, channel and background conditions. We developed a Korean broadcast news speech recognizer. We used a morpheme-based dictionary and a language model to reduce the out-of-vocabulary (OOV) rate. We concatenated the original morpheme pairs of short length or high frequency in order to reduce insertion and deletion errors due to short morphemes. We used a lexicon with multiple pronunciations to reflect inter-morpheme pronunciation variations without severe modification of the search tree. By using the merged morpheme as recognition units, we achieved the OOV rate of 1.7% comparable to European languages with 64k vocabulary. We implemented a hidden Markov model-based recognizer with vocal tract length normalization and online speaker adaptation by maximum likelihood linear regression. Experimental results showed that the recognizer yielded 21.8% morpheme error rate for anchor speech and 31.6% for mostly noisy reporter speech.

Keywords: *Continuous speech recognition, Broadcast news transcription, Language model, Morpheme-based speech recognition*

I. Introduction

Broadcast news transcription was a cutting-edge research topic during the last few years in the speech recognition field because broadcast news signals have many adverse conditions in speaking mode and noise compared with the conventional large vocabulary continuous speech recognition [1,2]. There can be two kinds of speaking modes in a broadcast news program: prepared speech by anchors and spontaneous speech by reporters or interviewers. Speech by anchors is mostly clean. But speech by reporters usually has noisy background signals: background noise, background music, or background speaker[2]. Speaking speed of broadcast news speech is faster than that of dictation by ordinary people because anchors and reporters are well trained to pronounce news scripts. For these reasons, broadcast news transcription is a more difficult task than a dictation task even though both tasks require speaker-independent large vocabulary continuous speech recognizers.

There has been much attention to broadcast news transcription because it is a good task to tackle the adverse conditions described above. The Defense Advanced Research Projects Agency (DARPA) has supported research on broadcast news transcription from 1996 to 1999[1,3,4]. Recent benchmark test results by the DARPA showed that the CU-HTK system achieved the

Corresponding author: Oh-Wook Kwon (ohwook@ucsd.edu)
Institute for Neural Computation, University of California, San Diego 9500 Gilman Drive, La Jolla, CA 92093-0523, USA

lowest word error rate (WER) 16.1% when recognition time is restricted to 10 times real time[3]. Without the time restriction, the best system showed 7.8% WER for clean speech (F0 focus condition), 15.1% for spontaneous speech (F1), 13.6% for noisy background (F4), and 13.5% when averaged over all conditions. For Japanese broadcast news speech recognition, a recent study using morpheme-based recognition units reported 19.7% WER for anchors and 38.2% WER for other speakers[5,6]. For Korean broadcast news speech recognition, only a few have reported: a speech recognizer for a limited weather domain[7] or an experimental system using syllable-based recognition units [17]. To the author's knowledge, there have been no reports that had achieved good accuracy on Korean broadcast news speech recognition in whole news domains.

For Korean broadcast news transcription, we require large vocabulary continuous speech recognition (LVCSR) systems. LVCSR systems for Korean have much higher out-of-vocabulary (OOV) rate than for European languages when the word phrase (eojeol) unit is used as recognition unit for LVCSR[8]. For European languages, the word unit in dictionaries and language modes is well defined and keeps the OOV rate to a manageable level. However, for agglutinative languages such as Korean, Japanese, Serbo-Crotian, the OOV rate becomes very high because a verb can have much more inflected forms than for European languages. In Japanese case, a sentence does not have any spaces for word phrase separation. When a word phrase is used as the basic recognition unit, the OOV words increases and recognition accuracy becomes worse consequently.

There have been a few efforts to use units other than word to cope with the high OOV problem in languages with heavy inflection. Although one can reduce vocabulary size by using a unit smaller than a word and thus can reduce the test-set perplexity, overall word recognition accuracy is not improved because the average phoneme count of recognition vocabulary words is small. Phoneme or syllable-based continuous speech recognition is inferior to morpheme or word-based speech recognition[9]. One can separate a word into morphemes and use the morphemes as the recognition units[10,11]. Furthermore one can concatenate a sequence of frequent words into an ensemble of words[12]. Recently automatic concatenation procedures based on the frequency, mutual information, log-likelihood, or perplexity criteria have been proposed[13,14].

We developed a broadcast news speech recognizer for Korean broadcast news. We used morpheme as the unit of language models to cope with high OOV rate. Morpheme-based unit has been often used for speech recognition of languages with inflections[6,8,10]. For Korean speech recognition, we may consider eojeol, morpheme, or syllable as a candidate of recognition units. We selected morpheme because we can reduce the OOV rate to a comparable level of European languages. A morphology analysis tool is modified so that the grapheme form of a morpheme is maintained and hence the original word can be obtained by concatenating morphemes. Then we concatenated short or frequent morphemes according to a linguistic knowledge-based rule and then added it in the vocabulary as a new unit. To get good recognition accuracy in morpheme-based speech recognition systems, we should have a good tagging system. More discussion on unit selection for Korean speech recognition is found in [9,16,17].

We evaluated performance of the Korean broadcast news speech recognizer using 2 broadcast news episodes. The recognizer showed 21.8% morpheme error rate for anchor speech and 31.6% for noisy reporter speech. The results are inferior to those for American English but encouraging because the size of speech database and text corpora is far smaller.

The organization of this paper is as follows. In Section II we describe speech databases and text corpora used in this work. In Section III we discuss acoustic modeling, language modeling, and decoding algorithms for morpheme-based LVCSR. In Section IV we present experimental results. Finally conclusions are drawn in Section V.

II. Collection of Broadcast Speech Data and Preprocessing of Text Corpora

The speech database consisted of 16 episodes of Korean

Broadcasting System (KBS) News selected during the period from July 1 to August 3, 1998. The speech data were obtained by extracting audio signals from videotapes [17]. Originally anchor and reporter speech, music, foreign speech was included. The average duration of a broadcast news episode was 3070 seconds (51 minutes). Of the 16-day data, two episodes of July 20 and 29, 1998 were used as a test set. The remaining 14 episodes were all used as a training set. Two anchors appeared on each episode. A male and a female anchors appeared on the test set. The two anchors in the test set were also in the training set. The test set included 53 reporters and 11 reporters were independent of the training set. We used only speech data from anchors and reporters for test and excluded weather forecasting. Table 1 shows the number of clean and noisy utterances for each date and speaker kind of the test set. The 174 utterances from anchors were all clean. The number of utterances from reporters was 571. The 67% of reporter utterances had background noise. The background noise included noisy signals from natural phenomena, other speech, and music. The speech data were manually segmented into utterances. Initial transcriptions were obtained from manuscripts supplied in a web site. The transcriptions have been corrected by listening to the speech data and noise marks were properly inserted into the transcriptions.

We gathered 2.5 years of broadcast news manuscripts from two broadcast companies, KBS and Munhwa Broadcasting Corporation (MBC). The gathered 6.6M eojeol manuscripts were too small to be used for reliable estimation of statistical language models. Therefore we gathered 2 years of newspaper articles from the Dongailbo and Chosunilbo newspaper articles from web sites. The text data should be preprocessed to be used as a text corpus for language modeling. First we should strip format-related parts and extract text information only. We converted Chinese characters into corresponding Korean characters. We converted symbol units for weights and measures into corresponding Korean units with the same pronunciation. Some symbols denoting the range were also converted to appropriate Korean texts. In Korean a number is pronounced differently depending on its usage is cardinal or ordinal. Therefore we identified a number as cardinal or ordinal and then converted it into appropriate pronunciation. The Chosunilbo texts had much more spacing errors compared with others because of line breaks within eojeol units. We corrected the spacing errors at line breaks using syllable trigram probabilities. The text corpora after preprocessing are summarized in Table 2. We discarded text data of KBS News from July 1998 because the test set of speech database was selected from that period. The amount of text corpora in this work is still less than the amount of text data used for the DARPA project, 550M[26].

Table 1. Test set.

Date	speaker	# of clean utterances	# of noisy utterances	# Total
98.07.20	anchor	88	0	88
	reporter	105	164	269
98.07.29	anchor	86	0	86
	reporter	82	220	302
# Total		361	384	745

Table 2. Text corpora.

Corpus	Duration	# of eojeols
KBS News	1996.09~1998.10	3.1M
MBC News	1996.08~1998.12	3.5M
Dongailbo	1996.10~1998.11	23M
Chosunilbo	1996.01~1998.11	52M

III. Morpheme-based Speech Recognition

Except language models we can use techniques for LVCSR same as for European languages. Language-specific parts in speech recognizers are defining phoneme-like units, classification of phoneme categories for decision trees, defining units for lexical tree construction and language modeling. We can define phoneme-like units and phoneme categories by referring to linguistic knowledge sources. Language modeling for European languages is relatively easier that for agglutinative languages or Chinese because there are clear clues for the word unit in European languages. In Japanese and Chinese, no segmentation is provided in the written text and hence the first step to speech recognition should be definition of word units. Even though a Korean sentence has spaces to segment word phrases fortunately, we still have the problem of defining appropriate word unit because we have too large OOV rate if the word unit is used as recognition units.

In the Korean language, a noun can have suffixes representing case and a verb is heavily inflected. There may be pronunciation change between two morphemes by some morphological rules, especially in case of two consecutive consonants. That is, a stem of a verb with the same meaning may have different graphemes depending on its following endings. We used a modified Korean morphology analyzer to produce a sequence of morphemes maintaining its phoneme information[18]. Thus the resulting morpheme sequence of a word can be concatenated to produce the word. Preliminary experiments showed that short morphemes contribute to large recognition errors. Therefore we merged morpheme units to reduce recognition errors due to short morphemes [8]. By merging, we can extend the context width of a language model. We merged the following morpheme pairs. A short morpheme with a single consonant was merged to the preceding morpheme. A suffix denoting noun or adjective forms of a verb was combined with the following particle if it existed. Otherwise it was merged to the preceding stem of the verb. Inflected forms of an auxiliary verb were used as recognition units without dividing into morphemes. That is, a word phrase of an auxiliary verb was used as a recognition unit. Compound nouns were divided into component nouns by the morphological analysis tool. Prefixes were not segmented into a recognition unit because each prefix is usually short and only a small number of nouns can be prefixed depending on contexts.

In Korean, pronunciation of a morpheme is subject to hard morphological rules depending on adjacent morphemes. When a word has several different pronunciations, we used multiple entries in the pronunciation dictionary and language model to cope with phonological variations between morpheme boundaries[8]. A similar approach was used for Japanese broadcast news transcription. A speech recognizer with a reading-dependent language model improved recognition accuracy in spite of increased OOV rate[5]. One can represent each morpheme in a graph form in order to reflect between-word phone variations[27]. In this case the transitions at word-initials and word-ends become very complex. We constructed a lexicon automatically by using a text-to-pronunciation tool[19].

We search a lexical tree in the first pass by using approximate trigram scores[20], applied vocal tract length normalization (VTLN) to the tree search result, search a flat lexical tree to use exact trigram scores, and construct and rescore lattice constructed from the second stage search result[21]. Assuming the current recognition result is correct, we perform speaker adaptation by maximum likelihood linear regression (MLLR). Then the above two stages, flat lexical tree search and lattice rescoring, are repeated to produce final recognition results. In the tree search, we distribute unigram scores in the search tree to apply language modeling as early as possible.

Figure 1 shows a lexicon and a search tree with cross-word triphone modeling used in the first pass assuming that 4 morphemes are to be recognized: three nouns '가', '갑', '곱' and 2 josas '이', '을'. We note that a josa cannot follow another josa in Korean. The square boxes denote dummy nodes that do not emit observation symbols. The name *l-c+r* denotes a phoneme */c/* with left context */l/* and right context */r/*. In this example, *l*, *c* and *r* represent an arbitrary phoneme in the lexicon: */a/*, */b/*, */eu/*, */g/*, */i/*, */l/* and */o/*. The S_{c+r} denotes a start node for context *c+r* and the E_r denotes an end node for right context *r*. For each end node at each time frame, we get a leaf node with the best score from all leaf nodes with the same right context phoneme. The double-lined ellipsoids denote word-initial nodes with dynamic triphone mapping[20]. A word-initial node dynamically inherits the last winning predecessor. Identification of a word-initial node is determined according to the preceding left phone. The leaf nodes are dynamically attached. In the case of single-phone words, all nodes are derived dynamically. At the transitions numbered in the tree, an approximate trigram score is added to accumulated score[20].

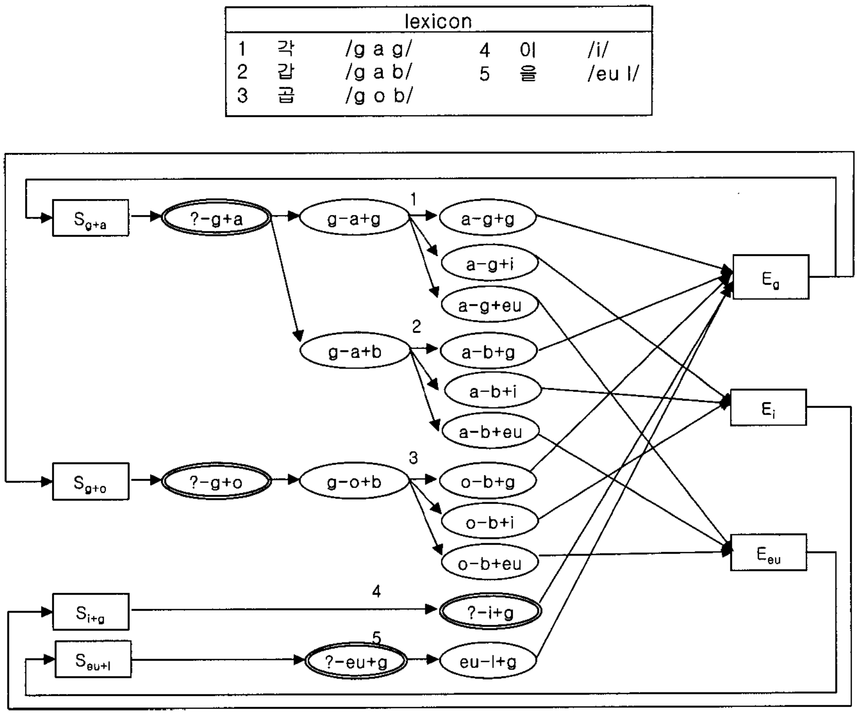

Figure 1. A lexicon and a search tree used in the forward tree search.

IV. Experimental Results

Speech signals were sampled at 16 kHz to produce 16 bit data. The speech signals were framed into 16 ms blocks at the rate of 10 ms. We extracted a feature vector for each frame comprising 12-order mel-frequency cepstrum coefficients (MFCC), first- and second-order differential coefficients, energy and first- and second-order differential energy. The cepstral coefficients are normalized on an utterance basis using cepstral mean subtraction. Then linear discriminant analysis (LDA) was performed to produce 24 dimensional feature vectors. For acoustic modeling, we used 40 basic phonemes including silence. For observation probability, we used senone-based acoustic modeling with cross-word coarticulation considered[7]. We used position-dependent quinphone modeling to model intraword and cross-word context dependency. For cross-word modeling, we only considered a single phoneme across word boundaries. We used 3,000 senones and each senone had a probability distribution defined over 16 Gaussian mixtures. For each model except silence, we used a 3-state hidden Markov model (HMM) without skip transition. We modeled silence using a 3-state HMM with all observation probability distributions shared. We clustered contexts by a decision tree. We used 47 phoneme categories (e.g., vowel, consonants, fricative, front-vowel, back-vowel, and so on) as questions in the decision tree.

Figure 2 shows the experimental setup used in this work. We used the JRTk recognition toolkit[21] to train the acoustic model and the SLM toolkit[22] to compute the trigram language model. We used a part-of-speech (POS) tagger[18] to segment eojeols into morphemes. A pronunciation dictionary was automatically obtained by using a morpheme-based grapheme-to-phoneme converter[19].

The vocabulary size was 64014 including human and nonhuman noise signals. We computed probabilities for bigram and trigram entries for which the number of samples are large than 2. We used the backoff smoothing method to estimate probabilities without samples[23]. When we used morphemes as recognition units, the OOV rate was 1.7%. The test-set perplexity was 196 when OOV words and all context cues (e.g., sentence-begin and

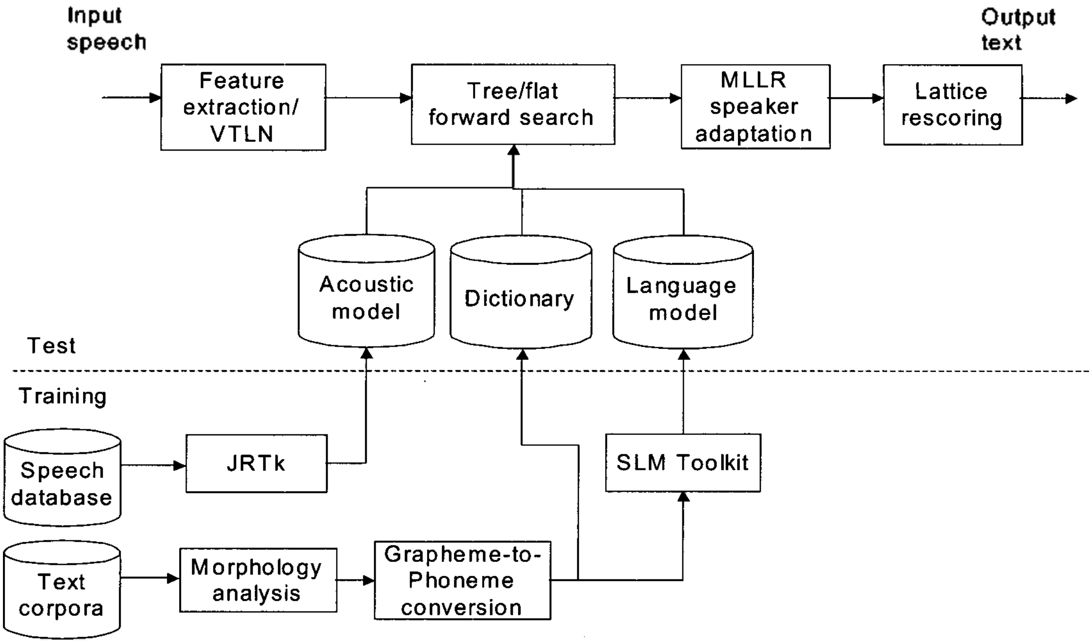

Figure 2. Block diagram of experiment setup.

sentence-end marks) were removed from calculation. The number of items in language models was 1.7M and 2.8M for bigram and trigram, respectively. The hit ratios were 52%, 35%, and 13% for trigram, bigram, and unigram, respectively. Compared with the DARPA HUB-4E experiments[24,25], the OOV rate and the perplexity are high. The OOV rate was 0.7% and the perplexity was 172 on 1996 HUB-4E evaluation set with 65185 words[24].

In the baseline recognizer, we used 30M eojeol texts including the KBS, MBC, and Dongailbo corpora to obtain vocabulary and trigram language model. Table 3 shows the language model used in the baseline recognizer. The OOV rate was 1.7%, which means that percentage of morphemes in the test set were not included in the recognition vocabulary. The performance of the baseline recognizer is shown in Table 4. In the table, TF denotes accuracy of tree forward search results, FF denotes morpheme accuracy of flat forward search results, LAT means lattice rescoring, and FINAL means word error rates after both VTLN and MLLR speaker adaptation have been applied. The ANC and REP represent anchor and reporter, respectively. We presented performance in the TF case for we cannot process speech signals backward in some applications where on-line processing is required (e.g., captioning a live broadcast news program). The best performance was obtained with language model weight 24 and insertion penalty 4. The average word (morpheme) error rate was 23.7% for anchor speakers 37.4% for reporter speakers when the beam size is 120. In this paper the WER for each speaker class was obtained by averaging WERs for the two episodes. When VTLN was applied after the first-stage tree forward search, there was a little

Table 3. Language model used in this work.

Vocabulary size		64014
OOV rate (%)		1.7
# entries	unigram	64k
	bigram	1.7M
	trigram	2.8M
Hit ratio (%)	unigram	13
	bigram	35
	trigram	52

Table 4. Word error rates (%) of the baseline system.

Date	Speaker	TF	VTLN	FF	LAT	FINAL
98.7.20	ANC	25.5	25.9	26.0	24.5	24.9
98.7.20	REP	35.5	35.9	35.2	34.8	31.7
98.7.29	ANC	23.2	22.3	21.4	21.2	20.7
98.7.29	REP	40.2	39.1	39.6	38.5	36.8
Average	ANC	23.7	24.1	23.7	22.7	22.8
	REP	37.4	37.5	37.4	36.7	34.3

improvement. When we applied both VTLN and MLLR speaker adaptation, approximately 5% of recognition errors were reduced.

In the next experiment, we used all text corpora prepared to obtain vocabulary and language models. In the experiment, we tested only anchor speech of 1998.07.20. The number of items in language models was 64,017, 3.5M, and 7.3M for unigram, bigram, and trigram, respectively. As shown in Table 5, for anchor speech, we achieved 5.6% relative improvement over the baseline recognizer. However for reporter speech, there was no improvement. The marginal improvement was due to many spacing errors in the added text corpus. The errors resulted in a large number of morphological analysis errors. We need more efforts to get clean text data. We did not continue the experiment with the speech data of 1998.07.29 because the above result did not yield significant improvements considering the increased complexity.

We added a question in the decision tree whether a phone is at boundaries of segmented morphemes. To be specific, the first phone of a particle and an ending was handled differently. We call it morpheme-boundary modeling. As shown in Table 6, performance without morpheme-boundary modeling was similar to performance with the conventional word-boundary modeling. Relative difference of performance was about 1%. Therefore we did not use morpheme-boundary question in the decision tree. But we note that word-boundary questions are still in use.

In the final experiment, we augmented the training set with 36-day broadcast news provided by ETRI. First we trained the recognizer without increasing the number of senones but we did not achieve any improvements. It implies that the number of senones in the recognizer was already optimized for best performance. Therefore we increased the number of senones to 5,000 so that a new recognizer can model acoustic feature with the augmented training data. As shown in Table 7, the new recognizer showed word error rate of 21.8% for anchor speech and 31.6% for noisy reporter speech.

Table 5. Word error rates (%) with increased text data.

Date	Speaker	FINAL
98.7.20	ANC	23.5
	REP	31.6
98.7.29	ANC	NA
	REP	NA

Table 6. Word error rates (%) with morpheme-boundary modeling.

Date	Speaker	TF	FINAL
98.7.20	ANC	35.3	23.8
98.7.20	REP	36.1	32.6
98.7.29	ANC	21.8	19.8
98.7.29	REP	40.5	36.4

Table 7. Word error rates (%) with increased speech data and number of senones.

Date	Speaker	TF	FINAL
98.7.20	ANC	24.5	23.3
98.7.20	REP	33.2	30.4
98.7.29	ANC	20.7	20.2
98.7.29	REP	37.2	33.7
Average	ANC	22.6	21.8
	REP	35.2	31.6

V. Conclusion

We investigated performance of Korean broadcast new transcription. To reduce the high OOV rate in Korean LVCSR, we used morpheme-based units for language modeling. The original morpheme pairs with short length or high frequency were merged into a larger unit to reduce insertion and deletion errors. We built a lexical search tree with multiple pronunciations to reflect inter-morpheme pronunciation variations according to the Korean phonology rule. This approach has the advantage that we do not need to modify the search tree. We evaluated performance of the recognizer using 2 broadcast news episodes. The speech signals of reporters mostly had noisy background. The baseline recognizer showed 23.7% WER for anchor speech and 37.4% for reporter speech. MLLR speaker adaptation reduced 5% of recognition errors. We increased the text corpora but performance was not improved. The added text data need further text normalization by correction of segmentation and orthographical errors. Morpheme-boundary modeling does not improve recognition accuracy but cross-word modeling with multiple pronunciations is good for acoustic modeling. We improved performance of the recognizer to 21.8% WER for anchor speech and 31.6% for reporter speech by increasing speech data and the number of acoustic model parameters. These results are encouraging because no special techniques were used for Korean. With increased training data and text corpora we expect better performance. Prospective future research areas include increasing speech data and text corpora, efficient implementation of multiple pronunciations, using condition-dependent speech recognition, improvement of speaker adaptation, and implementation of a total system with speech segmentation capability for on-line operation.

Acknowledgment

This research was partly supported by the Brain Science & Engineering Research Program of the Korea Ministry of Science and Technology.

References

1. D. S. Pallet, "Overview of the 1997 DARPA speech recognition workshop," *Proc. 1997 DARPA Speech Recognition Workshop*, Feb. 1997.
2. J. S. Garofolo, J. G. Fiscus, W. M. Fisher, "Design and preparation of the 1996 HUB-4 broadcast news benchmark test corpora," *Proc. 1997 DARPA Speech Recognition Workshop*, Feb. 1997.
3. D. S. Pallet, J. G. Fiscus, J. S. Garofolo, A. Martin, M. A. Przybocki, "1998 Broadcast News Benchmark Test Results," *Proc. 1999 DARPA Broadcast News Workshop*, Feb. 1999.
4. D. S. Pallett, J. Fiscus, M. Przybocki, "Broadcast News 1999 Test Results," *Proc. 2000 DARPA Speech Transcription Workshop*, May, 2000.
5. K. Ohtsuki, S. Furui, N. Sakurai, A. Iwasaki, Z. P. Zhang, "Improvements in Japanese Broadcast News Transcription," *Proc. 1999 DARPA Broadcast News Transcription*, Feb. 1999.
6. K. Ohtshuki, T. Matsuoka, T. Mori, K. Yoshida, Y. Taguchi, S. Furui, K. Shirai, "Japanese large-vocabulary continuous-speech recognition using a newspaper corpus and broadcast news," *Speech Communication* 28, pp. 155-166, 1999.
7. H. J. Yu, H. Kim, J. S. Choi, J. M. Hong, K. S. Park, J. S. Lee, H. Y. Lee, "Automatic recognition of Korean broadcast news speech," *Proc. ICSLP'98*, Sydney, Australia, Dec. 1998.
8. O. W. Kwon, K. Hwang, J. Park, "Korean large vocabulary continuous speech recognition using pseudomorpheme units," *Proc. EUROSPEECH'99*, Budapest, Hungary, Sept. 1999.
9. O. W. Kwon, "Performance of LVCSR with morpheme-based and syllable-based recognition units," *Proc. ICASSP 2000*, pp. 1567-1570, June 2000.
10. P. Geutner, "Using morphology towards better large-vocabulary speech recognition systems," *Proc. ICASSP' 95*, Detroit, USA, May 1995.
11. P. Scheytt, P. Geutner, A. Waibel, "Serbo-Crotian LVCSR on the dictation and broadcasting news domain," *Proc. ICASSP'98*, Seattle, USA, May 1998.
12. L. M. Tomokiyo, K. Ries, "An automatic method for learning a Japanese lexicon for recognition of spontaneous speech," *Proc. ICASSP'98*, Seattle, USA, May 1998.
13. H. K. J. Kuo, W. Reichl, "Phrased-based language models for speech recognition," *EUROSPEECH'99*, Budapest, Hungary, Sept. 1999.
14. L. M. Tomokiyo, K. Ries, "An automatic method for learning a Japanese lexicon for recognition of spontaneous speech," *ICASSP'98*, Seattle, USA, May 1998.
15. K. Ries, F. D. Buo, A. Waibel, "Class phrase models for language modeling," *ICSLP'96*, Philadelphia, USA, Oct. 1996.
16. D. Kiecza, T. Schultz, A. Waibel, "Data-driven determination of appropriate dictionary units for Korean LVCSR," *Proc. International Conference on Speech Processing (ICSP'99)*, pp. 323-327, Aug. 1999.
17. G. S. Lee, A. Waibel, "Korean broadcast news speech recognition using HMM," *Proc. International Conference on Speech Processing (ICSP'99)*, Aug. 1999.
18. J. H. Kim, Lexical Disambiguation with Error-Driven Learning, Ph. D. dissert. Dept. Computer Science, Korea Advanced Institute of Science and Technology, 1996.
19. J. Jeon, S. Cha, M. Chung, J. Park, K. Hwang, "Automatic generation of Korean pronunciation variants by multistage applications of phonological rules," *Proc. ICSLP'98*, Sydney, Australia, Dec. 1998.
20. M. K. Ravishankar, Efficient Algorithms for Speech Recognition, PhD dissert., School of Computer Science, Carnegie Mellon Univ., 1996.
21. M. Finke, P. Geutner, H. Hild, T. Kemp, K. Ries, M. Westphal, "The Karlsruhe-Verbmobil speech recognition engine," *Proc. ICASSP'97*, Munich, Germany, 1997.
22. P. Clarkson, R. Rosenfeld, "Statistical language modeling using the CMU-Cambridge toolkit," *Proc. EUROSPEECH'97*, pp. 2707-2710, 1997.
23. S. M. Katz, "Estimation of probabilities from sparse data for the language model component of a speech recognizer," *IEEE Trans. Acoustics, Speech, and Signal Processing*, vol. 35, pp. 400-401, 1987.
24. R. Bakis, S. Chen, P. Gopalakrishnan, R. Gopinath, S. Maes, L. Polymenakos, and M. Franz, "Transcription of Broadcast News Shows with the IBM Large Vocabulary Speech Recognition System," *Proc. 1997 DARPA Speech Recognition Workshop*, Feb. 1997.
25. P. Beyerlein, X. Aubert, R. Haeb-Umbach, M. Harris,

Dietrich Klakow, A. Wendemuth, Sirko Molau, Michael Pitz, A. Sixtus, "The Philips/RWTH System for Transcription of Broadcast News," Proc. DARPA Broadcast News Transcription, Feb. 1999.

26. J. L. Gauvain,L. Lamel, G. Adda, M. Jardino, "The LIMSI 1998 HUB-4E Transcription system," *Proc. DARPA Broadcast News Transcription*, Feb. 1999.

27. H. J. Yu, H. Kim, J. M. Hong, M. S. Kim, J. S. Lee, "Large vocabulary Korean continuous speech recognition using a one-pass algorithm," *Proc. ICSLP 2000*, Oct. 2000.

[Profile]

• Oh-Wook Kwon

Oh-Wook Kwon received the B.S. degree in electronic engineering from Seoul National University, in 1986, and the M.S. and Ph.D. degrees from the Korea Advanced Institute of Science and Technology (KAIST) in 1988 and 1997, respectively. From 1988 he was with the Electronics and Telecommunications Research Institute, Daejeon, Korea as a principal member of technical staff in the Spoken Language Processing Team, working on spontaneous speech translation in multimedia environment and speech input/output processing for human-computer interface. From 2000 to 2001, he was with the Brain Science Research Center, KAIST as a research professor. Since 2001, he has been with the Institute for Neural Computation of University of California, San Diego. His research interests include speech recognition, speech signal processing, pattern recognition, and language processing.

• Alex Waibel

Alex Waibel is a Professor of Computer Science at Carnegie Mellon University, Pittsburgh and at the University of Karlsruhe (Germany). He directs the Interactive Systems Laboratories at both Universities with research emphasizing speech recognition, handwriting recognition, language processing, speech translation, machine learning and multimodal and multimedia interfaces. At Carnegie Mellon, he also serves as Associate Director of the Language Technology Institute and as Director of the Language Technology PhD program. He was one of the founding member of the CMU's Human Computer Interaction Institute (HCII) and continues on its core faculty. Dr. Waibel was one of the founders of C-STAR, the international consortium for speech translation research and served as its chairman from 1998-2000. His team has developed the JANUS speech translation system, the JANUS speech recognition toolkit, and a number of multimodal systems including the meeting room, the Genoa Meeting recognizer and meeting browser. Dr. Waibel received the B.S. in Electrical Engineering from the Massachusetts Institute of Technology in 1979, and his M.S. and Ph.D. degrees in Computer Science from Carnegie Mellon University in 1980 and 1986. His work on the Time Delay Neural Networks was awarded the IEEE best paper award in 1990, and his work on speech translation systems the "Alcatel SEL Research Prize for Technical Communication" in 1994.